

DOCTORAL THESIS

Open Environmental Data Assimilation Under Unknown Uncertainty and Multiple Spatio-Temporal Scales

Lizaveta Miasayedava

TALLINN UNIVERSITY OF TECHNOLOGY
DOCTORAL THESIS
13/2024

Open Environmental Data Assimilation Under Unknown Uncertainty and Multiple Spatio-Temporal Scales

LIZAVETA MIASAYEDAVA



TALLINN UNIVERSITY OF TECHNOLOGY
School of Information Technologies
Department of Computer Systems

The dissertation was accepted for the defence of the degree of Doctor of Philosophy (Computer and Systems Engineering) on March 13, 2024

Supervisor: Associate Professor Jeffrey Andrew Tuhtan,
Department of Computer Systems, School of Information Technologies,
Tallinn University of Technology
Tallinn, Estonia

Co-supervisor: Jaanus Kaugerand, PhD
Department of Software Science, School of Information Technologies,
Tallinn University of Technology
Tallinn, Estonia

Opponents: Professor Shinji Fukuda,
Water Resources Planning Laboratory,
Tokyo University of Agriculture and Technology,
Tokyo, Japan

Assistant Professor Anastasija Nikiforova,
Institute of Computer Science, Chair of Software Engineering,
University of Tartu,
Tartu, Estonia

Defence of the thesis: March 27, 2024, Tallinn

Declaration:

Hereby I declare that this doctoral thesis, my original investigation and achievement, submitted for the doctoral degree at Tallinn University of Technology, has not been submitted for any academic degree elsewhere.

Lizaveta Miasayedava

signature

Copyright: Lizaveta Miasayedava, 2024
ISSN 2585-6901 (PDF)
ISBN 978-9916-80-125-3 (PDF)
DOI <https://doi.org/10.23658/taltech.13/2024>

Miasayedava, L. (2024). *Open Environmental Data Assimilation Under Unknown Uncertainty and Multiple Spatio-Temporal Scales* [TalTech Press]. <https://doi.org/10.23658/taltech.13/2024>

TALLINNA TEHNIKAÜLIKOOL
DOKTORITÖÖ
13/2024

Keskkonna avaandmete assimilatsioon tundmatu määramatuse ning erinevate aeg-ruumi skaalade korral

LIZAVETA MIASAYEDAVA

Contents

List of Publications	6
Author's Contributions to the Publications	7
Abbreviations.....	8
Terms	9
1 Introduction	10
1.1 Background and Motivation	10
1.2 Benefits and Challenges in Open Environmental Data Assimilation	12
1.3 Research Questions	14
1.4 Contributions of the Thesis	15
1.5 Thesis Organization	16
2 State of the Art	17
2.1 A Brief Overview of Environmental Monitoring	17
2.2 Challenges in Using Open Government Data	18
2.3 Automated Environmental Compliance Monitoring and Reporting	19
2.4 Data Assimilation of Ambient Air Quality Data	19
2.5 Conclusion on the State of the Art	20
3 Automated Environmental Compliance Monitoring Using the Internet of Open Government Data and Things	21
3.1 Background.....	21
3.2 Data Sources.....	21
3.3 Methods	22
3.4 Results	23
3.4.1 Environmental Compliance Estimation Service	23
3.4.2 Data Quality Analysis.....	25
3.4.3 Proposed Solutions for Open Data Quality Challenges	26
3.5 Conclusion	26
4 Lightweight Urban Air Quality Data Assimilation.....	27
4.1 Background	27
4.2 Data Sources.....	27
4.3 Methods	29
4.4 Results	31
4.5 Conclusion	33
5 Open Pan-European Urban Air Quality Data Assimilation.....	34
5.1 Background	34
5.2 Data Sources.....	34
5.3 Methods	34
5.4 Results	37
5.5 Conclusion	39

6 Conclusions	41
List of Figures	43
List of Tables	44
References	45
Abstract	56
Kokkuvõte	57
Appendix 1	59
Appendix 2	71
Appendix 3	89
Curriculum Vitae	110
Elulookirjeldus	112

List of Publications

The present Ph.D. thesis is based on the following publications that are referred to in the text by Roman numbers.

- I Lizaveta Miasayedava, Keegan McBride, and Jeffrey A. Tuhtan. Automated Environmental Compliance Monitoring of Rivers with IoT and Open Government Data. *Journal of Environmental Management*, 303:114283, February 2022
- II Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty. *Environmental Modeling & Assessment*, June 2023
- III Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Open Data Assimilation of Pan-European Urban Air Quality. *IEEE Access*, 11:84670–84688, August 2023

Other Related Publications

This work is also published/presented in adjacent fields but does not form a main part of this thesis.

1. Jeffrey A. Tuhtan, Elizaveta Dubrovinskaya, Lizaveta Miasayedava, Vishwajeet Pattanik, Jürgen Soom, Bernd Mockenhaupt, Cornelia Schütz, Christian Haas, and Philipp Thumser. Smart Fish Counter for Monitoring Species, Size, Migration Behaviour and Environmental Conditions. In *The 2022 International Symposium on Ecohydraulics*, pages 1–4, 2022
2. Jeffrey A Tuhtan, Lizaveta Miasayedava, and Gert Toming. Data Assimilation of Acoustic Doppler Velocimeter and Total Pressure Sensors. Presentation at the 40th IAHR World Congress, 2023

Author's Contributions to the Publications

- I I was the main author, aggregated the data for the analysis, developed the methodology, designed and developed the software, prepared the visualization, and wrote the manuscript.
- II I was the lead author, aggregated the data, executed both the experimental design and analysis, designed, developed, and validated the algorithms and software, produced visualizations, and wrote the manuscript.
- III I was the first author, aggregated the data, designed and performed experiments and analyses, designed, developed, and validated the algorithms and software, created visualizations, and wrote the manuscript.

Abbreviations

AERCM	Automated Environmental Regulatory Compliance Monitoring
AQ	Air Quality
DA	Data Assimilation
EEA	European Environment Agency
EFCES	Environmental Flows Compliance Estimation Service
EIA	Impact Assessment Directive
EIC	Environmental Intelligence Cycle
EU	European Union
GDPR	General Data Protection Regulation
IoOGDT	Internet of Open Government Data and Things
IoT	Internet of Things
LSDA	Least-Squares Data Assimilation
MAU	Mean Absolute Uncertainty
OEDA	Open Environmental Data Assimilation
OGD	Open Government Data
RLS	Recursive Least Squares
RMSE	Root Mean Squared Error
RQ	Research Question
SILAM	System for Integrated modeLLing of Atmospheric composition
UI	User Interface
WFD	Water Framework Directive

Terms

AR(1)	First-order linear autoregression model
DA1	LSDA of two sources with known uncertainties, for the same temporal and spatial scales
DA2	LSDA of two sources with unknown uncertainties, for the same temporal and spatial scales
DA3	LSDA of two sources with unknown uncertainties, for the same temporal and different spatial scales (with spatial calibration)
DA4	LSDA of two sources with unknown uncertainties, for different temporal and spatial scales (with temporal and spatial calibration)
R(1)	First-order linear regression model
S-DA	Sequential LSDA using data from a single source with unknown uncertainty
S-DA4	Sequential DA4

1 Introduction

1.1 Background and Motivation

The global impacts of pollution have a substantial negative impact on mortality rates and reduce the quality of life [6]. At the policy level, the European Union (EU) has developed frameworks including the European Green Deal [7], Zero-Pollution Action Plan [8], and EU Environmental Impact Assessment (EIA) Directive [9], all of which rely heavily on robust and reliable environmental monitoring. Accordingly, the key to implementing these ambitious frameworks is ensuring that data are accurate and timely, enabling informed decision-making and policy implementation. This work introduces a new concept, automated environmental regulatory compliance monitoring (AERCM), which offers substantial advances over the state-of-the-art environmental monitoring paradigm. Specifically, the advantages of AERCM are that it:

- provides accurate, continuous data for informed environmental compliance monitoring and forecasting;
- enhances transparency in environmental regulatory compliance, management and policy-making;
- allows for local and large-scale health impact assessments, including climate change effects, considering pollution-related risks;
- improves public awareness and stakeholder collaboration.

Despite recent advancements in environmental monitoring, existing approaches may not fulfil all these needs. Therefore, AERCM aims to address these limitations and provide a comprehensive solution to environmental monitoring and compliance challenges.

The concept of AERCM stems from the nexus of relationships between decision-making, innovation, data security and privacy, and data integration, as illustrated in Fig.1. The motivation behind AERCM lies in addressing these interconnected aspects.

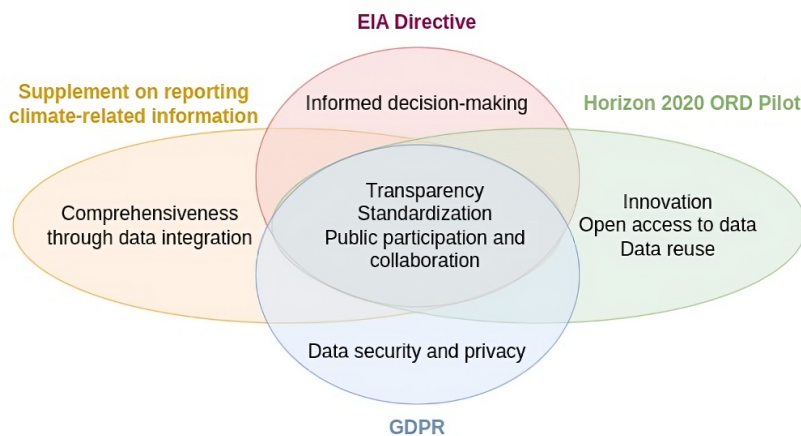


Figure 1 – AERCM lies at the nexus of relationships needed to address the ambitious goals of EU regulatory frameworks associated with environmental monitoring and compliance and can be visualized as sets of concurrent and inter-related agendas, with the common goals of improved transparency, and standardization facilitating public participation and collaboration.

In particular, the EU EIA Directive [9] mandates the identification and mitigation of negative impacts on the environment for environmental protection, which relies heavily on relevant environmental data during the assessment process. When policy is successfully implemented, it ensures that environmental pollution risks and stressors are integrated into the decision-making process. The EU guidelines outlined in the supplement on reporting climate-related information [10] also highlight that the enhancement of the quality and comprehensiveness of environmental reporting, must be achieved by integrating data from a broad range of sources including government agencies, research institutions and private organizations by employing standards or frameworks for the collection, management, and reporting of environmental data. The General Data Protection Regulation (GDPR) [11] guidelines encourage data security measures (such as encryption, access controls, and secure storage) and privacy measures (such as anonymization) to protect and minimize the exposure of personal data necessary for environmental research. Finally, the Horizon 2020 Open Research Data Pilot [12] promotes the principles of open research data management: responsible collection, management, and sharing of environmental data to make it findable, accessible, interoperable, and reusable for research and innovation, policy development, and public awareness.

All of the previously mentioned policies and guidelines encourage open access to environmental data and information to enhance transparency, and stimulate public engagement in governance, collaboration and innovation, which can lead to faster progress and economic growth. Thus, AERCM can implement the values outlined in the environmental policies and guidelines by merging several environmental data sources to obtain comprehensive and reliable estimates in an accessible and scalable way.

Previous studies highlight the need for implementing AERCM as a response to emerging environmental challenges. Subsequently, these studies demonstrate how integrating and assimilating in-situ and remote environmental monitoring and modeling data can bolster AERCM. For example, in the international fishing sector, on-board sensors paired with satellite-based remote sensing ensure adherence to fishing quotas and avoidance of restricted zones [13]. In urban development, ground-based sensors capture data on air quality (AQ), noise levels, and traffic patterns, whereas remote sensing through high-resolution satellite imagery monitors construction activities, helping to confirm alignment with zoning regulations and protection of restricted areas [14]. In the construction sector, spatial interpolation and noise propagation models are assimilated with field data collected by drones and noise sensors to continuously estimate noise levels within and outside the construction site, facilitating compliance with noise regulations [15]. Finally, the assimilation of observational and model-based data harmonizes anthropogenic land-use CO₂ flux estimations, supporting precise monitoring of climate change mitigation commitments [16] at the global scale.

To ensure transparency and accessibility of environmental monitoring and reporting, this dissertation further proposes **open environmental data assimilation (OEDA)** as a new and cost-efficient approach to data assimilation (DA) reusing existing, publicly available environmental data sources to support AERCM, as shown in Fig.2.

OEDA unleashes the potential of existing, open in-situ and remote sensing data and numerical models to substantially improve both the coverage and reliability of available data by incorporating potentially more accurate and comprehensive data across both spatial and temporal scales [17, 18]. This in turn can facilitate additional methodological advancements and cost-sharing [19, 20, 21]. Eventually, OEDA can allow for the automated analysis of complex environmental processes across multiple scales to identify regions needing enhanced data collection and model refinement [19].

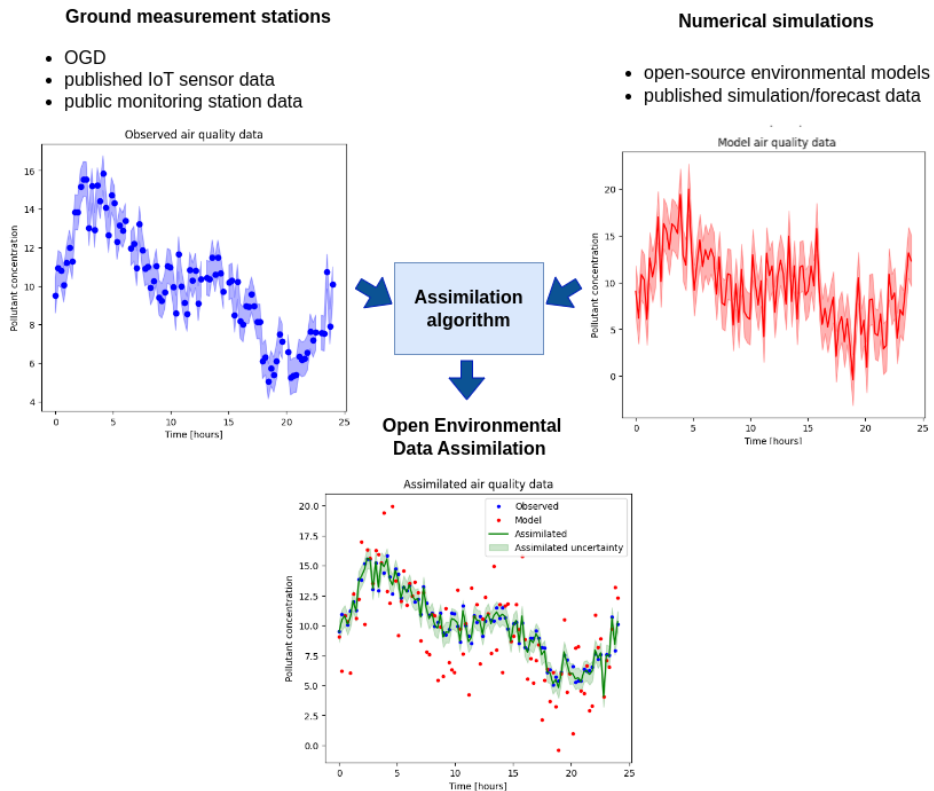


Figure 2 – Schematic representation of the OEDA process: Assimilating physical observations (top right, blue) and numerical model data (top left, red) from open data sources and with different errors and uncertainty bounds, to produce an analysis estimate of better quality and optimized uncertainty (bottom, green).

In summary, the integration of OEDA within the AERCM concept, as depicted in Fig. 3 is essential to provide a technologically-driven alignment of existing environmental monitoring and reporting policies and guidelines. OEDA facilitates the use of open data, its processing, and assimilation as part of the larger AERCM system. Specifically, OEDA enhances the quality of open data that facilitates more reliable information generation through analysis, compliance forecasting, and reporting, meeting the diverse needs of multiple end-users. This dissertation develops and explores real-world implementations of the OEDA concept using open environmental monitoring data first within Estonia and subsequently at the pan-European level.

1.2 Benefits and Challenges in Open Environmental Data Assimilation

OEDA integrates diverse publicly accessible environmental data into modeling frameworks to enhance the comprehensiveness and reliability of the final estimates. In other words, when applied to open data, DA algorithms facilitate the retrieval of the best possible estimate given the openly available data sources.

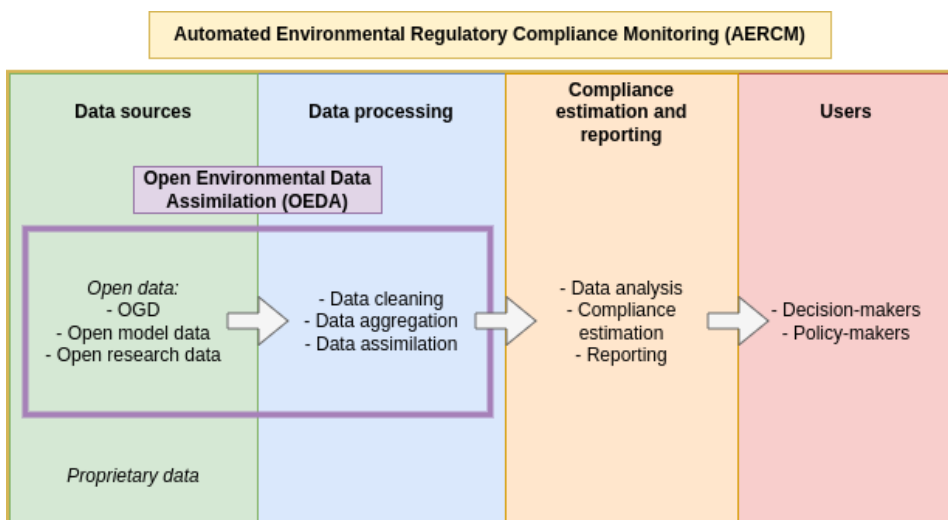


Figure 3 – Representation of the AERCМ concept including OEDA, which encompasses open data collection and data processing. AERCМ includes the additional stages of compliance estimation and reporting, ultimately the reports are used for decision- and policy-making.

The "best possible estimate" can be characterized considering the improvements in data quality parameters [22, 23, 24] which include:

- **accuracy**, which results in lower errors when compared with the true data values;
- **completeness**, by reducing the amount of missing data;
- **precision**, which minimizes the parameter value uncertainty.

DA algorithms can impute missing data and minimize uncertainty to improve the accuracy of the final estimates [25]. However, it is first important to note that the extent of improvement after applying DA is largely dependent on the following considerations:

- **Selection of appropriate algorithms:** Assimilating large amounts of data of different types and scales creates multiple challenges regarding algorithm selection and might lead to significant computational challenges [26, 27, 28].
- **Data quality of input data sources:** Low data quality from input data sources might lead to low-quality assimilation results, which without validation may result in misleading decision-making [29, 30].
- **Temporal and spatial scales of input data sources:** Environmental models often operate at large spatial scales (several km), whereas DA uses local observations at point locations. This necessitates knowledge of appropriate calibration mechanisms, which in many cases are not available [31].
- **Relevance of uncertainty estimates of input data sources:** DA solves the problem of uncertainty minimization. To do so, this requires the uncertainty contributions from the initial conditions, parameters, measurements, and process errors to quantify the contribution of the input data sources to the output assimilation results [23, 32]. Therefore, DA is dependent on the availability of uncertainty estimates, which are missing in many cases of practical importance.

Thus the successful implementation of OEDA is dependent on data sources of satisfactory quality and choosing appropriate algorithms for data imputation, calibration of data sources at different temporal and/or spatial scales, and for uncertainty quantification. When all of these pre-conditions are met, the successful implementation of OEDA can lead to benefits aligning with the concept of AERCM outlined in Fig.1:

- **Informed decision-making:** OEDA informs policy- and decision-making by providing more comprehensive and reliable data [9].
- **Transparency and accessibility:** As the data used are openly accessible, it encourages transparency, fostering a collaborative environment for the public, researchers, and policymakers. This can enhance public trust in scientific findings and environmental policies based on the data [12].
- **Cost-effectiveness through shared infrastructure:** The sharing of infrastructure and collaborative DA efforts allow for greater cost-effectiveness. Shared costs and efforts enable smaller programs and agencies to benefit from advanced modeling techniques and extensive datasets they might not have had access to otherwise [19, 20, 21].
- **Enriched scientific understanding:** By employing algorithms that blend large- and small-scale estimates from global and regional data, OEDA can help in obtaining a better understanding of environmental processes. This in turn identifies gaps in current models and points to areas where additional data collection is needed [19].

1.3 Research Questions

The main challenges to implement OEDA, as identified in the previous section, include the presence of missing and inaccurate data, the mismatch of temporal and spatial scales, and the need to estimate the uncertainty of input data. To address these challenges, this dissertation poses the following three research questions (RQ):

RQ1 Which data processing methods are the most suitable to improve data completeness, accuracy, and precision of open environmental monitoring and modeling data?

RQ2 Can data assimilation be applied at different spatial and temporal scales using sources without uncertainty?

RQ3 Are computationally lightweight assimilation methods suitable for large-scale open environmental monitoring data?

RQ1 seeks to identify methods to enhance the quality of open environmental data by addressing challenges associated with missing, inaccurate and imprecise data. The solutions to these challenges are key for OEDA implementation, as they strengthen data reliability for informed decision- and policy-making in the environmental domain.

RQ2 considers the challenges of assimilating environmental data with scale mismatches and unknown a priori uncertainty. This RQ is integral to the OEDA framework, as it seeks to develop and refine algorithms capable of assimilating disparate data sources even in the absence of explicit uncertainty information. Such algorithms would not only need to manage the complexities arising from scale mismatches, but also reliably estimate and incorporate unknown uncertainties. Successfully addressing this RQ would enable the OEDA framework to assimilate a broader range of environmental data sources, thereby

enhancing the accuracy and applicability of its outputs for comprehensive environmental analysis and decision-making.

RQ3 focuses on the feasibility and effectiveness of the proposed DA methods within the context of large-scale open environmental monitoring networks, such as those utilized for urban AQ monitoring in Europe. Ultimately, the success of OEDA depends on the performance in real-world applications, which include evaluating the DA algorithm performance, comparing results with pan-European scale reference data and analyzing the potential for further improvements by future researchers.

1.4 Contributions of the Thesis

The contributions of the thesis consist of three first-author journal manuscripts, each focusing on addressing the three RQs, and are summarized in Table 1. The details of the contributions are described in publications **I**, **II**, and **III**. A summary of how these contributions related to the RQs is included in Chapters 3 and 4. The summarized answers to each of the three RQs are provided in Chapter 6 of this dissertation.

Table 1 – Publications (Publ.) and corresponding contributions to each of the three RQs. The open data and code repository are provided for each publication via the GitHub link provided in the reference.

Publ.	Contributions	RQs
I	<p>C1.1. Identification of benefits and challenges in AERCM.</p> <p>C1.2. Introduction and promotion of the IoOGDT concept and open data reuse as an alternative to new infrastructure deployment.</p> <p>C1.3. Design and implementation of a web UI for the AERCM solution for Estonian rivers using OGD.</p> <p>Repository: [33]</p>	RQ1
II	<p>C2.1. Identification of benefits and challenges in OEDA.</p> <p>C2.2. Introduction and implementation of the algorithms DA2, DA3 - lightweight data-driven preprocessing and assimilation methods for data with unknown uncertainty estimates and varying spatial scales.</p> <p>C2.3. Performance validation of the algorithms DA2, DA3 using Tallinn AQ monitoring station (OGD) and Internet of Things (IoT) sensor data.</p> <p>Repository: [34]</p>	RQ2
III	<p>C3.1. Introduction and implementation of the algorithms S-DA, DA4, S-DA4 - lightweight data-driven preprocessing and assimilation methods for data with unknown uncertainty estimates and varying temporal and spatial scales.</p> <p>C3.2. Performance validation of the algorithms DA3, S-DA, DA4, S-DA4 using pan-European urban AQ monitoring station data (OGD) to improve data quality across scales.</p> <p>Repository: [34]</p>	RQ2, RQ3

1.5 Thesis Organization

This dissertation is structured into five main chapters as follows:

Chapter 1 offers a comprehensive introduction to the field of study. It provides the motivation behind the research, identifies benefits and challenges in OEDA, presents the main RQs, and lays out the significant contributions made through this work.

Chapter 2 provides a review of the current landscape in environmental monitoring and OGD utilization. It also gives an in-depth look into AERCM and the DA of ambient AQ data. The chapter concludes with a synthesis of the state of the art, identifying existing knowledge gaps.

Chapter 3 explores the integration of OGD with the IoT to advance AERCM. It addresses specific challenges in AERCM implementation and suggests potential solutions, facilitating improved and streamlined data utilization through OEDA. This chapter also includes the methods and results described in publication **I**.

Chapter 4 examines the challenges in implementing OEDA and introduces a lightweight open data-driven DA framework, demonstrating enhanced data quality for tackling these challenges. It also details the methods and results described in publication **II**.

Chapter 5 introduces new methods to the previously developed lightweight, data-driven DA framework and discusses their validation using an extended dataset from real-world, pan-European urban AQ monitoring. Additionally, this chapter presents the methods and results described in publication **III**.

Finally, **Chapter 6** serves as a conclusion for the dissertation, summarizing the key objectives and the RQs that drove this study, alongside the solutions provided through the development and validation of novel DA algorithms and methods. It reiterates on the significance of the study, reviews its implications, and suggests future research directions in the field.

2 State of the Art

2.1 A Brief Overview of Environmental Monitoring

Environmental monitoring plays a crucial role in studying, understanding and managing the dynamics of our environment, especially under the rapidly changing conditions driven by human activities. It serves as a foundational tool, providing essential data for researchers and decision-makers to assess the state of the environment, evaluate management strategies, and foster informed decision-making for sustainable resource use and conservation [35]. The effectiveness of environmental monitoring is significantly heightened when it is meticulously designed and executed, ensuring that the data collected is accurate, timely, and reliable [36, 37].

However, the accuracy and utility of environmental monitoring can be compromised by poorly implemented physical systems and data processing methods. Common issues include improperly calibrated equipment or inadequate deployment locations can yield misleading data [38, 39]. Furthermore, the lack of real-time data retrieval mechanisms, such as wireless sensor networks, can result in timing failures and delays, crucially impacting decision-making processes, particularly in environmental emergencies [40]. The absence of standardized data collection methods and protocols can also lead to inconsistent data that is difficult to compare and analyze [41]. Additionally, failing to integrate more recent data sources and technologies like remote sensing may result in less efficient and comprehensive monitoring [42, 43].

Environmental modeling complements monitoring, principally by addressing gaps and flaws in monitoring data. Techniques include numerical, machine learning-driven, and multi-source spatial data modeling frameworks [44], along with evolutionary computational and fuzzy rule-based models [45]. Despite challenges in model accuracy and reliability due to assumptions and parameter uncertainty [46], the combination of comprehensive modeling with monitoring data can significantly improve the reliability and accuracy of state estimates. DA is a methodology that integrates these approaches, continuously incorporating new observations to enhance model predictions and reduce prediction uncertainty [47, 48]. DA methods, including optimal interpolation, statistical, variational, ensemble, and hybrid methods, are employed across disciplines such as geophysics, hydrology, meteorology, and oceanography [49, 50, 51].

The current state of the art in academia, industry, and the public sector can be summarized in the following manner:

Academia: In academia, the shift towards employing complex and computationally intensive DA methods, often used in numerical weather prediction and environmental modeling, represents the cutting edge of research [27]. Global models, like the System for Integrated modeLLing of Atmospheric coMposition (SILAM) used in this work, play a crucial role in regional and global AQ and pollutant dispersion studies [52]. SILAM's publicly available data is invaluable for environmental research, aiding in the development of new applications and testing of pioneering modeling and DA techniques.

Industry: The commercial sector has significantly advanced environmental monitoring through the development and implementation of proprietary technologies and sensor networks, integrating sophisticated data processing algorithms for efficient and comprehensive monitoring [53]. Enterprises in several sectors are actively deploying sensor networks capable of collecting a diverse array of environmental data. These networks, often equipped with advanced IoT capabilities, facilitate real-time data transmission and analysis [40, 43]. The immediacy of this data allows for quicker responses to environmental changes, leading to more effective management strategies.

Additionally, industries are leveraging artificial intelligence and machine learning to analyze the collected data. These technologies enable the prediction of environmental trends and potential hazards, facilitating proactive measures [44, 45]. For instance, predictive modeling based on sensor data can anticipate air pollution levels, water quality changes, or the impacts of industrial activities on local ecosystems [53].

Through integration of IoT and artificial intelligence in environmental monitoring, companies can monitor, predict, and respond to environmental issues more efficiently and effectively. A prime example is IBM's Green Horizons initiative, demonstrating how large-scale sensor data combined with artificial intelligence can be utilized for advanced environmental predictions, showcasing the industry's active role in managing and mitigating environmental impacts through technology [53].

Public sector: Governments and public agencies, such as the European Environment Agency (EEA), are adopting open data policies to enhance monitoring, thereby improving data accessibility and standardization [54]. This integration of advanced DA methods and open data policies across sectors reflects a significant shift towards collaborative and integrated approaches to environmental monitoring, enhancing the ability to effectively predict and respond to environmental changes.

2.2 Challenges in Using Open Government Data

Governments are pivotal in the production and collection of data across various domains, positioning them as primary providers of valuable information to the public [55]. Open government data (OGD) typically comprises digital categorical and numerical data and information that governments release. This data plays a crucial role in analyzing public policies and fostering transparency and accountability, which are essential for combating corruption [56, 57]. Additionally, OGD is instrumental in generating economic and social value [58, 59], supporting sustainability-driven behavioral changes [60], and enhancing citizen collaboration and participation [61]. The focus of OGD lies in its disclosure, accessibility, and reuse, prompting many countries to develop laws and policies that ensure proper indexing and availability of data [62, 63, 64].

Despite the apparent benefits of OGD in promoting innovation, transparency, and citizen participation, government agencies often exhibit reluctance in opening their datasets and integrating data publication into their daily operations [55]. A significant portion of the datasets advertised as open on government websites may not be readily accessible to the public, and many published resources primarily offer informational content rather than granular source data, thus potentially limiting their practical utility [65]. This hesitation arises from various challenges, including ensuring a continuous data supply, providing access to necessary open data infrastructures, and allocating sufficient resources like hardware, software, and finances [66, 60]. Concerns about privacy, data quality, and interoperability also play a significant role in deterring agencies from making data openly available [67, 56, 68]. However, it's worth noting that advancements in OGD initiatives, such as the EU Open Data Directive [69], have aimed to address these challenges by promoting the publication of more accessible and granular datasets [70].

The effectiveness of OGD initiatives in driving innovation heavily relies on data quality, influenced by the processes of collection, management, and dissemination [55], governance frameworks [71], and the capabilities of data providers and users [72]. Data quality is determined by attributes such as completeness, accuracy, timeliness, and consistency [62]. However, challenges like inconsistent government policies, lack of standard formats, interoperability issues, and privacy concerns [72, 71] can impede these initiatives. To overcome these challenges, robust data verification [73], standardized metadata [74], and reg-

ular quality assessments [75] are essential, ensuring the data remains accurate and reliable for innovation and decision-making purposes.

2.3 Automated Environmental Compliance Monitoring and Reporting

AERCM systems play a pivotal role in managing the complex environmental regulations and commitments faced by various industries [76]. These systems are integral in promoting sustainable practices across multiple sectors, including the commercial fishing industry [13], chemical production [77], urban development planning [14], manufacturing [78], construction [15], and tourism [79].

The implementation of AERCM systems tailored for environmental monitoring presents several challenges:

- the limited availability, completeness, coverage, and quality of data can hinder accurate compliance estimation [80];
- the labour-intensive nature of implementing and maintaining compliance management systems across different jurisdictions [81];
- the complexity of providing both detailed and summary information through compliance software [82];
- challenges in record keeping and communicating detected problems across multiple facilities [82, 81];
- technical and financial barriers in deploying monitoring systems, particularly in low-resource countries [80, 77].

Overcoming these challenges requires thorough planning, appropriate compliance software, continuous monitoring, and collaborative efforts among organizations. Despite not being widely adopted in official compliance processes, AERCM systems offer significant benefits, such as cost savings through streamlined processes, reduced manual labor, and ensured adherence to regulatory requirements [83, 84, 85]. Automated monitoring capabilities allow organizations to quickly identify and address compliance issues, minimizing environmental damage and reducing the risk of non-compliance penalties [86, 87]. Additionally, AERCM leads to improved resource allocation and optimization, as it enables better identification of underutilized data resources [83].

2.4 Data Assimilation of Ambient Air Quality Data

Effectively addressing the challenges posed by AERCM necessitates the utilization of DA. DA plays a critical role in improving the accuracy of AQ data by integrating observed data with predictive models [88, 89]. This process results in a more precise representation of the atmosphere, which is vital for air pollution management and public health [90, 91]. Consequently, DA aligns seamlessly with AERCM's goals of refining data essential for understanding and managing environmental health [92, 93].

DA techniques are particularly effective in managing the myriad sources of uncertainty inherent in AQ data. For instance, when integrating chemical transport model simulations with predictions, discrepancies between the predicted and observed data often point to uncertainties in aspects like model emissions, spatial resolution, chemical reaction mechanisms, process parameterizations, and measurement errors [89]. Moreover, the use of coupled chemistry-meteorology models allows for the assimilation of both meteorological

and chemical data, encompassing meteorological uncertainties for a more comprehensive approach to these challenges [18].

A significant benefit of DA is its capacity to improve AQ forecasts in atmospheric chemistry models. Ensuring accuracy in model predictions is imperative, especially considering the complexities introduced by errors in meteorological fields, urban structures, and the spatial transportation of pollutants in urban AQ simulations. Recent advancements advocate for the integration of more detailed environmental and meteorological interactions and the use of cutting-edge computational methods alongside remote sensing data to mitigate modeling errors and enhance the accuracy of AQ predictions [94, 95].

This work will use OGD from local Estonian and pan-European urban AQ monitoring ground stations and a global numerical model to develop and test computationally lightweight DA methods that are suitable for low-powered microcontrollers used by IoT devices. By enhancing the precision of AQ forecasts and effectively tackling the uncertainties in environmental modeling, these new and sophisticated techniques will play an indispensable role in guiding informed decision-making for air pollution control and public health protection.

2.5 Conclusion on the State of the Art

Environmental monitoring is a critical component across various sectors, offering essential data for effective resource management and planning. The EEA has underscored the urgent need for robust environmental monitoring in Europe, given the region's significant environmental challenges [54]. However, environmental monitoring programs frequently encounter gaps and flaws in data, necessitating the integration of data from diverse sources such as global models, remote sensing, and wireless sensor networks to ensure comprehensive data collection [96, 97].

Systematic environmental data collection can be resource-intensive, but the utilization of existing, under-utilized open data presents a viable solution to reduce costs and minimize the need for additional infrastructure [98, 99, 100]. In this regard, OGD initiatives play a significant role. Nonetheless, the effectiveness of these initiatives largely depends on the accessibility and quality of the data provided [101, 102, 103].

DA techniques are increasingly used for enhancing the quality of environmental data. These techniques effectively bridge the gap between observed data and model predictions, thereby improving accuracy and reliability [104, 91]. However, the success of DA is contingent upon the quality of the input data. Challenges such as missing data and varying data resolutions can significantly impact the outcomes of DA processes [105, 106]. Consequently, accurate estimation of uncertainties is crucial for effective DA implementation [107, 108]. These uncertainty estimates are fundamental for informed decision-making, which is critical to environmental management and compliance monitoring. For example, automated compliance monitoring systems are becoming increasingly important for adhering to environmental regulations and can offer significant cost-saving benefits [109, 110, 111].

In conclusion, environmental monitoring remains a cornerstone of informed decision-making and effective management in the face of growing natural and human-induced uncertainty. The integration of DA and effective uncertainty estimation, coupled with the leveraging of open data will be essential to address the complexities and increasing volume of environmental data, thereby enhancing the overall quality and reliability of environmental monitoring efforts.

3 Automated Environmental Compliance Monitoring Using the Internet of Open Government Data and Things

3.1 Background

The EU Water Framework Directive (WFD) requires the calculation of environmental flows (eflows) to maintain water quality, yet its implementation varies across regions due to complex ecosystem dynamics, ambiguous regulations, and resource limitations [112, 113, 114]. Traditional methods for estimating eflows are often grounded in extensive ecological research, which, while thorough, can be time-consuming and challenging to scale [115]. To overcome these challenges, the study outlined in I proposes the innovative integration of OGD with IoT data to facilitate continuous AERCM within the framework of the EU WFD. This approach aligns with regulatory values, as depicted in Fig. 1, and represents a significant stride in enhancing environmental monitoring efficiency.

The study introduces the concept of the Internet of Open Government Data and Things (IoOGDT). This concept merges existing OGD with IoT-collected data to create an automated system for AERCM. By doing so, it fosters open data reuse, streamlines the monitoring process through automation, and mitigates costs associated with new investments in data collection and processing, utilizing existing data infrastructure (C1.2). A practical application of this system is demonstrated with Estonia's national river gauging station network, used for monitoring eflows in rivers as mandated by the EU WFD [112]. The implementation of IoOGDT, particularly in facilitating continuous access to river OGD through IoT technology, automates compliance monitoring. This automation primarily involves the retrieval and analysis of data already published on the OGD portal, rather than the initial provision of new data. Integrating IoT data with OGD enhances relevance, accuracy, timeliness, and usability [116]. IoT networks provide continuously collected, timestamped data which, when combined with OGD, yields timely insights critical for environmental monitoring. This integration ensures rapid responses to environmental changes and adherence to policy requirements [117].

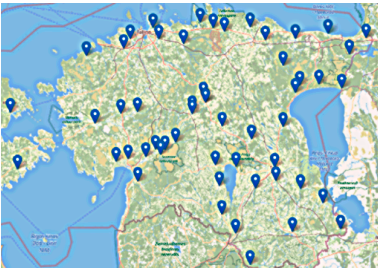
The IoOGDT framework underscores the synergistic potential between government-released data and IoT. OGD can amplify the impact of IoT data, while IoT can drive improvements in the quality and utility of OGD. This reciprocal relationship fosters reliable and innovative environmental monitoring solutions [118, 119, 116, 117]. IoOGDT enables the establishment of a unified framework for managing, processing and interpreting diverse data sets. This framework facilitates the creation of comprehensive overviews, enables meaningful comparisons across regions, and offers flexibility to scale and adapt to emerging data sources or methodologies.

Extending beyond its current application, IoOGDT holds the potential for broader environmental management contexts. By harnessing the power of real-time data and OGD, it can be pivotal in addressing various environmental challenges, from urban planning to climate change mitigation. The continuous evolution and integration of IoT technology with OGD are poised to unlock new frontiers in environmental monitoring, offering scalable, cost-effective, and dynamic solutions to meet the ever-growing demands of sustainable environmental governance.

3.2 Data Sources

The study is based on the repurposing of open Estonian national river discharge monitoring data as detailed in Table 2. A total of 54 river gauging stations were used, whose data ranged from 1867 to 2020 and included the daily mean flow rates and water levels.

Table 2 – Map, sources, variables, and time periods of the river data used in publication 1.

Map	Data Source	River Data
 <p>Map displaying 54 gauging stations in Estonia, data sources for 1.</p>	River OGD, managed by the Estonian Ministry of the Environment [120].	Daily mean flow rates [m ³ /s], water levels [cm]. Time period: from 1867 to 2020.
	Unpublished river data, provided by the Estonian Ministry of the Environment.	Daily minimum, mean, maximum flow rates [m ³ /s], water temperatures [°C], water levels [cm]. Time period: from 2009 to 2018.

3.3 Methods

The study utilizes the Environmental Intelligence Cycle (EIC) depicted in Fig. 4. This cycle represents a design science research methodology tailored for the iterative development and validation of intelligent environmental technology solutions [121, 122].

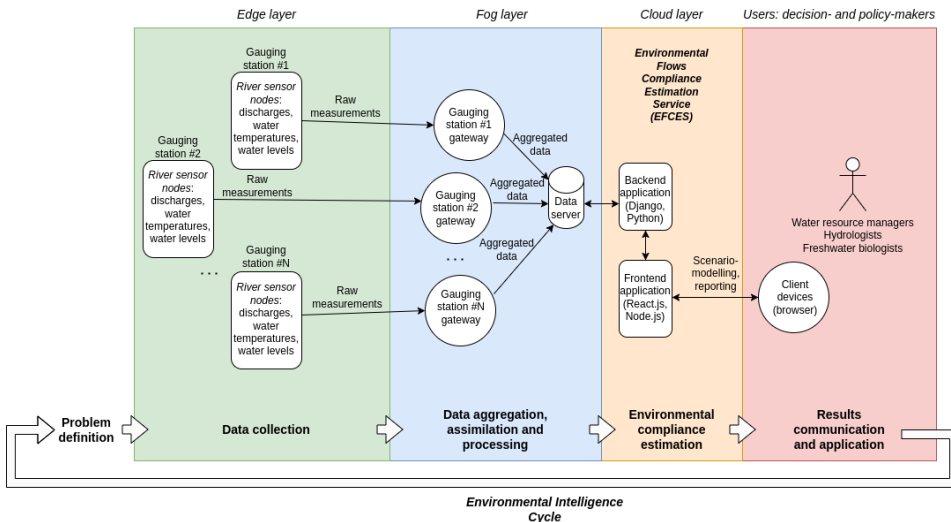


Figure 4 – Visual representation of the EIC methodology and its corresponding IoT architecture for the AERCM system for river eflows, adapted from publication 1.

The data flow and processes within the IoT system are shown in Fig. 4, which is designed for river eflow estimation, specifically the determination of exceedance levels as required by environmental regulations, specifically the EU WFD. The figure also delineates how multiple processes and architectural layers relate to the EIC methodology. The cycle initiates with the problem definition phase, which identifies stakeholders, sets decision-support goals, and ascertains the necessary data for the problem at hand (e.g., the requirements of water managers, hydrologists, and biologists). Subsequently, an IoT system

is implemented to automate the problem's resolution. The IoT system comprises three layers:

- The *Edge layer* used for data collection. In the provided illustration, this layer consists of multiple gauging stations, each containing river sensor nodes that monitor attributes such as discharges (flow rates), water temperatures, and water levels. These raw measurements are then relayed to the *Fog layer*.
- The *Fog layer* is charged with data aggregation, assimilation, and processing. These operations are proposed to be executed by the gateways affiliated with their respective gauging stations. Subsequently, this processed data is conveyed to a central data server, which accumulates aggregated data from all gateways and forwards it for in-depth analysis in the *Cloud layer*.
- The *Cloud layer* serves as the hub for data analysis and compliance assessment. For the eflows AERCM, the study conceptualizes and formulates the Environmental Flows Compliance Estimation Service (EFCES), a web service that implements eflow formulas and interprets results essential for scenario modeling and reporting. EFCES comprises a backend application (Django, Python) that manages data extraction from the data server, compliance assessment, and interaction with the frontend application. The frontend application (React.js, Node.js) offers a user interface (UI) for scenario modeling and result exportation for reporting purposes.

End-users, including water resource managers, hydrologists, and freshwater biologists, retrieve results and insights from EFCES using their client devices, accessed via a browser. These outcomes provide information for decision- and policy-making. Following this, a fresh iteration of the EIC is launched by revising the specifications in the "Problem definition" phase. This constitutes a perpetual loop where feedback can reconfigure the initial problem statement, thereby reinitiating the cycle. Finally, it is worth mentioning that the data collection phase may include deploying IoT sensors, monitoring stations, satellites, or modeling infrastructure for simulations or forecasts. Additionally, to reduce costs associated with new infrastructure, the study I also explores repurposing existing river data released as OGD, transitioning the system into the loOGDT system - a new concept that was first introduced in this study.

3.4 Results

3.4.1 Environmental Compliance Estimation Service

As a result of the research, a cloud-based component of the loOGDT system was developed to automate the estimation and interpretation of eflows compliance. This work provides EFCES - a custom, containerized web application built with Django/React. The application is based on OGD taken from 54 river gauging stations in Estonia and offers a suite of features for water managers, including:

- acquisition and manipulation of hydrological data;
- estimation of eflows;
- interpretation of compliance results;
- customization of estimation parameters;
- exporting compliance data.

EFCEs functions provide the required eflows calculation methods, allowing users to set the parameters for eflows estimation. Calculations are performed on the backend (Python, Django), and visual results are displayed on the frontend (React). This provides water managers with an efficient tool for generating estimations, visualizations, and reports. End users, including water resource managers, hydrologists, and biologists, access the system via the EFCEs interface to retrieve reports and test different estimation methods using available river data. EFCEs demonstrated its functionality using publicly available hydrological station data from Estonia, implementing environmental regulations to calculate minimum eflows based on a 95% exceedance probability, as detailed in publication I as established in the Estonian environmental regulation [123]:

$$Q_{env} = Q_{desc}[p(N + 1)], \quad (1)$$

where Q_{env} represents the eflow discharge (or volumetric flow rate) [m^3/s], Q_{desc} is an array of observed discharge values sorted in descending order [array of m^3/s], p denotes the probability of exceedance for an observed discharge value (ranging from 0 to 1); for instance, for a 95% exceedance probability, $p = 0.95$, N is the total number of observations in Q_{desc} , $[\cdot]$ is the operator for rounding the index to the nearest integer and retrieving the value from the array using this index.

The demonstration includes an example of compliance analysis that can be performed with EFCEs. The interface allows for the calculation of the percentage of days in noncompliance with the chosen eflows regulation (such as Eq. 1). This compliance summary was calculated and described for three differently-sized Estonian rivers over various bioperiods, and the summary table is presented in Table 3.

Table 3 – Eflows compliance summary for three Estonian rivers in 2009-2018 (summarized from publication I).

River	Average percentage of days in noncompliance per bioperiod [%]			
	Overwintering (Jan-Feb)	Spring Spawning (Mar-Jun)	Rearing and Growth (Jul-Sep)	Fall Spawning (Oct-Dec)
Narva (large)	33.39	11.89	30.33	26.63
Piusa (medium)	8.31	1.56	28.59	5.43
Puhajõgi (small)	3.73	4.02	18.48	6.96

The percentages are calculated based on the total number of days within each bioperiod, highlighting the relative duration of noncompliance that could potentially impact the river ecosystem during key biological periods. Additionally, EFCEs produces compliance visualizations, such as the one in Fig. 5, and offers a user manual on GitHub [34]. Fig. 5 presents discharge and water temperatures for the year 2016 at the Aesoo station. The x-axis represents time, the left y-axis represents discharge [m^3/s], and the right y-axis represents temperature [$^{\circ}C$]. Compliant and non-compliant discharges are marked with green and red dots, respectively, and the eflow threshold is with a dotted blue line. Water temperatures are shown with an orange line. The user interface allows for the customization of measurements and includes markers for key biological periods for in-depth analysis. EFCEs's testing on different rivers demonstrated the variability in compliance, highlighting the complex dynamics of river ecosystems and the impact of climate change on hydrological management.

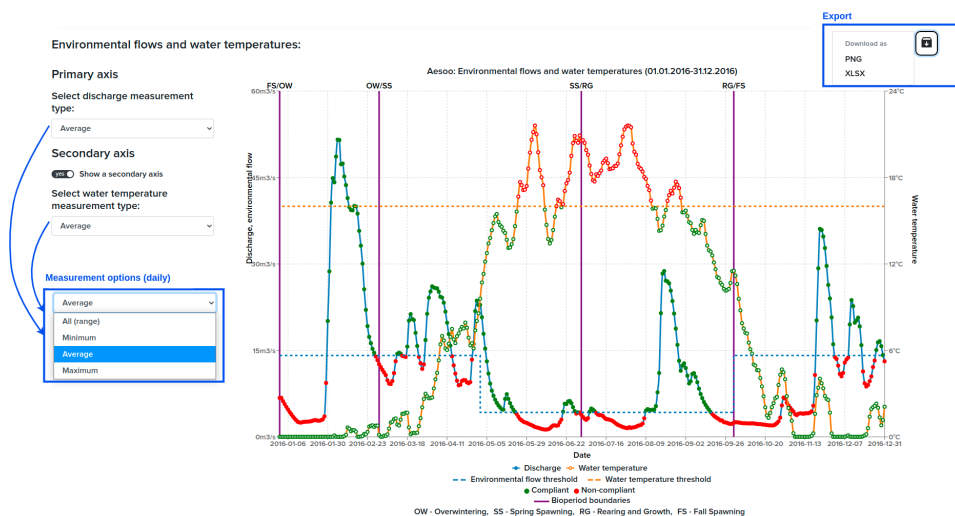


Figure 5 – Plot of eflows compliance indicating daily water discharge and temperatures with environmental threshold markers for the Aesoo station in 2016, generated by EFCES as developed in publication 1.

3.4.2 Data Quality Analysis

The data used for the development of EFCES highlighted challenges related to formatting differences, missing data, and inconsistencies. These issues were caused by measurement errors, manual processing mistakes, and data revisions. A comparison between "Open data" and "Requested data" from 2009-2018 for 52 common stations revealed varied percentages of missing data. Inconsistencies were identified by comparing the minimum, mean, and maximum values of daily time series of discharges and water levels. These findings are summarized in Table 4.

Table 4 – Overview of missing and inconsistent data in Estonian hydrological datasets: river data obtained from the government on request and OGD for the period from January 01, 2009, to December 31, 2018. Summarized from publication 1.

	Missing Data [%]		Inconsistent Data [%]	
	Requested data	OGD	Requested data	OGD
Discharge	10.61	2.11	16.89	48.47
Water levels	7.42	2.64	4.92	46.24

Challenges with data quality from different sources, including missing data and inconsistencies could result from several factors, including uncalibrated sensing devices and adverse environmental conditions. These issues underscore the need for meticulous data validation, a task that is particularly arduous for secondary users of the data. The study acknowledges the benefits of reusing existing data to bypass the costs of new environmental monitoring infrastructure. Yet, it cautions that repurposed OGD must meet rigorous standards of quality to serve new purposes effectively. When properly managed, such data can significantly improve the efficiency of environmental monitoring efforts, economizing on expenses.

3.4.3 Proposed Solutions for Open Data Quality Challenges

The study in I identifies key data quality challenges impacting the performance of the loOGDT system and proposes strategies to improve open environmental data, essential for effective environmental policy and decision-making (C1.1). During the implementation of the Estonian AERCM case, the following challenges and solutions were identified:

Data standardization: To address inconsistencies and fragmentation in datasets, such as those impacting environmental flow metrics, uniform data formats and protocols are necessary. Standardization, supported by ecological research, alongside OGD consolidation and API development for data sharing, can ensure consistency in metrics.

Automated continuous monitoring: Stale data in OGD repositories hinder continuous monitoring. By integrating IoT systems to continuously update OGD with current data, continuous monitoring can maintain data relevance and support proactive analysis.

Data contextualization: The lack of context in OGD often leads to misinterpretation. Enriching datasets with extensive metadata can clarify and enhance their utility.

Quality control and data assimilation: Sensor data quality is frequently compromised by inaccuracies, data gaps, malfunctions, and connectivity issues. Implementing rigorous quality control protocols, including regular maintenance and calibration, and deploying redundant sensors are crucial for data integrity. Additionally, DA techniques can address gaps and inconsistencies in OGD and IoT data, improving overall data completeness, accuracy, and precision.

3.5 Conclusion

The research outcomes in publication I were shown to directly address the challenges and opportunities associated with open environmental data, specifically those pertaining to RQ1.

RQ1: "Which data processing methods are the most suitable to improve data completeness, accuracy, and precision of open environmental monitoring and modeling data?"

Answer: A comprehensive DA framework is recommended to improve the completeness, accuracy, and precision of open environmental data. This approach effectively combines observational data with model predictions to refine estimates continually, thereby enhancing the robustness and reliability of environmental monitoring and modeling outcomes.

Addressing data quality challenges in environmental monitoring, the study suggests standardization, automated updates, and sensor maintenance to enhance OGD reliability, advancing RQ1's objectives. It leverages the AERCM approach for river eflows, in line with EU regulations, optimizing resources through the integration of IoT and OGD. Notable contributions include the development of the loOGDT framework (C1.2), a custom web UI for monitoring Estonian river eflows (EFCEs, C1.3), and architecture of the automated national-scale eflows compliance monitoring system (C1.1), operating within the EIC framework. Future research will extend these systems for broader regulatory compliance monitoring and stakeholder engagement, improving river management to support the EU WFD's ecological objectives.

4 Lightweight Urban Air Quality Data Assimilation

4.1 Background

Urban AQ monitoring and modeling can benefit from DA to fill in missing data and to improve the forecasting and reporting accuracy of ambient AQ, which is required for environmental compliance monitoring and reporting [124, 125, 27]. The most common AQ parameters reported are sulfur dioxide (SO₂), nitrogen dioxide (NO₂), carbon monoxide (CO), ozone (O₃) and particulate matter (PM_{2.5} and PM₁₀). The values of these parameters vary widely across multiple temporal and spatial scales, and are critical to localize and mitigate pollution sources and estimate associated health risks [125]. Yet, as publication II notes, no single source— be it IoT sensors, monitoring stations, numerical models, or satellites — can independently offer a complete and accurate dataset for urban environments [126, 127, 128, 129]. Such completeness is vital for dependable urban AQ monitoring and modeling, which is a cornerstone of effective regulatory compliance [124, 130].

Errors in AQ data can stem from multiple factors, including model discrepancies, measurement inaccuracies, and issues with data representation. Uncertainty is often quantified as error metrics from repeated measurements or simulations, such as standard deviations under the assumption of Gaussian error distribution, or it might be deduced from quality certificates or device specifications [131, 132]. The challenge lies in fully quantifying uncertainty due to the vast array of error sources, compounded by insufficient data on these sources, lack of knowledge about them, or the data collection and generation process itself [133, 134].

Fundamentally, DA techniques merge observations with numerical simulations—both error-prone—to estimate the true state of the environment. The goal is to reduce analysis errors by balancing background estimates from simulations with observational data [135]. These DA techniques require the weighting of data sources according to their uncertainties [136]. However, many sources, especially open data and IoT sensors, lack a priori uncertainty estimates.

Methods including Kalman filters, variational techniques, ensemble approaches, and hybrid methods are typically employed to improve data quality by integrating diverse sources [137, 138]. These, however, hinge on known uncertainty estimates — often missing from constantly updated numerical models and open-access data. Furthermore, the differing temporal and spatial scales of the data sources necessitate calibration operators that may not always be at hand. In response, the study presented in II proposes an approach to enable DA even without known uncertainty estimates by deriving these estimates from the recursively built model of the data itself.

4.2 Data Sources

The study assimilated hourly fixed-point air pollution observations detailed in Table 5 and corresponding hourly 0.2° grid forecasts from the SILAM public archive (retrieved in ZARR format from [139], transformed into CSV for processing). The data prepared for assimilation is available in [34]. The observations are obtained from the publicly available data (downloaded from [140] in CSV format) from the Liivalaia AQ monitoring station (59°25.86'N, 24°45.6'E), employing different professional measurement devices Horiba analyzers and a Met One BAM 1020, and two IoT sensors from Tallinn Smart City [141], located 60 meters and 700 meters from the station. The IoT sensors, which cost approximately 25 USD, were mounted on street lights at a height of 3.5 m and powered by solar energy. Each grid cell corresponds to a single hourly value, potentially leading to discrepancies between fixed-point observations and the spatial scale of the grid cells.

Table 5 – Observation sources, variables, and time periods of the air pollution data used in publication II from assimilation with the SILAM data. CO - carbon monoxide, NO₂ - nitrogen dioxide, O₃ - ozone, SO₂ - sulfur dioxide, PM_{2.5} - particulate matter ≤ 2.5 μm, PM₁₀ - particulate matter ≤ 10 μm.

Air Pollutant [μg/m ³]	Data Source	Time Period
CO	AQ monitoring station (OGD, managed by Estonian Environmental Research Center [140])	October 12, 2021 - November 10, 2021 ("Fall")
NO ₂		
O ₃		
SO ₂		
PM _{2.5}		
PM ₁₀	IoT sensors (60 and 700 meters away from the station; sourced from the Tallinn Smart City, managed by Thinnect [141])	October 12, 2021 - November 10, 2021 ("Fall")

The model data was sourced from the SILAM [52], an atmospheric dispersion model developed by the Finnish Meteorological Institute [142]. This model delivers a 0.2° grid of 4-day hourly air pollutant forecasts, with results refreshed daily and kept in a 30-day public archive [139], integrating several transport routines and transformation modules. The algorithms assimilate observations from Table 5 with the numerical simulations from the SILAM grid, as illustrated in Fig. 6.

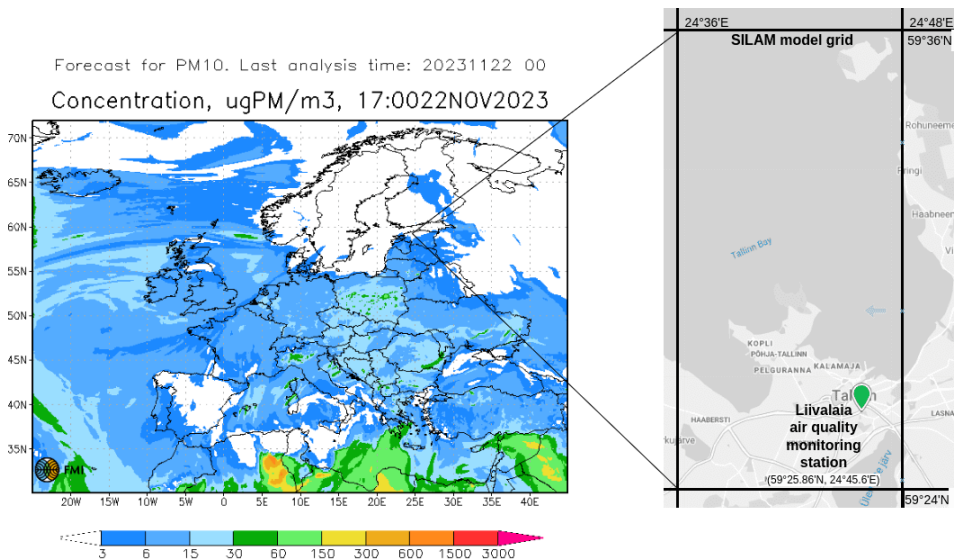


Figure 6 – Map showing the Liivalaia AQ monitoring station within the 0.2° SILAM model grid cell in Tallinn, Estonia, as used in publication II. The fixed-point station is set against the backdrop of the large-scale SILAM grid, underscoring the contrast in spatial scales. The SILAM grid was retrieved following [142].

4.3 Methods

The study introduces a new, lightweight framework for assimilating urban ambient AQ data without a priori uncertainty estimates. The goal is to enhance data quality in terms of completeness, precision, and accuracy, aligning with the focus of **RQ2**. The framework uses the least-squares data assimilation (LSDA) algorithm and estimates regression-based uncertainties recursively from the input data values.

In particular, the general LSDA formulas are computed as follows:

$$\begin{aligned}
 x_a &= k \cdot x_{obs} + (1 - k) \cdot x_m = x_m + k \cdot (x_{obs} - x_m), \\
 \sigma_a^2 &= (k \cdot \sigma_{obs})^2 + ((1 - k) \cdot \sigma_m)^2, \\
 k &= \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{obs}^2},
 \end{aligned} \tag{2}$$

where x_{obs} represents the observation (measurement), x_m - the background (model) estimate (simulation), x_a - the analysis estimate, k - a coefficient characterizing the contribution of x_{obs} to x_a , σ_m - the uncertainty of the background (model) estimate x_m , σ_{obs} - the uncertainty of the observation x_{obs} , σ_a - the uncertainty of the analysis estimate x_a .

When utilizing open data, both σ_m and σ_{obs} are frequently unknown, making the application of the traditional LSDA algorithm impossible. Furthermore, either x_m or x_{obs} may be absent, necessitating the imputation of missing data prior to employing LSDA. Instead, the study recommends estimating the unknown σ_m and σ_{obs} as errors ε_m and ε_{obs} from the recursively predicted values (termed "regression-based uncertainties"). The prediction is performed by 1st-order recursive least squares (RLS)-based linear (auto)regression filters.

By using a 1st-order RLS-based autoregressive (AR(1)) filter, instead of the unknown σ , the regression-based uncertainty ε of a data source (which can be applied to both x_{obs} and x_m) is proposed to be calculated as:

$$\begin{aligned}
 x^{pred}[t] &= w_1 \cdot x[t-1] + w_0, \\
 \varepsilon[t] &= |x[t] - x^{pred}[t]|,
 \end{aligned} \tag{3}$$

where $x^{pred}[t]$ represents the estimated value (predicted by AR(1)) at time t , w_0 and w_1 are the coefficients obtained from the RLS algorithm at time $t-1$ (when correlating $x[t-2]$ with $x[t-1]$), $x[t-1]$ is the value (either actual or imputed) at the previous time step $t-1$, $\varepsilon[t]$ denotes the prediction error at time t , $x[t]$ signifies the actual value at time t .

In Eq. 3, $x^{pred}[t]$ is only suggested for estimating $\varepsilon[t]$, and should not replace $x[t]$ when available. Otherwise, $x^{pred}[t]$ can act as an imputed value for $x[t]$. The coefficients w_0 and w_1 are updated post-prediction, implying that the prediction at time t utilizes coefficients determined at time $t-1$. To manage issues of differing temporal scales and the absence of a calibration operator, analogous procedures involving a 1st-order RLS-based linear regression (R(1)) filter can be executed. Initially, Eq. 2 should integrate the calibration of one data source to another (e.g., model values aligned with observations as x_m^c). By substituting the unknown σ_{obs} and σ_m with the regression-based uncertainties ε_{obs} and ε_m and adding calibration indicators, Eq. 2 can be modified as follows:

$$\begin{aligned}
x_a &= k \cdot x_{obs} + (1 - k) \cdot x_m^c = x_m^c + k \cdot (x_{obs} - x_m^c), \\
\epsilon_a^2 &= (k \cdot \epsilon_{obs})^2 + ((1 - k) \cdot \epsilon_m^c)^2, \\
k &= \frac{(\epsilon_m^c)^2}{(\epsilon_m^c)^2 + \epsilon_{obs}^2},
\end{aligned} \tag{4}$$

To derive the calibrated values x_m^c and its uncertainty ϵ_m^c , the R(1) filter is deployed. This filter matches model values to observations. Since the input values already possess uncertainty estimates, they are propagated in accordance with the rules of uncertainty (error) propagation:

$$\begin{aligned}
x_m^c[t] &= h_1 \cdot x_m[t] + h_0, \\
\epsilon_m^c[t] &= |h_1| \cdot \epsilon_m[t] + |x_{obs}[t - 1] - x_m^c[t - 1]|,
\end{aligned} \tag{5}$$

where $x_m^c[t]$ represents the calibrated (predicted by R(1)) model value at time t , h_0 and h_1 are coefficients from the RLS algorithm at time $t - 1$ (when fitting $x_m[t - 1]$ with $x_{obs}[t - 1]$), $x_m[t]$ is the non-calibrated model value at time t , $\epsilon_m[t]$ denotes the uncalibrated model AR(1) prediction error at time t , $\epsilon_m^c[t]$ signifies the calibrated and propagated model prediction error at time t , $x_{obs}[t]$ represents the observation value (primary data source) at time t .

In Eq. 5, $x_m^c[t]$ is calibrated linearly to align more closely with $x_{obs}[t]$. Consequently, the uncertainty of this calibrated value is a summation of the calibrated model uncertainty $|h_1| \cdot \epsilon_m[t]$ and the regression-based uncertainty of the calibration R(1) filter. As the calibration R(1) filter utilizes h_0 and h_1 adjusted at time $t - 1$, the calibration R(1) error $|x_{obs}[t - 1] - x_m^c[t - 1]|$ at time $t - 1$ is accordingly considered as a measure of how well the calibration performed for the coefficients.

Eqs. 2-5 were incorporated into algorithms **DA1**, **DA2**, and **DA3** depicted in Fig. 7.

DA1 is a traditional LSDA algorithm using known uncertainty estimates (see Eq.2), and **DA2** estimates unknown uncertainties using Eq.3 for each data source and uses them for Eq.2. **DA2** is designed for data sources with the same temporal and spatial scales (that do not require any calibration). Therefore, when the data sources have different spatial scales, spatial calibration using Eq.5 is recommended to be applied after the initial uncertainty estimation using Eq.3. This serves as the basis for the algorithm **DA3**, which uses the adjusted form of the traditional LSDA algorithm described by Eq. 4.

The validation involved comparing the performance of algorithms **DA2** and **DA3** when assimilating hourly simulations from the SILAM model with hourly observations from the Liivalaia AQ monitoring station. Performance was assessed using the root mean squared error (RMSE) and mean absolute uncertainty (MAU) metrics. A lower RMSE indicates closer alignment to the reference values. The RMSE is computed pair-wise for each of the AQ variables, comparing station observations, model simulations, and assimilated values, and is defined by Eq.6 as:

$$\text{RMSE}(x_1, x_2) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_1[i] - x_2[i])^2}, \tag{6}$$

where x_1 and x_2 are data value vectors of length n from the two sources, i - index variable.

To compare the regression-based uncertainties, the MAU metric was computed with Eq.7 for each AQ variable for station and sensor observations, model simulations, and

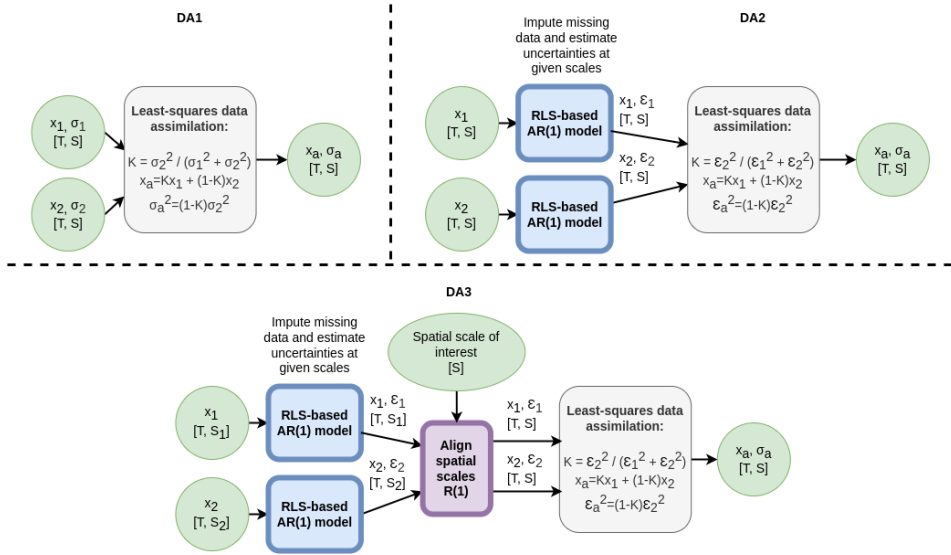


Figure 7 - Visual representation of the introduced **DA1**, **DA2** and **DA3** algorithms, derived from publication II.

assimilated values as:

$$\text{MAU}(\varepsilon) = \frac{1}{n} \sum_{i=1}^n |\varepsilon[i]|, \quad (7)$$

where ε is a vector of length n containing regression-based uncertainties $\varepsilon[i]$, with i as the index variable, each calculated using Eq. 5.

Finally, it should be noted that n in both Eq. 6 and Eq. 7 represents the total duration of the experiment (the number of hours in the corresponding time period) and is used for the overall assessment of the experiment.

4.4 Results

The study first analyzed AQ variables (CO, SO₂, PM_{2.5}, NO₂, O₃, PM₁₀) [$\mu\text{g}/\text{m}^3$] at the Liivalaia AQ station in Tallinn, Estonia, comparing station observations ("Station"), the SILAM data ("Model"), and assimilation results from **DA2** ("DA2") and **DA3** ("DA3") during October-November 2021 ("Fall") and January-February 2022 ("Winter"). The obtained RMSE and MAU metrics were detailed in Tables 1 and 2 of the manuscript II.

The RMSE between "Station" and "Model" was higher than the RMSE between "Station" and either "DA2" or "DA3". "DA3" consistently had the lowest RMSE but the highest MAU compared to "DA2" for both seasons across all the AQ variables. In October-November 2021 ("Fall"), PM₁₀ concentrations [$\mu\text{g}/\text{m}^3$] were assimilated with **DA2** and **DA3** using data from the Liivalaia station ("Reference Sensor"), IoT sensors at 60 meters ("IoT Sensor 60 m") and 700 meters ("IoT Sensor 700 m") from the station, and the SILAM model data ("Model"). For the station and IoT data, RMSE and MAU metrics were also calculated and described in Table 3 and Table 4 of the manuscript II, respectively.

Similar to the results of the first scenario, for all observation sources, **DA3** consistently outperformed **DA2** in terms of RMSE, though at the expense of a higher MAU. The examples of the validation scenarios with explanations are demonstrated in Table 6.

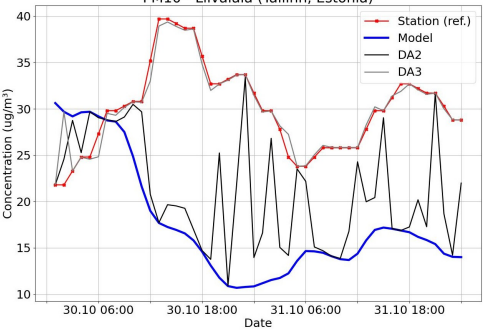
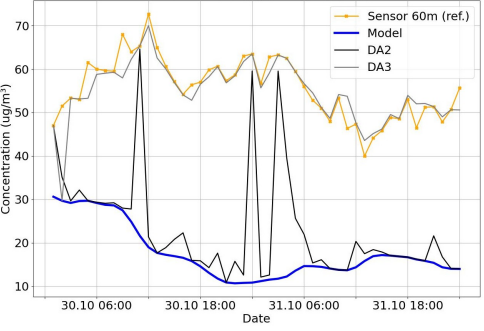
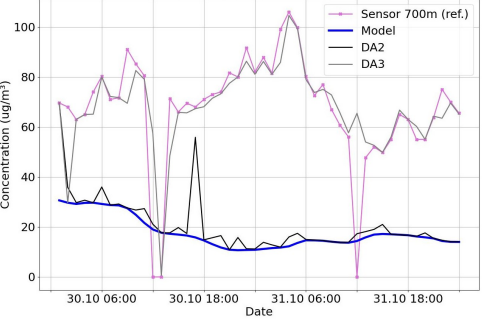
Visual example of a validation scenario	Performance metrics
<p style="text-align: center;">DA2 vs DA3: Reference Sensor:</p>  <p style="text-align: center;">IoT Sensor (60 meters from Reference):</p>  <p style="text-align: center;">IoT Sensor (700 meters from Reference):</p>  <p>Validation scenarios for the DA2 and DA3 algorithms over the period from October 30, 2021, to November 1, 2021, in assimilating hourly PM10 concentrations at Liivalaia, Tallinn, Estonia. "Station", "Sensor 60m", "Sensor 700m" correspond to x_{obs} "Model" - x_m, "DA2" - $x_a(DA2)$, "DA3" - $x_a(DA3)$. x_{obs} is used as the reference (ref.) for comparison; derived from publication II.</p>	<p style="text-align: center;">Performance metrics</p> <p>Accuracy metrics:</p> $r_{RMSE} = \frac{RMSE(x_a(DA2); x_{obs})}{RMSE(x_a(DA3); x_{obs})}$ <p>Precision metrics:</p> $r_{MAU} = \frac{MAU(\epsilon_a(DA2))}{MAU(\epsilon_a(DA3))}$ <p>Reference Sensor: For the visual example: $r_{RMSE} = 9.038$, $r_{MAU} = 0.694$ (DA2 is 9.038 times less accurate and 1.441 times more precise than DA3) From October 12, 2021, to November 10, 2021 (from II): $r_{RMSE} = 5.875$, $r_{MAU} = 0.417$ (DA2 is 5.875 times less accurate and 2.398 times more precise than DA3) From January 27, 2022, to February 25, 2022 (from II): $r_{RMSE} = 6.608$, $r_{MAU} = 0.554$ (DA2 is 6.608 times less accurate and 1.805 times more precise than DA3)</p> <p>IoT Sensor 60m: For the visual example: $r_{RMSE} = 8.344$, $r_{MAU} = 0.365$ (DA2 is 8.344 times less accurate and 2.740 times more precise than DA3) From October 12, 2021, to November 10, 2021 (from II): $r_{RMSE} = 1.200$, $r_{MAU} = 0.287$ (DA2 is 1.200 times less accurate and 3.484 times more precise than DA3)</p> <p>IoT Sensor 700m: For the visual example: $r_{RMSE} = 3.552$, $r_{MAU} = 0.160$ (DA2 is 3.552 times less accurate and 6.250 times more precise than DA3) From October 12, 2021, to November 10, 2021 (from II): $r_{RMSE} = 1.484$, $r_{MAU} = 0.213$ (DA2 is 1.484 times less accurate and 9.001 times more precise than DA3)</p>

Table 6 – Examples of comparative analysis of DA2 vs DA3 algorithms for hourly PM10 concentrations in Liivalaia (Tallinn, Estonia). Performance metrics assess accuracy and precision against the reference sensor values. Data source: II.

The overall validation results indicate that the assimilation results for both **DA2** and **DA3** algorithms were successfully able to reduce the uncertainty (MAU) relative to the uncertainty of the input data sources. **DA2** generally yields lower uncertainty than **DA3**, as evidenced by the lower MAU of **DA2** results. However, **DA3** might achieve greater accuracy (evident from the lower RMSE of **DA3** results compared to **DA2** results) because of its calibration step, albeit with a potential increase in uncertainty. In this context, the accuracy of **DA3** hinges on the accuracy of the reference data source used for calibration. Thus, in **DA3**, calibrating one data source (model) to another (observations) reduces the error between the reference and assimilated values, thereby enhancing accuracy. When assimilating model data with observations from IoT sensors, the magnitude of errors depends on the difference between station observations and IoT sensor measurements. Station observations exhibited the lowest uncertainty. **DA3** was most accurate for the station and least accurate for the most distant IoT sensor.

4.5 Conclusion

The study in publication II discusses the benefits and challenges of OEDA in the urban AQ monitoring domain. While improved data quality stands as a significant benefit, challenges such as assimilating data across different spatial and temporal scales and the absence of known uncertainty estimates are predominant. Furthermore, these challenges are often compounded by the requirement for substantial knowledge of the underlying models (C2.1).

These considerations have consequently informed the development of **RQ2**:

RQ2: "Can data assimilation be applied at different spatial and temporal scales using sources without uncertainty?"

Answer: The study introduces two lightweight DA methods, **DA2** and **DA3**, which are suitable for continuous execution and provide uncertainty estimates for LSDA. **DA2** performs LSDA using data sources of the same temporal and spatial scales without calibration. **DA3** includes a calibration step to handle data of different spatial scales. The uncertainty estimation methods are anchored in the propagated prediction errors from the AR(1) and R(1) models. These approaches align with contributions C2.2 and C2.3 in the thesis, which focus on the development and performance validation of these algorithms. The findings demonstrate the feasibility of applying DA in contexts where data sources lack pre-established uncertainty metrics, thereby expanding the potential of DA in future automated environmental monitoring and compliance reporting applications.

In a test case in Tallinn (C2.3), **DA2** achieved the minimum uncertainty, while **DA3** provided the least error when compared to the reference measurements. Additionally, both **DA2** and **DA3** seamlessly impute missing data using AR(1) filter predictions when no input value is available, enhancing the output dataset's completeness.

The **DA2** and **DA3** algorithms provide a standardized approach to manage missing uncertainty estimates in IoT devices with limited processing capacities. Using simple 1st-order models and linear operators, they are suitable for IoT applications. Their main objective is to improve urban AQ monitoring by incorporating IoT sensors and leveraging existing open data sources, thus broadening the coverage of existing networks and potentially lowering costs. Future work will focus on applying **DA2** and **DA3** to larger networks to further refine the quality of environmental data in terms of accuracy, precision and completeness.

5 Open Pan-European Urban Air Quality Data Assimilation

5.1 Background

Cities monitor their AQ to reduce pollution and plan for emergencies, combining ground data with computer models for better accuracy. The EEA provides a public AQ database [143] that, when paired with models using DA methods, improves analysis while considering data uncertainties [137, 144].

This chapter focuses on the study featured in publication **III**, which, in contrast to prior research on new IoT-based sensor networks [145, 146], aims to reuse OGD from the EEA AQ dataset [143]. The study demonstrates that, on a European scale, data from large-scale models such as SILAM [52] can be utilized without an in-depth understanding of the model itself and can be assimilated with the EEA AQ data to improve the accuracy of the estimates by using the proposed novel DA methods.

Using open data sources for DA presents challenges, particularly when uncertainty estimates are absent and the temporal and spatial scales differ [147], **II**. This research provides methods for estimating uncertainties using chained 1st-order recursive least squares (RLS) filters. The propagated errors from these filters have been shown to serve as data-driven uncertainty estimates for DA algorithms. The work does not analyze Europe's AQ but offers a method to improve the quality of existing data using DA with open numerical simulations. The goal is to enhance data quality using only timestamped air pollutant values and their location coordinates, without the need for additional information or uncertainty estimates.

This study extends prior work by introducing new algorithms for DA when data sources have differing spatial and temporal scales. Similar to the work in **II**, the work **III** addresses the need to improve the accuracy and completeness of urban ambient AQ monitoring and modeling through DA with the cost-efficient reuse of OGD. The latter work also validates all algorithms using OGD from European urban ambient AQ monitoring stations.

5.2 Data Sources

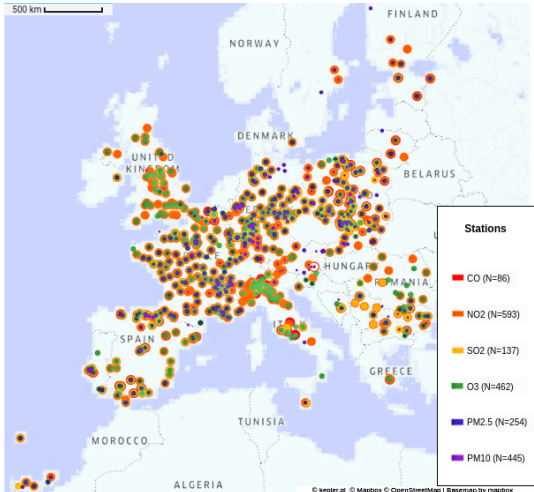
The study assimilated hourly fixed-point surface observations from the open EEA AQ data, exported in CSV format [143], and numerical simulations from the SILAM. The latter were retrieved in ZARR format from [139] and transformed into CSV for processing. SILAM is a global model that features an hourly 0.2° grid. The data prepared for assimilation is available in [34].

The EEA AQ dataset includes details on monitoring networks, stations, measurements, and assessment configurations. Specific filtering criteria such as station type (background) and area type (urban) were applied, ensuring less than 20% missing data, to select stations for validation. The distribution and numbers of the selected AQ monitoring stations are illustrated in Table 7. For validation, hourly values from the data sources were averaged over 24 hours to retrieve daily values. Data from each station were assimilated with simulation results from the corresponding SILAM grid cell. The same model was previously used for a single monitoring station (Liivalaia station referenced in publication **II**).

5.3 Methods

This study, found in publication **III**, presents three new DA methods outlined in Fig. 8, in addition to the previously introduced algorithms developed and validated in publication **II**:

Table 7 – Map and variables of air pollutants from European monitoring stations, the data from which were used in publication III.

Map	Air Pollutant	Number of Stations
 <p>Distribution of AQ monitoring stations in Europe, the data from which was assimilated in III.</p>	CO (carbon monoxide)	86
	NO2 (nitrogen dioxide)	593
	O3 (ozone)	462
	SO2 (sulfur dioxide)	137
	PM2.5 (particulate matter $\leq 2.5 \mu\text{m}$)	254
	PM10 (particulate matter $\leq 10 \mu\text{m}$)	445

- **S-DA**: sequential LSDA for a single source with unknown uncertainty, eliminating the need for a second data source to estimate uncertainty;
- **DA4**: LSDA for two sources with unknown uncertainties that have different temporal and spatial scales;
- **S-DA4**: sequential LSDA using results from **DA4** and its predictions for **S-DA**, suitable for two sources with unknown uncertainties and different temporal and spatial scales.

Each of the DA algorithms can be differentiated based on the number of data sources used, the presence of uncertainty estimates, and types of calibration. All the proposed DA algorithms (excluding **DA1**, which is a standard LSDA) utilize conventional 1st-order RLS filters for uncertainty estimation, spatial and temporal calibration, sequential estimation, and data imputation. The complexity of a single RLS filter is $O(L^2)$ with $L = 2$. The coefficients of the 1st-order linear equation are recursively refitted at each time step, resulting in constant time complexity, $O(1)$, which does not vary with more iterations. The proposed DA algorithms employ several RLS filters for 1st-order linear regression models to perform preprocessing tasks for LSDA, the numbers and execution times of which are shown in Table 8. The execution time was estimated using a computer with an Intel(R) Core(TM) i7-8565U CPU @1.80GHz \times 8 and 16Gb RAM (details are in publication III).

Each RLS-based 1st-order model is initially initialized with zero error and an identity state transition matrix. These are refitted at each step based on the input data values. The difference between the predicted and actual outcomes (the error) is computed and used to adjust the RLS model's parameters.

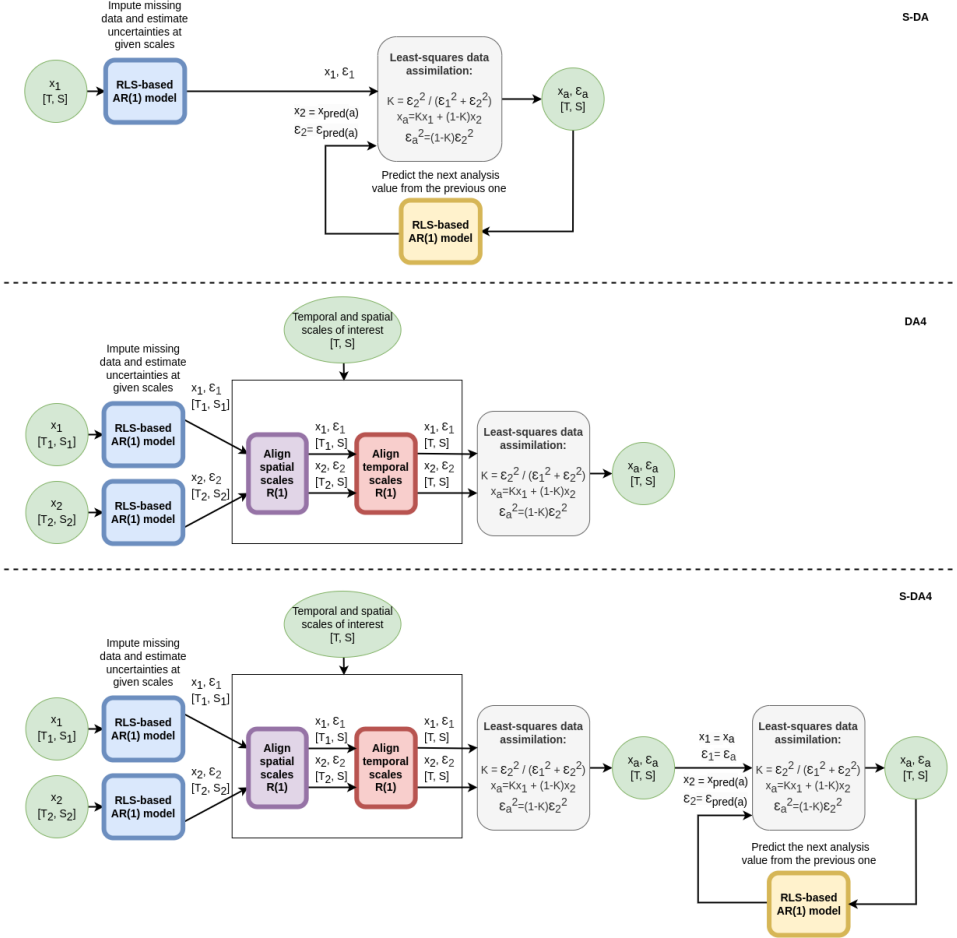


Figure 8 – Visual representation of the introduced S-DA, DA4 and S-DA4 algorithms, derived from publication III.

The RLS-based 1st-order autoregression AR(1) model is utilized to estimate initial uncertainties. This model predicts a current value using a previous one. When current data is missing, the RLS prediction fills the gap; otherwise, the actual value is used. The deviation from the prediction provides a regression-based uncertainty estimate (see Eq.3). This model is applied to each data source. In addition, the R(1) model is used for the spatial calibration of two datasets. It takes the outputs of the AR(1) model and aligns one data source to another with a different spatial scale. The error from the AR(1) model is adjusted according to uncertainty propagation rules and combined with the R(1) error (see Eq.5).

While DA2 solely employs the AR(1) model to impute missing data and estimate uncertainties, DA3 uses both AR(1) and R(1) models. The use of both models in DA3 is vital to add a spatial calibration step to one of the data sources. This ensures that data from both sources match in terms of scale and can be effectively compared or combined with LSDA (see Eq.4). Subsequently, LSDA is applied to obtain a refined analysis estimate considering uncertainties and calibration. When the temporal scales differ, temporal calibration is applied. In III, temporal calibration is illustrated using hourly and daily values, but this approach can be generalized to other temporal scales.

Table 8 – Average execution time of 1 iteration of the proposed DA algorithms. The execution time was estimated using a computer with an Intel(R) Core(TM) i7-8565U CPU @1.80GHz × 8 and 16Gb RAM.

Algorithm	Number of RLS Filters	Execution Time (ms)
DA2	2	0.056
S-DA	2	0.058
DA3	3	0.077
DA4	4	0.100
S-DA4	5	0.126

As detailed in publication III, hourly data can be converted to daily data by averaging the values over a 24-hour period. The "Recursive daily average estimator" (described in Algorithm 1 of publication III) processes hourly data, continuously updating daily averages. After receiving 24 data points, which equate to a full day, the daily average is recorded, and the estimator begins anew for the subsequent day. For a mix of hourly and monthly data, the transformation must account for the total hours in a month. To convert daily data back to hourly, an RLS-based 1st-order R(1) model is employed. This model fits daily data to predict hourly values, akin to the previously described R(1) model for spatial calibration. **DA4** conducts LSDA by calibrating both temporal (time-related) and spatial (space or location-related) scales. The data sources undergo AR(1) data imputation, R(1) spatial and temporal calibration, and are then assimilated with LSDA.

To estimate sequential data points while considering previous data, the RLS-based AR(1) model for sequential estimation is utilized. **S-DA** applies LSDA to assimilate a new observation that has been processed through AR(1) for data imputation and uncertainty estimation with the previously acquired **S-DA** analysis result. After each LSDA procedure run, another RLS-based AR(1) model fits the relationship between the current analysis result and the previous one, making a refined prediction for LSDA with a new observation in the subsequent step. Compared to **S-DA** updating the coefficients using incoming data, **S-DA4** uses the output from **DA4** for sequential estimation.

5.4 Results

The developments in addressing **RQ2** led directly to the exploration of **RQ3**: evaluating the performance of the algorithms in refining the accuracy of urban air pollution estimates from European cities. The aim was to ascertain whether the algorithms stemming from **RQ2** could contribute to more accurate AQ monitoring, which is imperative for both public health and regulatory compliance.

In addition to the validation and comparison of algorithms **DA2** and **DA3** in publication II, the study III validated and compared the performance of algorithms **S-DA**, **DA3**, **DA4**, and **S-DA4** using data from European AQ monitoring stations (EEA AQ dataset). The results of **S-DA** were compared with **DA3**, and **DA4** results were compared with **S-DA4**. For each AQ variable, results were averaged across all stations. Performance was evaluated and compared using the ratios of two metrics: RMSE (see Eq.6) and MAU (see Eq.7), derived by dividing the metrics of one DA method by those of the other. The examples of the validation scenarios for one of the AQ variables (O3 - ozone) are presented in Table 9.

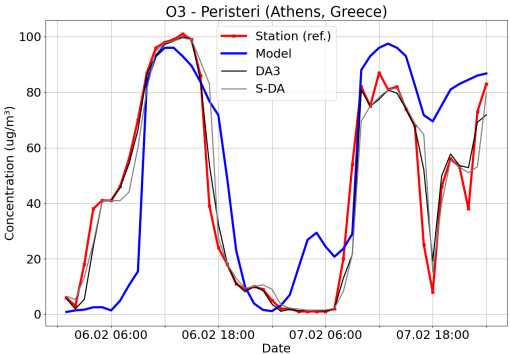
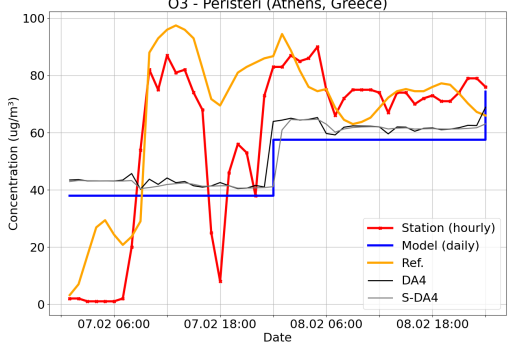
Visual example of a validation scenario	Performance metrics
<p style="text-align: center;">DA3 vs S-DA:</p>  <p style="text-align: center;">O3 - Peristeri (Athens, Greece)</p> <p>Validation scenario for the DA3 and S-DA algorithms over the period from February 6, 2022, to February 8, 2022, in assimilating hourly ozone concentrations at Peristeri, Athens, Greece. "Station" corresponds to x_{obs}, "Model" - x_m, "DA3" - $x_{a(DA3)}$, "S-DA" - $x_{a(S-DA)}$. x_{obs} is used as the reference for comparison; derived from publication III.</p>	<p style="text-align: center;">Accuracy metrics:</p> $r_{RMSE} = \frac{RMSE(x_{a(S-DA)}; x_{obs})}{RMSE(x_{a(DA3)}; x_{obs})}$ <p style="text-align: center;">Precision metrics:</p> $r_{MAU} = \frac{MAU(\epsilon_{a(S-DA)})}{MAU(\epsilon_{a(DA3)})}$ <p>For the visual example: $r_{RMSE} = 1.406$, $r_{MAU} = 0.885$ (DA3 is 40.6% more accurate and 11.5% less precise than S-DA)</p> <p>Averaged for N=462 stations from January 27, 2022, to February 25, 2022 (from III): $r_{RMSE} = 1.367$, $r_{MAU} = 0.902$ (DA3 is 36.7% more accurate and 9.8% less precise than S-DA)</p>
<p style="text-align: center;">DA4 vs S-DA4:</p>  <p style="text-align: center;">O3 - Peristeri (Athens, Greece)</p> <p>Validation scenario for the DA4 and S-DA4 algorithms over the period from February 6, 2022, to February 8, 2022, in assimilating hourly and daily ozone concentrations at Peristeri, Athens, Greece. "Station" corresponds to x_{obs}^h, "Model" - x_m^d, "Ref." - x_m^h, "DA4" - $x_{a(DA4)}$, "S-DA4" - $x_{a(S-DA4)}$; derived from publication III.</p>	<p style="text-align: center;">Accuracy metrics:</p> $r_{RMSE}^{d \rightarrow h} = \frac{RMSE(x_a; x_m^h)}{RMSE(x_m^d; x_m^h)}$ <p style="text-align: center;">Precision metrics:</p> $r_{MAU} = \frac{MAU(\epsilon_{a(S-DA4)})}{MAU(\epsilon_{a(DA4)})}$ <p>For the visual example: DA4: $r_{RMSE}^{d \rightarrow h} = 0.922$ S-DA4: $r_{RMSE}^{d \rightarrow h} = 0.950$ (DA4 is 7.8% and S-DA4 is 5.0% more accurate than flat-line extrapolated daily values; DA4 is 2.8% more accurate than S-DA4) $r_{MAU} = 0.582$ (DA4 is 41.8% less precise than S-DA4)</p> <p>Averaged for N=462 stations from January 27, 2022, to February 25, 2022 (from III): DA4: $r_{RMSE}^{d \rightarrow h} = 0.868$ S-DA4: $r_{RMSE}^{d \rightarrow h} = 0.885$ (DA4 is 13.2% and S-DA4 is 11.5% more accurate than flat-line extrapolated daily values; DA4 is 1.7% more accurate than S-DA4) $r_{MAU} = 0.614$ (DA4 is 38.6% less precise than S-DA4)</p>

Table 9 – Examples of comparative analysis of DA3 vs S-DA and DA4 vs S-DA4 algorithms for hourly ozone concentrations in Peristeri (Athens, Greece). Performance metrics assess accuracy and precision against reference values. Data source: publication III.

The comparative analysis of **S-DA** and **DA3** utilized hourly observations from the EEA AQ dataset, combined with hourly SILAM simulation data. Reference data for algorithms **S-DA** and **DA3** was based on observations from the monitoring stations. For all the AQ variables examined, **S-DA** proved to be less accurate yet more precise than **DA3**. The performance of the **DA4** and **S-DA4** algorithms was validated by changing one of the hourly data sources to daily (e.g. model data x_m). The assimilation results of the daily data x_m^d with the hourly data from another source x_{obs}^h are then compared with the initial reference hourly data x_m^h .

While the results in terms of accuracy for **DA4** and **S-DA4** vary: **DA4** was more accurate for PM2.5, NO2, O3 and PM10, and **S-DA4** was more accurate for CO and SO2. For all the variables, **S-DA4** is more precise than **DA4** due to the sequential estimation loop, reducing uncertainty through another round of DA. The results for all the AQ variables are summarized in Table 2 (for **S-DA** and **DA3**) and Table 3 (for **DA4** and **S-DA4**) of the original manuscript III.

5.5 Conclusion

The study offers a thorough evaluation of the performance of all proposed DA methods in assimilating open AQ data sources, including EEA AQ monitoring station data and SILAM's large-scale model simulations, which vary in temporal and spatial scales and lack input uncertainty estimates.

Contributions **C3.1** and **C3.2** are addressed in the study based on publication III. **C3.1** covers the introduction and implementation of lightweight, data-driven preprocessing and assimilation methods (**S-DA**, **DA4**, **S-DA4**). **C3.2** focuses on the performance validation of these algorithms using pan-European urban AQ monitoring station data (OGD) to improve data quality across scales. The introduced algorithms **S-DA**, **DA4** and **S-DA4** further facilitate answering **RQ2**:

RQ2: "Can data assimilation be applied at different spatial and temporal scales using sources without uncertainty?"

Answer: The study addresses **RQ2** by introducing novel DA algorithms such as **S-DA**, **DA4**, and **S-DA4** (in addition to **DA2** and **DA3** from publication II), which are designed to preprocess and assimilate urban AQ data across varying temporal and spatial scales without the need for prior uncertainty estimates. In particular, **S-DA** is designed to perform DA using one data source with unknown uncertainties, **DA4** and **S-DA4** - using the data sources of both varying temporal and spatial scales (compared to **DA3**). These methods improve the accuracy by utilizing a data-driven approach to estimate uncertainties, which is crucial for effective assimilation under the given constraints.

S-DA vs **DA3** and **DA4** vs **S-DA4** are rigorously compared. **S-DA** is found to exhibit a higher error but lower uncertainty than **DA3**. Moreover, **DA4** and **S-DA4** show similar performance levels, with **DA4** noted for its computational efficiency and **S-DA4** for improved accuracy under specific conditions, especially during rapid data changes. These findings underscore the feasibility of applying DA methods to diverse urban AQ data sets, paving the way for more sophisticated environmental monitoring and analysis.

The results from publication III affirmatively answer **RQ3**:

RQ3: "Are computationally lightweight assimilation methods suitable for large-scale open environmental monitoring data?"

Answer: The findings from publication III positively respond to **RQ3**, demonstrating the suitability of lightweight DA methods for large-scale environmental monitoring, specifically in the context of European urban AQ data. The validation confirms that the DA methods can be applied to data of varying scales, effectively improving the accuracy of the final estimate. It demonstrates that the algorithms can enhance the accuracy of open European urban AQ estimates without the need for in-depth knowledge of the models, calibration operators, or prior uncertainty estimates. The introduced methods successfully address missing data, ensuring dataset completeness — a crucial aspect of environmental monitoring. RMSE metrics are used to assess accuracy by measuring the differences between the assimilation outputs and the reference data. The choice of a specific DA method is informed by several factors: the spatial and temporal scales of the input data, the availability (or lack) of a priori uncertainty information, and the need to accommodate rapid variations in environmental data. This versatility and adaptability underline the algorithms' potential to support robust, data-driven environmental monitoring and decision-making on a large scale.

Future research will build on this foundation, targeting the creation of dynamic maps, optimization of urban AQ monitoring placements, and the inclusion of a broader range of data sources.

6 Conclusions

The primary objectives of this dissertation were to address three RQs to improve data completeness, accuracy, and precision in environmental monitoring. This was accomplished by developing lightweight assimilation algorithms using OGD from varied spatial and temporal scales without a priori uncertainty estimates. The efficacy of the proposed algorithms was assessed using Estonian local and pan-European urban air pollution estimates. This work, consisting of publications **I**, **II** and **III** provides compelling answers to these questions through the development and validation of novel DA algorithms and methodologies.

The work aimed to develop a general approach to estimating uncertainty based solely on the input data values, without reference to the model's structure or the source of observations. The SILAM model was selected since it has publicly available access to data archives, and this does not limit the algorithms' applicability to other models or domains.

The proposed DA methods show promise for transforming AQ monitoring using open data, affordable sensors, and numerical models, which could substantially reduce the need for new infrastructure. Their adaptable and scalable design, supported by open-source code, paves the way for further research and innovation. Emphasizing the importance of developing "lightweight" methods suitable for IoT devices, the study addresses the challenge of unknown uncertainties by effectively using regression-based uncertainties. To juxtapose the research outputs of this dissertation with the current state of the art in academia, industry, and the public sector, the following practices in DA and environmental monitoring can be noted to highlight the specific advances made in this dissertation:

Academia: In academia, the utilization of DA methods has been traditionally focused on complex and often computationally intensive models, such as those described by [27] in their comprehensive review of DA methods for numerical weather prediction. The advances in modeling techniques are driving the future of DA towards more complex, non-linear methods and a growing integration of machine learning, though it faces challenges in estimating uncertainty [148]. The lightweight DA algorithms developed in this dissertation, such as **DA2**, **DA3**, **DA4**, **S-DA**, and **S-DA4**, represent a significant departure from this norm, offering simpler, more agile alternatives that are suitable for continuous applications and IoT devices and including automatic uncertainty estimation.

Industry: In the industry, environmental monitoring often relies on proprietary technologies and sensor networks with integrated data processing capabilities. Companies like IBM and their Green Horizons initiative use large-scale sensor data to predict AQ [53]. The methods introduced in this dissertation provide a cost-effective, open-source alternative that can be integrated with existing IoT infrastructures, democratizing access to advanced monitoring capabilities.

Public sector: The public sector, particularly environmental agencies like the EEA, has been increasingly adopting open data policies to enhance environmental monitoring. The dissertation's approach aligns with these policies by leveraging open data from EEA and other sources, improving upon the traditional methods that often require detailed model knowledge or proprietary data [54].

These points underscore the novelty and applicability of the research presented in this dissertation, placing it at the forefront of current environmental monitoring practices. Notably, the outcomes of this work are expected to impact not only academic research, but are likely to have substantial applications in industry and in the public sector. This is due to the ability of the proposed DA algorithms to be integrated easily into existing monitoring systems, providing enhanced data quality for ongoing research and development in the field. Furthermore, this dissertation lays the groundwork for future research directions:

- expanding the application of the proposed algorithms to dynamic air pollution mapping;
- formulating methodologies for optimal sensor placements based on uncertainties;
- validating the proposed algorithms on different types of environmental data sources;
- extending support for different data types and additional input parameters;
- implementing these methods on mesh networks of low-cost IoT devices to enhance system-wide consistency, reliability, and accuracy.

In summary, this dissertation advances the field of environmental monitoring and compliance reporting by introducing cost-effective, data-agnostic, and scalable DA algorithms that improve data quality on multiple fronts. These algorithms are particularly effective for assimilating disparate data sources and are validated to improve the reliability of open European urban air pollution estimates. Open-source code and validation scenarios are provided to ensure reproducibility and to encourage future research in this vital area. By addressing the challenges associated with assimilating multi-source, multi-scale environmental data, this work offers major insights and practical answers for adapting to a future where precise, timely, and extensive environmental monitoring is urgently required.

List of Figures

1	AERCM lies at the nexus of relationships needed to address the ambitious goals of EU regulatory frameworks associated with environmental monitoring and compliance and can be visualized as sets of concurrent and inter-related agendas, with the common goals of improved transparency, and standardization facilitating public participation and collaboration.....	10
2	Schematic representation of the OEDA process: Assimilating physical observations (top right, blue) and numerical model data (top left, red) from open data sources and with different errors and uncertainty bounds, to produce an analysis estimate of better quality and optimized uncertainty (bottom, green).....	12
3	Representation of the AERCM concept including OEDA, which encompasses open data collection and data processing. AERCM includes the additional stages of compliance estimation and reporting, ultimately the reports are used for decision- and policy-making.....	13
4	Visual representation of the EIC methodology and its corresponding IoT architecture for the AERCM system for river eflows, adapted from publication I.	22
5	Plot of eflows compliance indicating daily water discharge and temperatures with environmental threshold markers for the Aesoo station in 2016, generated by EFCES as developed in publication I.	25
6	Map showing the Liivalaia AQ monitoring station within the 0.2° SILAM model grid cell in Tallinn, Estonia, as used in publication II. The fixed-point station is set against the backdrop of the large-scale SILAM grid, underscoring the contrast in spatial scales. The SILAM grid was retrieved following [142].	28
7	Visual representation of the introduced DA1 , DA2 and DA3 algorithms, derived from publication II.	31
8	Visual representation of the introduced S-DA , DA4 and S-DA4 algorithms, derived from publication III.	36

List of Tables

1	Publications (Publ.) and corresponding contributions to each of the three RQs. The open data and code repository are provided for each publication via the GitHub link provided in the reference.....	15
2	Map, sources, variables, and time periods of the river data used in publication I.	22
3	Eflows compliance summary for three Estonian rivers in 2009-2018 (summarized from publication I).....	24
4	Overview of missing and inconsistent data in Estonian hydrological datasets: river data obtained from the government on request and OGD for the period from January 01, 2009, to December 31, 2018. Summarized from publication I.	25
5	Observation sources, variables, and time periods of the air pollution data used in publication II from assimilation with the SILAM data. CO - carbon monoxide, NO ₂ - nitrogen dioxide, O ₃ - ozone, SO ₂ - sulfur dioxide, PM _{2.5} - particulate matter $\leq 2.5 \mu\text{m}$, PM ₁₀ - particulate matter $\leq 10 \mu\text{m}$	28
6	Examples of comparative analysis of DA ₂ vs DA ₃ algorithms for hourly PM ₁₀ concentrations in Liivalaia (Tallinn, Estonia). Performance metrics assess accuracy and precision against the reference sensor values. Data source: II.	32
7	Map and variables of air pollutants from European monitoring stations, the data from which were used in publication III.	35
8	Average execution time of 1 iteration of the proposed DA algorithms. The execution time was estimated using a computer with an Intel(R) Core(TM) i7-8565U CPU @1.80GHz \times 8 and 16Gb RAM.....	37
9	Examples of comparative analysis of DA ₃ vs S-DA and DA ₄ vs S-DA ₄ algorithms for hourly ozone concentrations in Peristeri (Athens, Greece). Performance metrics assess accuracy and precision against reference values. Data source: publication III.....	38

References

- [1] Lizaveta Miasayedava, Keegan McBride, and Jeffrey A. Tuhtan. Automated Environmental Compliance Monitoring of Rivers with IoT and Open Government Data. *Journal of Environmental Management*, 303:114283, February 2022.
- [2] Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty. *Environmental Modeling & Assessment*, June 2023.
- [3] Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Open Data Assimilation of Pan-European Urban Air Quality. *IEEE Access*, 11:84670–84688, August 2023.
- [4] Jeffrey A. Tuhtan, Elizaveta Dubrovinskaya, Lizaveta Miasayedava, Vishwajeet Pattana-ik, Jürgen Soom, Bernd Mockenhaupt, Cornelia Schütz, Christian Haas, and Philipp Thumser. Smart Fish Counter for Monitoring Species, Size, Migration Behaviour and Environmental Conditions. In *The 2022 International Symposium on Ecohydraulics*, pages 1–4, 2022.
- [5] Jeffrey A Tuhtan, Lizaveta Miasayedava, and Gert Toming. Data Assimilation of Acoustic Doppler Velocimeter and Total Pressure Sensors. Presentation at the 40th IAHR World Congress, 2023.
- [6] European Commission. Horizon Europe Work Programme 2023-2024: Health. Work programme, European Commission, 3 2023. European Commission Decision C(2023) 2178 of 31 March 2023.
- [7] European Commission. Delivering the European Green Deal. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/european-green-deal/delivering-european-green-deal_en. [Accessed 25-09-2023].
- [8] European Commission. Pathway to a Healthy Planet for All: EU Action Plan: 'Towards Zero Pollution for Air, Water and Soil'. Communication to the European Parliament, The Council, The European Economic and Social Committee and The Committee of the Regions, 5 2021. COM(2021) 400 final, SWD(2021) 140 final - SWD(2021) 141 final.
- [9] European Parliament and Council of the European Union. Directive 2014/52/EU of the European Parliament and of the Council of 16 April 2014 amending Directive 2011/92/EU on the assessment of the effects of certain public and private projects on the environment. Official Journal of the European Union, 4 2014.
- [10] European Commission. Communication from the commission: Guidelines on non-financial reporting: Supplement on reporting climate-related information. Official Communication, 6 2019. C(2019) 4490 final.
- [11] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation), 2016. OJ L 119, 4.5.2016, p. 1–88.

- [12] European Commission. Open access - H2020 Online Manual. https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm. [Accessed 25-09-2023].
- [13] David A. Kroodsma, Juan Mayorga, Timothy Hochberg, Nathan A. Miller, Kristina Boerder, Francesco Ferretti, Alex Wilson, Bjorn Bergman, Timothy D. White, Barbara A. Block, Paul Woods, Brian Sullivan, Christopher Costello, and Boris Worm. Tracking the global footprint of fisheries. *Science*, 359:904 – 908, 2018.
- [14] Tarek Rashed and Carsten Jürgens. Remote Sensing of Urban and Suburban Areas. volume 10 of *Remote Sensing and Digital Image Processing*, Dordrecht, 2010. Springer. Corpus ID: 127796208.
- [15] Gitaek Lee, Seonghyeon Moon, Jaehyun Hwang, and Seokho Chi. Development of a real-time noise estimation model for construction sites. *Advanced Engineering Informatics*, 58:102133, 2023.
- [16] Giacomo Grassi, Clemens Schwingshackl, Thomas Gasser, Richard A. Houghton, Stephen A. Sitch, Josep G. Canadell, Alessandro Cescatti, Philippe Ciais, Sandro Federici, Pierre Friedlingstein, Werner A. Kurz, Maria J. Sanz Sanchez, Raúl Abad Viñas, Ramdane Alkama, Selma Bultan, Guido Ceccherini, Stefanie Falk, Etsushi Kato, Daniel Kennedy, Jürgen Knauer, Anu Korosuo, Joana Melo, Matthew J. McGrath, Julia Nabel, Benjamin I Poulter, Anna A. Romanovskaya, Simone Rossi, Hanqin Tian, Anthony P. Walker, Wenping Yuan, Xu Yue, and Julia Pongratz. Harmonising the land-use flux estimates of global models and national inventories for 2000–2020. *Earth System Science Data*, 2023.
- [17] Miranti Indri Hastuti, Ki-Hong Min, and Ji-Won Lee. Improving radar data assimilation forecast using advanced remote sensing data. *Remote Sensing*, 15:2760, 2023.
- [18] Marc Bocquet, Hendrik Elbern, Henk J. Eskes, Marcus Hirtl, Rahela Žabkar, Gregory R. Carmichael, Johannes Flemming, Antje Inness, Mariusz Pagowski, Juan Luis Pérez Camaño, Pablo E. Saide, Roberto San José, Mikhail Sofiev, Julius Vira, Alexander A. Baklanov, Claudio Carnevale, Georg Grell, and Christian Seigneur. Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models. *Atmospheric Chemistry and Physics*, 15:5325–5358, 2014.
- [19] Arianna Valmassoi, Jan Keller, Daryl T. Kleist, Stephen J. English, Bodo Ahrens, Ivan Bašták Ďurán, Elisabeth Bauernschubert, Michael G. Bosilovich, Masatomo Fujiwara, Hans Hersbach, Lili Lei, Ulrich Löhnert, Nabir Mamnun, Cory R. Martin, Andrew M. Moore, Deborah Niermann, Juan Jose Ruiz, and Leonhard Scheck. Current challenges and future directions in data assimilation and reanalysis. *Bulletin of the American Meteorological Society*, 2022.
- [20] Sipeng Shen. Editorial: Integrative approaches to analyze cancer based on multi-omics. *Frontiers in Genetics*, 13, 2022.
- [21] Huang Ding, Fang Cui, Zhi jia Wang, H. Zhou, and Wei dong Chen. A comparison study of the application of data assimilation in the short-term prediction of radiation and precipitation. *IOP Conference Series: Earth and Environmental Science*, 136, 2018.

- [22] Marc E. Ridler, Nils van Velzen, Stef Hummel, Inge Sandholt, Anne Katrine Vinther Falk, Arnold W. Heemink, and Henrik Madsen. Data assimilation framework: Linking an open data assimilation library (openda) to a widely adopted model interface (openmi). *Environmental Modelling & Software*, 57:76–89, 2014.
- [23] Hamze Dokoohaki, Bailey D Morrison, Ann M. Raiho, Shawn Paul Serbin, Katie Zarada, Luke Dramko, and Michael Dietze. Development of an open-source regional data assimilation system in pecan v. 1.7.2: application to carbon cycle reanalysis across the contiguous us using sipnet. *Geoscientific Model Development*, 2022.
- [24] Carsten Montzka, Valentijn R. N. Pauwels, Harrie-Jan Hendricks-Franssen, Xujun Han, and Harry Vereecken. Multivariate and multiscale data assimilation in terrestrial systems: A review. *Sensors (Basel, Switzerland)*, 12:16291 – 16333, 2012.
- [25] Jude Lubega Musuuza, David Gustafsson, Rafael Pimentel, Louise Crochemore, and Ilias G. Pechlivanidis. Impact of satellite and in situ data assimilation on hydrological predictions. *Remote Sensing*, 12:811, 2020.
- [26] Ghada Y. El Serafy, Martin Verlaan, Stef Hummel, Albrecht H. Weerts, and Juzer F. Dhondia. OpenDA Open Source Generic Data Assimilation Environment and its Application in Process Models. 2010.
- [27] A. Carrassi, M. Bocquet, L. Bertino, and G. Evensen. Data assimilation in the geosciences: An overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5):e535, 2018.
- [28] Kyung Hwa Cho, Yakov A. Pachepsky, Mayzonee Ligaray, Yong Sung Kwon, and Kyunghyun Kim. Data assimilation in surface water quality modeling: A review. *Water research*, 186:116307, 2020.
- [29] Francesco Avanzi, Simone Gabellani, Fabio Delogu, Francesco Silvestro, Edoardo Cremonese, Umberto Morra di Cella, Sara Maria Ratto, and Hervé Stevenin. S3m 5.1: a distributed cryospheric model with dry and wet snow, data assimilation, glacier mass balance, and debris-driven melt. *Geoscientific Model Development*, 2021.
- [30] Yuqiong Liu and Hoshin V. Gupta. Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework. *Water Resources Research*, 43, 2007.
- [31] William A. Lahoz and Philipp Schneider. Data assimilation: making sense of earth observation. *Frontiers in Environmental Science*, 2, 2014.
- [32] Xin Huang, Dan Lu, Daniel M. Ricciuto, Paul J. Hanson, Andrew D. Richardson, Xuehe Lu, Ensheng Weng, Sheng Nie, Lifen Jiang, Enqing Hou, Igor Steinmacher, and Yiqi Luo. A Model-Independent Data Assimilation (MIDA) module and its applications in ecology. *Geoscientific Model Development*, 2021.
- [33] Lizaveta Miasayedava. EEFlows: A Django/React web application for environmental flows estimation. <https://github.com/effie-ms/eeflows>, 2021.
- [34] Lizaveta Miasayedava. RLS-based data assimilation with unknown uncertainty. <https://github.com/effie-ms/rls-assimilation>, 2023.
- [35] Hakan Alphan. Analysis of landscape changes as an indicator for environmental monitoring. *Environmental Monitoring and Assessment*, 189:1–10, 2016.

- [36] Roger H. Green. Statistical and nonstatistical considerations for environmental monitoring studies. *Environmental Monitoring and Assessment*, 4:293–301, 1984.
- [37] Gary M Lovett, Douglas A. Burns, Charles T. Driscoll, Jennifer C. Jenkins, Myron J. Mitchell, Lindsey E Rustad, James B. Shanley, Gene E. Likens, and Richard A. Haeuber. Who needs environmental monitoring. *Frontiers in Ecology and the Environment*, 5:253–260, 2007.
- [38] Eric Biber. The problem of environmental monitoring. *University of Colorado Law Review*, 83:1, 2010.
- [39] José M. Barceló-Ordinas, Messaoud Doudou, Jorge García-Vidal, and Nadjib Badache. Self-calibration methods for uncontrolled environments in sensor networks: A reference survey. *Ad Hoc Networks*, 88:142–159, 2019.
- [40] Gonçalo de Jesus, António Casimiro, and Anabela Oliveira. A survey on data quality for dependable monitoring in wireless sensor networks. *Sensors (Basel, Switzerland)*, 17, 2017.
- [41] Richard Gliklich, Nancy A. Dreyer, and Michelle B. Leavy. Data collection and quality assurance. 2014.
- [42] George Dan Mois, Silviu Corneliu Folea, and Teodora Sanislav. Analysis of three iot-based wireless sensors for environmental monitoring. *IEEE Transactions on Instrumentation and Measurement*, 66:2056–2064, 2017.
- [43] Shi Lan, Miao Qilong, and Jinglin Du. Architecture of wireless sensor networks for environmental monitoring. *2008 International Workshop on Education Technology and Training & 2008 International Workshop on Geoscience and Remote Sensing*, 1:579–582, 2008.
- [44] Shawky Mansour, Mohammed Ali K. Al-Belushi, and Talal Al-Awadhi. Monitoring land use and land cover changes in the mountainous cities of oman using gis and ca-markov modelling techniques. *Land Use Policy*, 91:104414, 2020.
- [45] Keith L. Downing. Using evolutionary computational techniques in environmental modelling. *Environmental Modelling and Software*, 13:519–528, 1998.
- [46] Neil D. Bennett, Barry F. W. Croke, Giorgio Guariso, Joseph H. A. Guillaume, Serena H. Hamilton, Anthony J. Jakeman, Stefano Marsili-Libelli, Lachlan T. H. Newham, John P. Norton, Charles Perrin, Suzanne A. Pierce, Barbara J. Robson, Ralf Seppelt, Alexey A. Voinov, Brian D. Fath, and Vazken Andréassian. Characterising performance of environmental models. *Environmental Modelling & Software*, 40:1–20, 2013.
- [47] Adrian Sandu and Tianfeng Chai. Chemical data assimilation—an overview. *Atmosphere*, 2:426–463, 2011.
- [48] Vladimir Penenko. Variational methods of data assimilation and inverse problems for studying the atmosphere, ocean, and environment. *Numerical Analysis and Applications*, 2:341–351, 2009.

- [49] Amy McNally, Kristi R. Arsenault, Sujay V. Kumar, Shraddhanand Shukla, Pete Peterson, Shugong Wang, Chris Funk, Christa D. Peters-Lidard, and James P. Verdin. A land data assimilation system for sub-saharan africa food and water security applications. *Scientific Data*, 4, 2017.
- [50] Mark Asch, Marc Bocquet, and Maëlle Nodet. *Data Assimilation: Methods, Algorithms, and Applications*. Society for Industrial and Applied Mathematics, 2016.
- [51] Wolfgang Kurtz, Andrei Lapin, Oliver S. Schilling, Qi Tang, Eryk Schiller, Torsten Braun, Daniel Hunkeler, Harry Vereecken, Edward A. Sudicky, Peter G. Kropf, Harrie-Jan Hendricks-Franssen, and Philip Brunner. Integrating hydrological modelling, data assimilation and cloud computing for real-time management of water resources. *Environmental Modelling & Software*, 93:418–435, 2017.
- [52] Finnish Meteorological Institute. SILAM v.5.7: System for Integrated modelLing of Atmospheric coMposition. <http://silam.fmi.fi/>, 2021.
- [53] IBM Research. IBM Research Green Horizons, 2015. <https://www.research.ibm.com/labs/china/greenhorizons/>.
- [54] European Environment Agency. The european environment — state and outlook 2020: Knowledge for transition to a sustainable europe, 2020. <https://www.eea.europa.eu/soer/2020>.
- [55] Judie Attard, Fabrizio Orlandi, Simon Scerri, and S. Auer. A systematic review of open government data initiatives. *Government Information Quarterly*, 32:399–418, 2015.
- [56] Teresa Scassa. Privacy and open government. *Future Internet*, 6:397–413, 2014.
- [57] Mila Gascó-Hernández, Erika G. Martin, Luigi Reggi, Sunyoung Pyo, and Luis Felipe Luna-Reyes. Promoting the use of open government data: Cases of training and engagement. *Government Information Quarterly*, 35:233–242, 2018.
- [58] Thorhildur Jetzek, Michel Avital, and Niels Bjørn-Andersen. Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, 9:100–120, 2014.
- [59] Mauricio Solar, Gastón Concha, and Luis Meijueiro. A model to assess open government data in public agencies. In *International Conference on Electronic Government*, 2012.
- [60] Anneke Zuiderwijk and Marijn Janssen. The negative effects of open government data - investigating the dark side of open data. *Proceedings of the 15th Annual International Conference on Digital Government Research*, 2014.
- [61] Barbara Ubaldi. Open government data: Towards empirical analysis of open government data initiatives. 2013.
- [62] Jan Kučera, Dušan Chlapek, and Martin Nečaský. Open government data catalogs: Current approaches and quality perspective. In *EGOVIS/EDEM*, 2013.
- [63] Aikaterini Yannoukakou and Iliana Araka. Access to government information: Right to information and open government data synergy. *Procedia - Social and Behavioral Sciences*, 147:332–340, 2014.

- [64] Omer Hassan Abdelrahman. Open government data: Development, practice, and challenges. In *Open Data*, chapter 2. IntechOpen, Rijeka, 2021.
- [65] Natasa Z. Veljkovic, Sanja Bogdanovic-Dinic, and Leonid Stoimenov. Benchmarking open government: An open data perspective. *Government Information Quarterly*, 31:278–290, 2014.
- [66] Tim Davies and Zainab Ashraf Bawa. The Promises and Perils of Open Government Data (OGD). *Journal of Community Informatics*, 8, 2012.
- [67] Marijn Janssen and Jeroen van den Hoven. Big and open linked data (bold) in government: A challenge to transparency and privacy? *Government Information Quarterly*, 32:363–368, 2015.
- [68] Hui-Ju Wang and Jin Lo. Adoption of open government data among government agencies. *Government Information Quarterly*, 33:80–88, 2016.
- [69] European Parliament and Council of the European Union. Directive (EU) 2019/1024 of the European Parliament and of the Council of 20 June 2019 on open data and the re-use of public sector information (recast), June 26 2019.
- [70] A. Nikiforova, N. Rizun, M. Ciesielska, C. Alexopoulos, and A. Miletic. Towards high-value datasets determination for data-driven development: A systematic literature review. In I. Lindgren et al., editors, *Electronic Government*, volume 14130 of *Lecture Notes in Computer Science*, Cham, 2023. Springer.
- [71] Tung-Mou Yang, Jin Lo, and Jing Shiang. To open or not to open? determinants of open government data. *Journal of Information Science*, 41:596 – 612, 2015.
- [72] Anneke Zuiderwijk and Marijn Janssen. Open data policies, their implementation and impact: A framework for comparison. *Government Information Quarterly*, 31:17–29, 2014.
- [73] Luigi Reggi and Sharon S. Dawes. Open government data ecosystems: Linking transparency for innovation with transparency for participation and accountability. In *International Conference on Electronic Government*, 2016.
- [74] Alejandro Sáez-Martín, Arturo Rosario, and María Pérez. An international analysis of the quality of open government data portals. *Social Science Computer Review*, 34:298 – 311, 2016.
- [75] Bernd W. Wirtz, Jan C. Weyerer, and Michael Rösch. Open government and citizen participation: an empirical analysis of citizen expectancy towards open government data. *International Review of Administrative Sciences*, 85:566 – 586, 2019.
- [76] Soren Hansen and Stamatis Fradelos. An integrated vessel performance system for environmental compliance. In *Trends and Challenges in Maritime Energy Management*, pages 185–198. Springer International Publishing, 2018.
- [77] Greg Haunschild and Suzanne Shelley. Environmental management: A better approach. *Chemical Engineering*, 112:44–49, 2005.
- [78] Ilhang Shin, Myung-gun Lee, and Woojin Park. Implementation of the continuous auditing system in the erp-based environment. *Managerial Auditing Journal*, 28:592–627, 2013.

- [79] Marianna Sigala. A supply chain management approach for investigating the role of tour operators on sustainable tourism: the case of TUI. *Journal of Cleaner Production*, 16:1589–1599, 2008.
- [80] Larissa S. May, Jean-Paul Chretien, and Julie A. Pavlin. Beyond traditional surveillance: applying syndromic surveillance to developing settings – opportunities and challenges. *BMC Public Health*, 9:242 – 242, 2009.
- [81] Richard Lee Jenks. Mastering environmental tasks using compliance-management tools. *Chemical Engineering*, 109:83–86, 2002.
- [82] Thorben Sandner, Matthias Kehlenbeck, and Michael H. Breitner. Visualization of automated compliance monitoring and reporting. *2010 Workshops on Database and Expert Systems Applications*, pages 364–368, 2010.
- [83] Thomas H. Beach, Jean-Laurent Hippolyte, and Yacine Rezgui. Towards the adoption of automated regulatory compliance checking in the built environment. *Automation in Construction*, 118:103285, 2020.
- [84] Heiko Henning Thimm. ICT Support of Environmental Compliance - Approaches and Future Perspectives. In *International Conference on Informatics for Environmental Protection*, 2016.
- [85] Matthias Kehlenbeck, Thorben Sandner, and Michael H. Breitner. Application and economic implications of an automated requirement-oriented and standard-based compliance monitoring and reporting prototype. *2010 International Conference on Availability, Reliability and Security*, pages 468–474, 2010.
- [86] Terence M. Yhip and Bijan M. D. Alagheband. Credit monitoring and compliance. In *The Practice of Lending*, pages 421–430, Springer Books, 2020. Springer.
- [87] Fredda E. Ackerman, R. Darrell Mounts, and David U. Thomas. Automated compliance systems. *Journal of Financial Regulation and Compliance*, 7:48–56, 1999.
- [88] Camillo Silibello, Andrea Bolignano, R. Sozzi, and Claudio Gariazzo. Application of a chemical transport model and optimized data assimilation methods to improve air quality assessment. *Air Quality, Atmosphere & Health*, 7:283–296, 2014.
- [89] Jia Xing, Siwei Li, Dian Ding, James T. Kelly, Shuxiao Wang, Carey J. Jang, Yun Zhu, and Jiming Hao. Data assimilation of ambient concentrations of multiple air pollutants using an emission-concentration response modeling framework. *Atmosphere*, 11, 2020.
- [90] Olga Lucía Quintero Montoya, Elías D. Nino-Ruiz, and Nicolas Pinel. On the mathematical modelling and data assimilation for air pollution assessment in the tropical andes. *Environmental Science and Pollution Research*, pages 1–20, 2020.
- [91] Caterina Buizza, César Quilodrán Casas, Philip Nadler, Julian Mack, Stefano Marrone, Zainab Titus, Clémence Le Cornec, Evelyn Heylen, Tolga Hasan Dur, Luis Baca Ruiz, Claire E. Heaney, Julio Amador Díaz López, K. S. Sesh Kumar, and Rossella Arcucci. Data Learning: Integrating Data Assimilation and Machine Learning. *Journal of Computer Science*, 58:101525, 2021.

- [92] Chi Vuong Nguyen, Lionel Soulhac, and Pietro Salizzoni. Source Apportionment and Data Assimilation in Urban Air Quality Modelling for NO₂: The Lyon Case Study. *Atmosphere*, 9:8, 2018.
- [93] Rolf H. Reichle. Data assimilation methods in the earth sciences. *Advances in Water Resources*, 31:1411–1418, 2008.
- [94] Tapio Schneider et al. Harnessing AI and computing to advance climate modelling and prediction. *Nature Climate Change*, 13:789–795, 2023.
- [95] Nicki Hickmon et al. Artificial intelligence for earth system predictability (ai4esp): Revolutionizing earth system science with ai technologies. *Environmental Research Letters*, 16(4):043010, 2021.
- [96] Hayden Dahmm. Major environmental data gaps remain, but progress is on the horizon. 2021. Accessed: 2023-11-03.
- [97] United Nations Environment Programme. Global environment outlook. 2019. Accessed: 2023-11-03.
- [98] Jennifer K. Whyte, Ana Mijić, Rupert J. Myers, Panagiotis Angeloudis, Michel-Alexandre Cardin, M. Stettler, and Washington Yotto Ochieng. A research agenda on systems approaches to infrastructure. *Civil Engineering and Environmental Systems*, 37:214–233, 2020.
- [99] Jun Ying, Xiaojing Zhang, Yiqi Zhang, and Svitlana Bilan. Green infrastructure: systematic literature review. *Economic Research-Ekonomska Istraživanja*, 35:343 – 366, 2021.
- [100] Heather T. Gold, Cara L. McDermott, Ties Hoomans, and Todd H. Wagner. Cost data in implementation science: categories and approaches to costing. *Implementation Science : IS*, 17, 2022.
- [101] Hui Zhang and Jianying Xiao. Quality assessment framework for open government data. *The Electronic Library*, 38:209–222, 2020.
- [102] Areti Karamanou, Petros Brimos, Evangelos Kalampokis, and Konstantinos A. Tarabanis. Exploring the quality of dynamic open government data using statistical and machine learning methods. *Sensors (Basel, Switzerland)*, 22, 2022.
- [103] Alfonso Quarati. Open government data: Usage trends and metadata quality. *Journal of Information Science*, 49:887 – 910, 2021.
- [104] Dishant M. Pandya, Bhasha Vachharajani, and Rohit Srivastava. A review of data assimilation techniques: Applications in engineering and agriculture. *Materials Today: Proceedings*, 2022.
- [105] Sibó Cheng, Che Liu, Yike Guo, and Rossella Arcucci. Efficient deep data assimilation with sparse observations and time-varying sensors. *Journal of Computational Physics*, 496:112581, 2024.
- [106] Sibó Cheng, César Quilodrán Casas, Said Ouala, Alban Farchi, Che Liu, Pierre Tandeo, Ronan Fablet, Didier Lucor, Bertrand Iooss, Julien Brajard, Dunhui Xiao, Tijana Janjić, Weiping Ding, Yike Guo, Alberto Carrassi, Marc Bocquet, and Rossella Arcucci. Machine learning with data assimilation and uncertainty quantification for dynamical systems: A review. *IEEE/CAA Journal of Automatica Sinica*, 10:1361–1387, 2023.

- [107] Claire Merker, Gernot Geppert, Marco Clemens, and Felix Ament. Estimating the uncertainty of areal precipitation using data assimilation. *Tellus A: Dynamic Meteorology and Oceanography*, 71, 2019.
- [108] Renata Julita Romanowicz and Peter C. Young. Data assimilation and uncertainty analysis of environmental assessment problems - an application of stochastic transfer function and generalised likelihood uncertainty estimation techniques. *Reliability Engineering & System Safety*, 79:161-174, 2003.
- [109] Katalin Gruiz and Éva Fenyvesi. Chapter 3 In-situ and real-time measurements in water monitoring: Site Assessment and Monitoring Tools, pages 181-244. 11 2016.
- [110] Zhu Liu, Taochun Sun, Ying Yu, Piyu Ke, Zhu Deng, Chenxi Lu, Da Huo, and Xiang Ding. Real-time carbon emission accounting technology toward carbon neutrality. *Engineering*, 2022.
- [111] Veronica Constantin. Automation technology plays a crucial role in environmental sustainability. *E+T Magazine*, October 2021.
- [112] European Parliament, Council of the European Union. Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000: establishing a framework for Community action in the field of water policy, 2000.
- [113] B. Gopal. A conceptual framework for environmental flows assessment based on ecosystem services and their economic valuation. *Ecosystem Services*, 21:53-58, 2016.
- [114] C. Pahl-Wostl, A. Arthington, J. Bogardi, S. Bunn, H. Hoff, L. Lebel, E. Nikitina, M. Palmer, N. Poff, K. Richards, et al. Environmental flows and water governance: managing sustainable water uses. *Current Opinion in Environmental Sustainability*, 5:341-351, 2013.
- [115] N.L. Poff, J. Allan, M. Bain, J.R. Karr, K. Prestegard, B. Richter, R. Sparks, and J.C. Stromberg. The natural flow regime. a paradigm for river conservation and restoration. *Bioscience*, 47:769-784, 1997.
- [116] F. Montori, L. Bedogni, and L. Bononi. A collaborative internet of things architecture for smart cities and environmental monitoring. *IEEE Internet of Things Journal*, 5:592-605, 2017.
- [117] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi. Internet of things for smart cities. *IEEE Internet of Things Journal*, 1:22-32, 2014.
- [118] M. Janssen, Y. Charalabidis, and A. Zuiderwijk. Benefits, adoption barriers and myths of open data and open government. *Information Systems Management*, 29:258-268, 2012.
- [119] T. Jetzek, M. Avital, and N. Bjorn-Andersen. Data-driven innovation through open government data. *Journal of Theoretical and Applied Electronic Commerce Research*, 9:100-120, 2014.
- [120] Ministry of the Environment. Estonian weather service: historical observation data. <https://www.ilmateenistus.ee/siseveed/ajaloolised-vaatlusandmed/>, 2020.

- [121] J.T. Mathis, E. Osborne, and S. Starkweather. Collecting Environmental Intelligence in the New Arctic. *NOAA Arctic Report Card*, 2018.
- [122] K. Peffers, T. Tuunanen, M. Rothenberger, and S. Chatterjee. A design science research methodology for information systems research. *Journal of Management Information Systems*, 24:45–77, 2007.
- [123] Ministry of the Environment. Specified requirements for the expansion of a water body, environmental monitoring related to the expansion, protection of aquatic life, dam, elimination of the expansion and lowering of the water level, and methodology for determining the minimum ecological flow. <https://www.riigiteataja.ee/akt/124092014001>, 2014. Accessed: 2021-05-07.
- [124] European Parliament and Council of European Union. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe, 2008.
- [125] World Health Organization. WHO Global Air Quality Guidelines: Particulate Matter (PM_{2.5} and PM₁₀), Ozone, Nitrogen Dioxide, Sulfur Dioxide and Carbon Monoxide. World Health Organization, 2021.
- [126] R. J. Evans. Gems: An airborne system for urban environmental monitoring, 2004.
- [127] L. Weissert et al. Low-cost sensors and microscale land use regression: data fusion to resolve air quality variations with high spatial and temporal resolution. *Atmospheric Environment*, 213:285–295, 2019.
- [128] H. H. A. Cotta, V. A. Reisen, P. Bondon, and P. R. P. Filho. Identification of redundant air quality monitoring stations using robust principal component analysis. *Environmental Modeling & Assessment*, 25:521–530, 2020.
- [129] K. Ben Youssef et al. Estimation of aerosols dispersion and urban air quality evaluation over malaysia using modis satellite. *International Journal of Advanced Scientific and Technical Research*, 3:229–238, 2016.
- [130] A. Kotsev, O. Peeters, P. Smits, and M. Grothe. Building bridges: experiences and lessons learned from the implementation of inspire and e-reporting of air quality data in europe. *Earth Science Informatics*, 8:353–365, 2014.
- [131] Joint Committee for Guides in Metrology. Evaluation of measurement data – guide to the expression of uncertainty in measurement. JCGM 100, 2008.
- [132] J. C. Damasceno and P. R. Couto. *Methods for evaluation of measurement uncertainty*. IntechOpen, Rijeka, 2018.
- [133] P. Holnicki and Z. Nahorski. Emission data uncertainty in urban air quality modeling – case study. *Environmental Modeling & Assessment*, 20:583–597, 2015.
- [134] A. Gressent, L. Malherbe, A. Colette, H. Rollin, and R. Scimia. Data fusion for air quality mapping using low-cost sensor observations: feasibility and added-value. *Environment International*, 143:105965, 2020.
- [135] F. Bouttier and P. Courtier. Data assimilation concepts and methods, 1999.
- [136] J. R. Taylor. *An Introduction to Error Analysis*. 1982.

- [137] Paul Hamer, Sam-Erik Walker, and Philipp Schneider. Appropriate Assimilation Methods for Air Quality Prediction and Pollutant Emission Inversion: An Urban Data Assimilation Systems Report, 2021.
- [138] A. Monteiro et al. Ensemble techniques to improve air quality assessment: focus on O₃ and PM. *Environmental Modeling and Assessment*, 18:249–257, 2012.
- [139] Finnish Meteorological Institute. SILAM v.5.7: System for Integrated modelLing of Atmospheric coMposition. Model and Data Access. <http://silam.fmi.fi/thredds/catalog.html>, 2022.
- [140] Estonian Environmental Research Centre. Estonian air quality. <http://airviro.klab.ee/>, 2021.
- [141] Thinnect. Smart city overview. <https://thinnect.com/smart-city-overview/>, 2019.
- [142] Finnish Meteorological Institute. Air quality forecasts. <https://en.ilmatieteenlaitos.fi/air-quality-forecasts>, 2021.
- [143] European Environment Agency. Download of air quality data. <https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm>, 2022.
- [144] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar. Comparative analysis of machine learning techniques for predicting air quality in smart cities. *IEEE Access*, 7:128325–128338, 2019.
- [145] B. Crawford, D. H. Hagan, I. Grossman, E. Cole, L. Holland, C. L. Heald, and J. H. Kroll. Mapping pollution exposure and chemistry during an extreme air quality event (the 2018 Kilauea eruption) using a low-cost sensor network. *Proceedings of the National Academy of Sciences of the United States of America*, 118(27):e2025540118, 2021.
- [146] D. Zhang and S. S. Woo. Real time localized air quality monitoring and prediction through mobile and fixed iot sensing network. *IEEE Access*, 8:89584–89594, 2020.
- [147] I. Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici. Uncertainty-aware deep learning architectures for highly dynamic air quality prediction. *IEEE Access*, 9:14765–14778, 2021.
- [148] Geir Evensen, Femke C. Vossepoel, and Peter Jan van Leeuwen. Data assimilation fundamentals: A unified formulation of the state and parameter estimation problem. *Data Assimilation Fundamentals*, 2022.

Abstract

Open Environmental Data Assimilation Under Unknown Uncertainty and Multiple Spatio-Temporal Scales

This dissertation focuses on advancing environmental monitoring through novel data assimilation methods. It addresses the challenges in enhancing data quality for environmental monitoring and modeling, particularly in open environmental data. Key contributions include the development of lightweight data assimilation algorithms that improve the accuracy of urban air pollution estimates without prior uncertainty estimates. These scalable algorithms are suitable for IoT devices and support continuous, accurate environmental data collection, aligning with EU environmental directives.

The work also emphasizes the importance of open environmental data assimilation for informed policy-making and compliance monitoring. It addresses research questions on improving data completeness, accuracy, and precision, and the efficacy of the new algorithms in improving air quality data.

Compared to existing methods, this dissertation offers cost-effective, open-source alternatives for environmental monitoring, promoting collaborative innovation and democratizing advanced monitoring capabilities. Future research directions include air pollution mapping, sensor placement optimization, and expanding the algorithms' applications. The research supports EU environmental policies and regulations and provides open-source code for ongoing research, enhancing environmental monitoring's precision, timeliness, and extent.

Kokkuvõte

Keskkonna avaandmete assimilatsioon tundmatu määramatuse ning erinevate aeg-ruumi skaalade korral

Käesolev doktoritöö keskendub keskkonnaseire edendamisele uuenduslike andmete assimileerimise meetodite abil. Töös käsitletakse keskkonnaseire ja modelleerimise valdkonna väljakutseid seoses andmete kvaliteedi parandamisega, pöörates erilist tähelepanu avatud keskkonnaandmetele. Peamised panused, eesmärgiga parandada linnakeskkonna õhusaaste mõõtetulemuste täpsust, hõlmavad kergekaaluliste algoritmide arendamist ilma määramatuse hinnanguta andmete assimileerimiseks. Loodud algoritmid on hästi skaaleruvad, vastavad ELi keskkonnadirektiividele, sobivad IoT seadmetele ja toetavad pidevat, täpset keskkonnaandmete kogumist.

Töö rõhutab ka avatud keskkonnaandmete assimileerimise olulisust teadliku poliitika kujundamisel ja sellest tulenevate nõuete ning regulatsioonide jälgimisel. Töö käsitleb uurimisküsimusi andmete täielikkuse, täpsuse ja kordustäpsuse parandamise ning uute algoritmide tõhususe kohta õhu kvaliteedi andmete parandamisel.

Võrreldes olemasolevate meetoditega pakub käesolev doktoritöö kuluefektiivseid, avatud lähtekoodiga alternatiive keskkonnaseireks, edendades koostööl põhinevat innovatsiooni ja demokratiseerides arenenud seirevõimalusi. Tulevased uurimissuunad hõlmavad õhusaaste kaardistamist, sensorite paigutuse optimeerimist ja algoritmide rakenduste laiendamist. Uurimus toetab ELi keskkonnapoliitikat ja -regulatsioone ning pakub avatud lähtekoodiga koodi käimasolevaks uurimistööks, suurendades keskkonnaseire täpsust, õigeaegsust ja ulatust.

Appendix 1

I

Lizaveta Miasayedava, Keegan McBride, and Jeffrey A. Tuhtan. Automated Environmental Compliance Monitoring of Rivers with IoT and Open Government Data. *Journal of Environmental Management*, 303:114283, February 2022



Contents lists available at ScienceDirect

Journal of Environmental Management

journal homepage: www.elsevier.com/locate/jenvman

Automated environmental compliance monitoring of rivers with IoT and open government data

Lizaveta Miasayedava^{a,1,*}, Keegan McBride^b, Jeffrey Andrew Tuhtan^c^a Research Laboratory for Proactive Technologies, Tallinn University of Technology, Akadeemia tee 15a, 12618, Tallinn, Estonia^b Hertie School's Centre for Digital Governance, Friedrichstraße 180, 10117, Berlin, Germany^c Department of Computer Systems, Tallinn University of Technology, Akadeemia tee 15a, 12618, Tallinn, Estonia

ARTICLE INFO

Keywords:

Environmental compliance monitoring
 Environmental flows
 Internet of things
 Open government data

ABSTRACT

Environmental monitoring of rivers is a cornerstone of the European Union's Water Framework Directive. It requires the estimation and reporting of environmental flows in rivers whose characteristics vary widely across the EU member states. This variability has resulted in a fragmentation of estimation and reporting methods for environmental flows and is exhibited by the myriad of regulatory guidelines and estimation procedures. To standardise and systematically evaluate environmental flows at the pan-European scale, we propose to formalise the estimation procedures through automation by reusing existing river monitoring resources. In this work, we explore how sensor-generated hydrological open government data can be repurposed to automate the estimation and monitoring of river environmental flows. In contrast to existing environmental flows estimation methods, we propose a scalable IoT-based architecture and implement its cloud-layer web service. The major contribution of this work is the demonstration of an automated environmental flows system based on open river monitoring data routinely collected by national authorities. Moreover, the proposed system adds value to existing environmental monitoring data, reduces development and operational costs, facilitates streamlining of environmental compliance and allows for any authority with similar data to reuse or scale it with new data and methods. We critically discuss the opportunities and challenges associated with open government data, including its quality. Finally, we demonstrate the proposed system using the Estonian national river monitoring network and define further research directions.

1. Introduction

The negative environmental impacts on societies and ecosystems are frequently driven by human activity and amplified by increasing climatic variability. The impacts may degrade the environment, decrease available natural resources, increase pollution levels, damage human health and well-being. Due to the importance of the environment and its role in societal and ecosystem health, governments have a persistent interest in regulating, monitoring and managing the state of the environment.

River ecosystems are one of the most heavily regulated, monitored, and governed aspects of the environment. Maintaining healthy river ecosystems is necessary to ensure a large number of benefits to society, such as the provision of the drinking water supply, irrigation, agriculture, power generation, transportation, and industry (Tickner et al.,

2017). However, to secure the health and benefits of these ecosystems, costly investments are needed for maintenance, governance, infrastructure, and regulation. Meanwhile, new technological developments are increasingly being highlighted and tested to offset or reduce traditionally high costs.

Environmental management is experiencing an increased interest in the use of disruptive technologies, such as the Internet of Things (IoT) and big data analytics, to improve efficiency, productivity, and effectiveness of service delivery. However, systematic approaches integrating technological advances into existing governance and regulatory frameworks remain largely absent for the monitoring and evaluation of river ecosystems. At the same time, such approaches are precisely what is needed for the governance of river ecosystems (Tickner et al., 2017). It is our belief that innovative technologies are necessary for better environmental management and monitoring, but their implementation alone

* Corresponding author.

E-mail addresses: lizaveta.miasayedava@taltech.ee (L. Miasayedava), mcbride@hertie-school.org (K. McBride), jeffrey.tuhtan@taltech.ee (J.A. Tuhtan).¹ Permanent address: Tallinn University of Technology, Tallinn, Estonia (Department of Computer Systems, Akadeemia tee 15a, 12618 Tallinn).

is not wholly sufficient to provide the urgently needed gains in effectiveness. To address this, we provide an Environmental Intelligence (EI) conceptualisation for environmental service delivery, which considers a broad variety of technical and non-technical social, political and economic challenges needed to address the growing scope of environmental impacts on societies and ecosystems (Shin, 2014; Mathis et al., 2018).

The proper management of negative impacts to river ecosystems requires ensuring environmental regulatory compliance in the face of increasing challenges related to environmental uncertainty, scientific (estimation methodologies), technological (e.g. missing equipment, failing sensors, infrastructure, and acquisition of high quality, up-to-date, and relevant data), and regulatory (e.g. lacking or missing policies). The monitoring of river ecosystem health requires complete and accurate information on the river flow regime, water usage, and the state of the aquatic ecosystems. However, it is often missing, of poor quality, or limited to a specific geographical scope (Pahl-Wostl et al., 2013).

In the European Union (EU), the Water Framework Directive (WFD) establishes the necessary cross-cutting water policies committing the EU member states to achieve a good status of water bodies, and serves as the fundamental framework for European water legislation (European Parliament/Council of the European Union, 2000). The compliance of rivers with the WFD involves the maintenance of the hydrological regime (water quantity and dynamics) sustaining aquatic ecosystems quantified by environmental (or ecological) flows (eflows) (European Commission, 2015).

The methods to estimate eflows vary depending on the situational complexity, available resources and user requirements (European Commission, 2015; Poff et al., 1997; Stamou et al., 2018; Zeiger and Hubbart, 2021; Parasiewicz et al., 2018). However, the geographical scope of their implementation is uneven, and the eflows requirements are not systematically assessed at a large scale, which is fundamental to improve the effectiveness of eflows estimation as a water management tool (Gopal, 2016; Mezger et al., 2019). The impediments to large-scale implementation include the complexity of ecosystem functions and habitat biodiversity, ambiguity and uncertainty over method choice, lack of transferable eflows rules and strategies, agility and scalability of the existing estimation system and legislation, insufficient resources and knowledge, lack of collaboration, commitment and support by governments and stakeholders, conflicts of interest, and many other (Pahl-Wostl et al., 2013; Opperman et al., 2018; Mezger et al., 2019; Espinoza et al., 2021; Gopal, 2016; Le Quesne et al., 2010; Parasiewicz et al., 2018; European Commission, 2015).

Organisational uncertainties and inefficiencies resulting from the complexity of eflows policies and laws have a significant effect on the implementation of the WFD objectives and commitment of executives (Taylor et al., 2021; Wineland et al., 2021; Manna and Moffitt, 2019). Therefore, we suggest complementing expensive and time-consuming ecological studies involving deep evaluation, complex habitat simulation and expert consultation with automated minimum viable eflows estimation and monitoring.

To enable automated eflows monitoring, we propose implementing automated, data-driven methods. In particular, simple hydrological methods use data routinely collected by the government hydrological services (e.g. flow rates, water depths and temperatures), and their implementation would allow for an automated minimum viable estimation procedure. In the scope of the work, we make use of the exceedance probabilities of the calculated eflows (minimum flow rates required by aquatic species) as a compliance criterion, which can further be updated and replaced with more detailed criteria as needed.

In this paper, we ask the following research question: "How can open government data be repurposed to enable automated eflows estimation and monitoring for river management at the national scale?". To answer this research question, we built and tested an IoT-based system for automated eflows estimation and monitoring based on the Estonian national river sensing network and its open government hydrological data. Methodologically, this paper follows a particular implementation

of the design science research methodology that also takes into account the environmental intelligence (EI) cycle. Though this paper focuses on the specific case of Estonia, the approach adopted here is well suited for new implementations in different country contexts. The core contribution of this paper is related to the design and demonstration of a system that enables automated monitoring and evaluation of eflows by repurposing existing open river monitoring data routinely collected by the government, followed by the introduction of the concept of the "Internet of Open Government Data and Things" (IoOGDT).

2. Background

Environmental managers have a unique role in large-scale environmental monitoring because they possess information relating to the state of the environment, meteorological observations and forecasts of different types and enforce environmental policies, legislation, plans, permits, licenses, and others. Currently, environmental regulations are unevenly enforced (Keith-Roach et al., 2014), and environmental regulatory technologies, which could promote the implementation of environmental regulations, remain rare.

In the water domain, the enforced objectives of the EU Water Framework Directive (WFD) to achieve good chemical and ecological statuses of all surface and groundwater bodies remain unachieved, and deadlines across Europe have been extended (European Commission, 2017a). The major reasons for non-compliance are large uncertainties associated with the definition of monitoring and assessment procedures of water protection areas, as well as the lack of adaptation strategies to the effects of climate change, including more frequently occurring flooding and droughts (European Commission, 2017a; Voulvoulis et al., 2017).

The monitoring of compliance of rivers with the WFD involves the estimation of the hydrological regime (water quantity and dynamics) sustaining aquatic ecosystems often quantified with environmental (or ecological) flows (eflows) (European Commission, 2015). Since governments already collect river data for national and regional monitoring, reporting and research and provide them to open access as open government data (OGD), the data can be reused for eflows estimation. Different types of river OGD can be fused (Ghamisi et al., 2019; Modafferi et al., 2013), merged with proprietary data and software (Gupta, 2012) to complement each other, calibrate sensors (Ferrer-Cid et al., 2020), validate the reliability of data (Ocio et al., 2019), estimate unmeasured locations with some form of spatial interpolation (Modafferi et al., 2013) and create new environmental data-driven solutions supporting decision- and policy-making. The real-time provision of river OGD allows for the automation of compliance monitoring of rivers that can serve the needs of various environmental policies and agencies (European Commission, 2017b, 2019).

Substantial literature has been devoted to the study of how OGD can be used to drive the creation of new innovative services (Janssen et al., 2012; McBride et al., 2019a; Jetzek et al., 2014). However, there remains a lack of studies that specifically examine OGD originating from the Internet of Things (IoT) networks and how these OGD may be utilised in a similar way to how it is done within this paper. This paper primarily differs from previously conducted research in that it specifically focuses on how OGD, generated by an IoT system, can be used to drive and enable the development of a targeted tool for environmental monitoring. When previously published papers do discuss IoT and OGD, the papers are almost exclusively related to the topic of smart cities, see, for example, Aguilera et al. (2017); Zanella et al. (2014); Ahlgren et al. (2016), or rather explore the generation and use of data from IoT devices (Montori et al., 2017; Borges Neto et al., 2015; Calbimonte et al., 2012). While such papers do, sometimes, focus on the utilization of OGD, they also often refer simply to data generated from IoT devices. These data are not conceptually the same as OGD. Furthermore, such papers are not necessarily focused on the immediate usage of OGD, generated from IoT networks, for a specific environmental monitoring task or application. In

this paper, this specific concept of OGD being generated and released from IoT networks, is defined/conceptualized as the Internet of Open Government Data and Things (IoOGDT). IoOGDT can be understood as data that have been collected or paid for by a government organisation, are generated via IoT and subsequently released as OGD (with an appropriate license, in a machine-readable, human-understandable, and freely reusable format). Data released in this way may help to improve and extend the effectiveness and value of the IoT system since anyone can take advantage of them to build a new useful service (Mergel et al., 2018).

As OGD may help with improving the quality, relevance, and impact of IoT, so too may IoT also help drive the improvement and use of OGD. The primary reason for this is that data coming via an IoOGDT system may be reasonably expected to be more relevant, accurate, timely, and useable than other standard sources of OGD. This is an important point because it is known that one of the primary reasons that individuals or citizens do not take advantage of OGD is due to lack of quality, relevance, or timeliness (Young and Yan, 2017; Zuiderwijk et al., 2012). It is also known that one of the highest predictors for future use of OGD is related to a positive previous experience with an OGD-based system (McBride et al., 2019b). Thus, it can be argued that OGD and IoT generate a symbiotic relationship, with each improving the other, and that by releasing OGD collected via IoT, it is possible to encourage the creation of new services, thereby simultaneously extending and improving the value of the IoT system.

3. Methodology

3.1. Procedures and data sources

In order to answer the research question of how to develop a system providing automated environmental compliance monitoring of rivers, we adopt a particular implementation of the design science research methodology (Peffer et al., 2007) - environmental intelligence (EI) cycle (Mathis et al., 2018), involving an iterative creation and application of an environmental technology artefact addressing a specific design problem and facilitating its better understanding. The adaptation of the EI cycle for the eflows compliance estimation is presented in Fig. 1.

The EI cycle can be useful for stakeholders of different sectors,

supporting various environmental focus areas, e.g. business risk management, critical infrastructure protection, military operations, as well as environmental regulatory compliance management stakeholders. Data collection can involve spatial and temporal environmental data of multiple types, used independently, in parallel, assimilated or fused. With the extraction of the functional evidence from the data, the new information is evaluated, communicated and applied. The steps of the EI cycle are iteratively revisited to respond to the needs of stakeholders.

To reduce implementation expenses, we suggest repurposing available river data. Rivers provide a variety of uses, including shipping, drinking water, and irrigation. And river monitoring data are widely available from local, regional, national and international institutions and are supported by domain experts with multi-generational knowledge. IoT-based river monitoring systems are now rapidly advancing beyond data logging services to provide data-driven solutions for pumping station control (Dong and Yang, 2020), increase the automation level of water-intensive agriculture (Puranik et al., 2019) and flood alert systems (Rani et al., 2020).

In this work, in order to estimate the compliance of river flows, we use the data collected by the Estonian Ministry of the Environment for weather reporting and water resource management obtained on request as well as publicly accessible Estonian river data - Estonian hydrological OGD. This paper uses hydrological observations from 54 Estonian gauging stations (see Fig. 2) from January 01, 2009 to December 31, 2018 with the following features:

- minimum/maximum/average daily water level [centimetres],
- minimum/maximum/average daily water temperature [°C],
- minimum/maximum/average daily discharge [meters cubed per second].

Open government data include longer time series of historical hydrological measurements from the gauging stations, starting from 1867 up to 2020 (see open datasets on Ministry of the Environment, 2020) and have the following features:

- average daily water level [centimetres],
- average daily discharge [meters cubed per second].

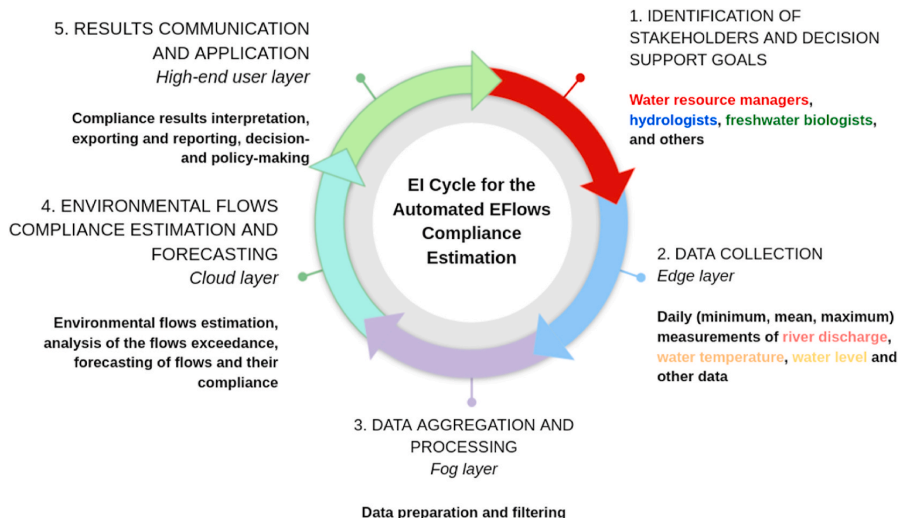


Fig. 1. Environmental Intelligence (EI) cycle for the automated environmental flows (eflows) compliance estimation. Each step of the cycle corresponds to a layer of the IoT-based system automating the environmental compliance monitoring.

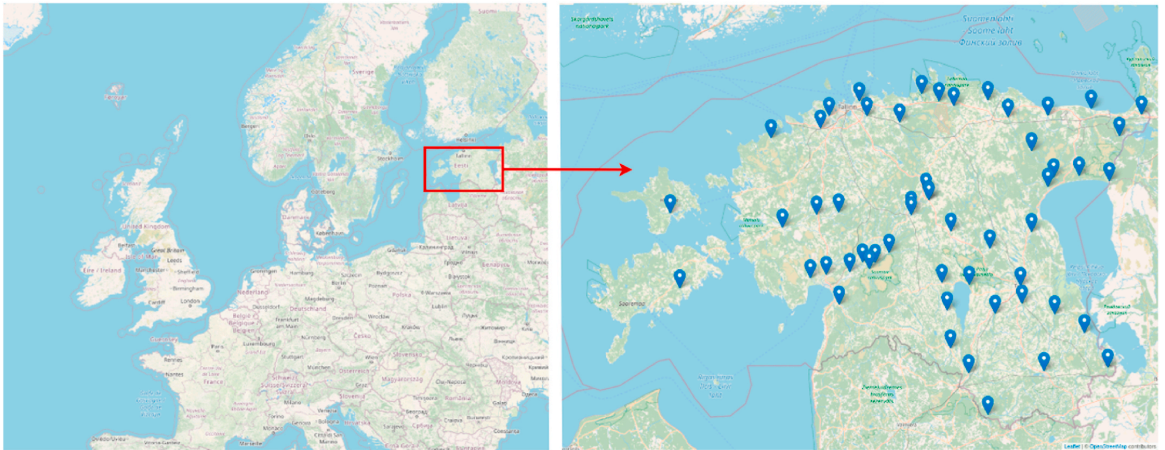


Fig. 2. Map showing the locations of the 54 sites which make up the Estonian river hydrological monitoring network, the map layer is retrieved from <https://www.openstreetmap.org>.

3.2. Assumptions and limitations

In this work, we demonstrate how automated compliance monitoring can be achieved in a scalable and practical manner by utilising IoT and simultaneously taking advantage of existing data. The limitations of the work include the limitations of methods used to estimate eflows, the necessity to establish procedures for how to use the results of assessment for regulatory purposes and the lack of availability of open IoT infrastructure.

Automatable methods of eflows estimations are often too simple for the sufficient evaluation of river ecological status, whereas current extensive field studies are expensive, time-consuming and cannot be carried out very often in many locations, which is necessary considering climate change conditions (Parasiewicz et al. (2018)). We acknowledge

the limitations of the methods and do not resolve them. Instead, we suggest complementing extensive manual assessments with a simplified automated assessment to extend the evaluation coverage and encourage the creation of more profound but automatable assessment methods.

Since there are no universally used methods of eflows assessment for regulatory purposes, there is a need to establish procedures and compliance rules. The proposed service provides a simple estimation and interpretation tool for environmental specialists, governmental authorities and other stakeholders. Still, the definition of specific rules for water management actions and measures are, however, out of the scope of the paper.

In this work, we reuse the data from existing river sensing infrastructure provided and maintained by the government for river monitoring purposes. We assume that the fog and edge layers are

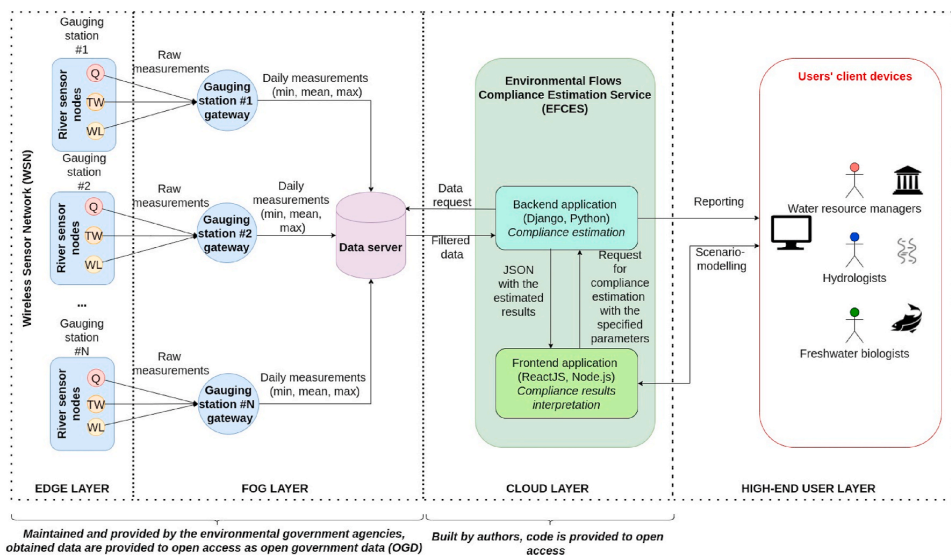


Fig. 3. Architecture of the proposed IoTGD real-time environmental flows (eflows) compliance monitoring system. The edge layer consists of the river sensor nodes which measure the flow rate (Q), water temperature (TW) and water level (WL). The raw measurements undergo data quality control and post-processing, the statistical results are pushed to the data server. The compliance estimation service runs at the cloud layer, which is accessed by high-end users.

implemented as a wireless sensor network, but it might be not always the case. Open access to data can add value to any existing infrastructure implementing a river monitoring system.

4. Results

4.1. System architecture

To automate the estimation and monitoring of compliance with eflows levels, we have designed an IoOGDT system, the architecture of which is presented in Fig. 3. The workflow of the system is defined by the EI cycle mentioned above (see Fig. 1).

The edge layer of the IoT system is represented by a Wireless Sensor Network (WSN) carrying out the collection of river data. The fog layer involves data aggregation and processing, which is organised through passing the collected data to the data server through sink nodes, or “gateways”. The collected measurement data are stored in the data server. The presented river data are commonly collected by governments and provided to open access. Reusing the data, we develop a cloud layer of the IoOGDT system - eflows compliance estimation service (EFCES). EFCES is a dockerised Django/React web application providing eflows compliance estimation and interpretation for reporting. The code repository of the web service is available open-source on GitHub, see <https://github.com/effie-ms/efflows>.

EFCES implements the following functional requirements:

1. Hydrological data acquisition and manipulation.
2. Environmental flows estimation.
3. Compliance estimation and interpretation.
4. Configuration of estimation parameters.
5. Export of the generated compliance information.

The estimation of eflows by EFCES is carried out similarly to existing eflows calculators (Smakhtin and Anputhas, 2006). EFCES allows setting the eflows estimation parameters (gauging stations, dates, methods to be used, and other parameters). The processing and calculations are implemented at the backend (Django) application of the service, and the visual interpretation of results is carried out by the frontend (React) application. The service also allows the export of the obtained information for reporting, representing a benefit to water managers who can simply use the tool to generate needed estimations, images, and reports.

On the high-end of the system, water resource managers, hydrologists, freshwater biologists, and other users use the system through the EFCES interface to access reports on the environmental compliance of rivers as well as test various methods and parameters of compliance estimation using the available river data.

4.2. Data quality

Data quality is essential to enable the automated processing of data. However, during the development of the service using the data from multiple sources, we have faced the problems of differences in formatting, missing data and inconsistencies in existing data. For the development of the eflows compliance estimation service, we used data provided by the government to open access (referred to as “OGD” - open government data) with preliminary quality control and a sample of data from the same owner without revision (referred to as “Requested data”).

OGD have historical time series of mean flow rates and water levels from 1867 to 2020. The availability of data is increasing with the installation of new gauging stations (see Fig. 4). However, downward fluctuations in the number of stations are associated with missing data.

The data obtained on request from the government correspond to the period from January 01, 2009 to December 31, 2018. Taking OGD from the same period, we have compared the percentages of missing data for that period from all the 52 stations common for both datasets, the results are presented in Table 1.

The approaches to impute (fill in) missing data vary. Since open and requested data correspond to the same locations, average flow rates and

Table 1

Missing data in Estonian hydrological datasets: river data obtained from the government on request and open government data (OGD). Each column corresponds to percentages of missing values in minimum, mean and maximum daily time series of flow rates (Q) and water levels (WL) from January 01, 2009 to December 31, 2018. NA - data not available.

Data source	Percentage of missing values [%]					
	Min Q	Mean Q	Max Q	Min WL	Mean WL	Max WL
Requested data	10.06	11.71	10.05	9.52	3.26	9.48
OGD	NA	2.11	NA	NA	2.64	NA

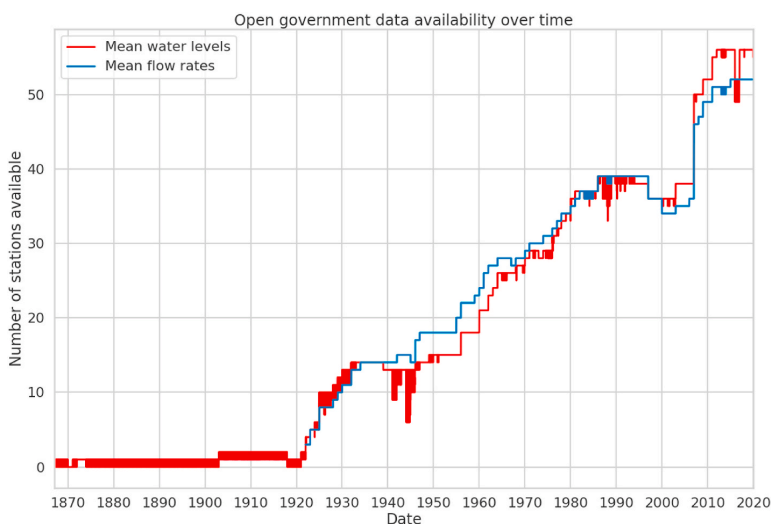


Fig. 4. Availability of daily river open government data (OGD) at Estonian gauging stations over time from 1867 to 2020.

water levels can be filled with the corresponding OGD levels and vice versa. At the same time, due to Estonia’s relatively uniform hydroclimate and homogeneous river morphologies, the river data of most Estonian rivers are highly correlated (see Table 2), allowing for the implementation of various imputation methods (e.g., see Ekeu-wei et al. (2018)).

Another problem occurring with data from multiple sources is inconsistencies caused by sensor failures, human errors from manual processing and revision of data. Given the minimum and maximum values from the requested data, we have carried out a simple comparison of average values with them, assuming that average values should be within the range of corresponding minimum and maximum values. The numbers of violations of the rule (inconsistencies when average values are out of range) were counted, the percentages are given in Table 3.

There are many different types of faults and inconsistencies that can affect further steps of analysis and modelling and consequent decision-making. The faults and inconsistencies in data can be caused by uncalibrated sensors, hardware failures or extreme environmental conditions. The validation of data quality has many challenges for the owners of the data collection infrastructure, and it can become even more challenging for parties reusing the data of invalidated quality.

The data from multiple sources can be combined to potentially their correctness and other quality parameters. In particular, with the provision of uncertainty estimates, well-calibrated redundant sensors can be fused or assimilated with numerical forecasts to provide a potentially better estimate with decreased uncertainty. The procedures to address the discussed problem are left for further research.

4.3. Demonstration

To demonstrate IoOGDT for eflows compliance estimation and monitoring, we used Estonian hydrological OGD, which is publicly available data from the Estonian national hydrological sensing network (see Fig. 2). Current Estonian environmental regulation (Ministry of the Environment, 2014) defines the national methodology for determining the minimum eflow for the ice-free period from May to October as calculating the average monthly minimum flow with a 95% exceedance probability (1).

$$q_{env} = q_{desc} \lceil p(N + 1) \rceil \tag{1}$$

where.

- q_{env} - eflow discharge (or volumetric flow rate) [m³/s],
- q_{desc} - array of observed discharge values sorted in the descending order [array of m³/s],
- p - probability of exceedance for an observed discharge value [0 to 1] (e.g. for 95% exceedance probability $p = 0.95$).
- N - total count of observations in q_{desc} .
- $\lceil \cdot \rceil$ - operator of rounding to the nearest integer.
- $[\cdot]$ - index operator (taking a value in an array by index).

The eflows calculation according to Formula (1) requires river discharge data in the form of volumetric flow rates. To estimate the compliance of river flows following the eflow threshold, a user needs to select a location and configure the estimation parameters. An example of

Table 2

The Pearson cross-correlation coefficients of time series from 52 Estonian gauging stations of both river data obtained from the government on request and open government data (OGD). Each column corresponds to Pearson cross-correlation ranges of coefficients of minimum, mean and maximum daily time series of flow rates (Q) and water levels (WL) from January 01, 2009 to December 31, 2018. SD - standard deviation, NA - data not available.

Data source	Pearson cross-correlation coefficients [mean ± sd (min; max)]					
	Min Q	Mean Q	Max Q	Min WL	Mean WL	Max WL
Requested data	0.71 ± 0.25 (−0.58; 0.98)	0.73 ± 0.24 (−0.45; 0.98)	0.70 ± 0.25 (−0.58; 0.98)	0.61 ± 0.22 (−0.29; 0.99)	0.62 ± 0.21 (−0.24; 0.99)	0.61 ± 0.21 (−0.26; 0.98)
OGD	NA	0.73 ± 0.24 (−0.45; 0.98)	NA	NA	0.62 ± 0.21 (−0.26; 0.98)	NA

Table 3

Inconsistent data in Estonian hydrological datasets: river data obtained from the government on request and open government data (OGD). Each column corresponds to percentages of inconsistent values identified from the comparison of minimum, mean and maximum values of daily time series of flow rates (Q) and water levels (WL). The percentages are calculated from the non-missing data in the period from January 01, 2009 to December 31, 2018.

Data source	Q	WL
Requested data	16.89% (out of 79.92% available)	4.92% (out of 88.15% available)
OGD	48.47% (out of 89.71% available)	46.24% (out of 88.79% available)

a plot generated by the developed environmental flows compliance estimation service (EFCES) is shown in Fig. 5. Refer to Supplementary material or README of the.

web-service repository (<https://github.com/effie-ms/eflows>) to view the user manual, screenshots of its graphical user interface and examples of other plots that can be generated.

The long-term management of rivers requires further analysis of continuous and cyclical trends, cumulative effects of eflows non-compliance in four different bioperiods.

For example, using formula (1), we calculate the number of all the occurred low flow events (below eflow thresh-old) - the total number of days of non-compliance per each bioperiod (bioperiods are described in Parasiewicz et al. (2018)) over 2009–2018. The results are presented for three Estonian rivers monitored at the following gauging stations: Narva linn (Narva river), Korela (Piusa river), and Toila-Oru (Pühajõgi river) located in eastern Estonia. The days of non-compliance are plotted in Fig. 6 where the “large” Narva river (corresponding to the 100th percentile as the largest river in Estonia) is shown as a black line, a grey line corresponds to the “medium” Piusa river (57th percentile) and the “small” river, Pühajõgi (20th percentile) is shown as a red line.

The results for the first “overwintering” bioperiod from January to February is shown in the top left panel of Fig. 6. During this bioperiod, the decreasing trend observed in Narva river is assumed to be related to climate change as there have been no major anthropogenic changes in the river during the period of observation. However, it can be seen that none of the other rivers shows a substantial change in the days of non-compliance. Therefore, decision-makers may wish to carry out targeted mitigation measures. This could be accomplished for example by routing more water flows to Narva river, reducing the number of days of non-compliance to as close to zero as possible, as during the overwintering bioperiod.

The “spring spawning” bioperiod runs from March to June, and is shown in the top left panel of Fig. 6. Here, the effects of increased winter seasonal air temperature and precipitation in Estonia have resulted in an earlier beginning and decrease in spring floods, and consequently longer dry periods. Again, the most impacted river is Narva river, which clearly shows a recent jump in the days of non-compliance from 2017. Water managers may therefore wish to target this region of the river for mitigation measures, for example by artificially creating more fish spawning habitats to help reduce the stress placed on fish due to the increase in days of eflow non-compliance in this bioperiod.

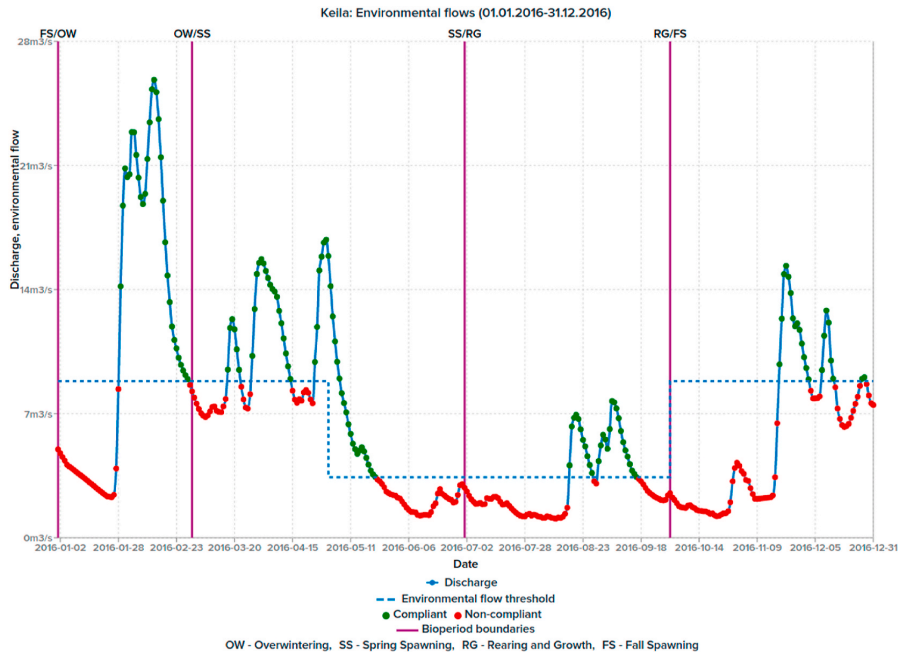


Fig. 5. Example of compliance estimation results for the Keila River in 2016 using the environmental flows compliance estimation service (EFCES). The blue curve corresponds to the flow rate of the river, the dashed blue line indicates the calculated eflow level. Compliance is indicated by colour. Red: the river flow is below the required eflow level, non-compliant. Green: the flow is at or above the required eflow level, compliant. The four major bioperiods for local fish populations used in the subsequent assessment of compliance are shown as vertical purple lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

The third bioperiod occurs during summer, and is referred to as “rearing and growth”. Considering all of the three rivers assessed in detail, a clear biannual pattern from 2011 onwards can be seen in the bottom left panel of Fig. 6. This pattern has been attributed to climate change and has been recognised to have a negative effect not only on fish populations but also on the water quality (chemical status) and aquatic habitats for other animal species as well (Kallis et al., 2017). In contrast to the first two bioperiods, the pattern of eflow non-compliance affects rivers of all sizes. Therefore the changes in Estonian river eflows observed in the rearing and growth bioperiod may present a large-scale problem for aquatic biodiversity. Especially concerning is the rapid increase in the days of non-compliance observed in 2018 for all three rivers evaluated.

Finally, the fourth “fall spawning” bioperiod is shown in the bottom right panel of Fig. 6. Once again, the large Narva river is clearly differentiated from the remainder of the other five rivers, with the highest number of days of eflows non-compliance from 2009 to 2018. In contrast to the rearing and growth bioperiod, there is no clear large scale pattern affecting all rivers. However, there is some commonality in peaks of non-compliance in 2010 and 2015 between Narva river and Pühajõgi. We believe that the explanation for this may be geographical, as both rivers are located in northeast Estonia, and are therefore subject to similar rainfall conditions.

Thus, the proposed system automates the estimation and reporting procedures, however, the analysis of the estimated results (as described above) for decision- and policy-making requires the expertise of water managers and other associated specialists, or further automated to assist planning procedures. In particular, the next step would be to add on the ability to simulate water management actions addressing the ecological conditions: restoration of the natural flow regime, decrease of water withdrawals to agricultural and industrial end-users, etc.

5. Discussion

Building new large-scale environmental monitoring networks and integrating the resulting data into bespoke decision-making support systems is expensive and time-consuming (Lovett et al., 2007), and the additional communication costs alone can be prohibitive (Liu et al., 2010). Instead, we propose that existing OGD from national hydrological services can be repurposed for eflows compliance estimation. A web-service demonstrator provides the first step in automating environmental compliance monitoring and evaluation. It allows authorities to allocate their limited resources to more important tasks such as strategic planning and enforcement (Qin, 2011). Using the publicly accessible environmental data, as well as combining it with the data from other sources, creates new opportunities for the reuse of open government data to develop new value-added services and tools. The real-time provision of environmental data to open access empowers data-driven services addressing sustainability and climate resilience challenges with timely reporting, critical for short-term high-risk events.

The implementation and deployment of intelligent environmental solutions supporting the successful implementation of environmental regulations must align with sustainability and climate resilience goals to protect and preserve the environment and enable disaster-resilient governance. The pressure on governments to ensure better water management practices under the increasing frequency of occurrence of water-related disasters can be alleviated using automated environmental monitoring and evidence-based decision making, risk assessment and planning. Environmentally intelligent services require not only a deep understanding of the domain and the problem to be addressed but also include stakeholder feedback after using the system in practice. Environmental regulatory compliance often requires decision- and policy-making under conflicting environmental, social and economic

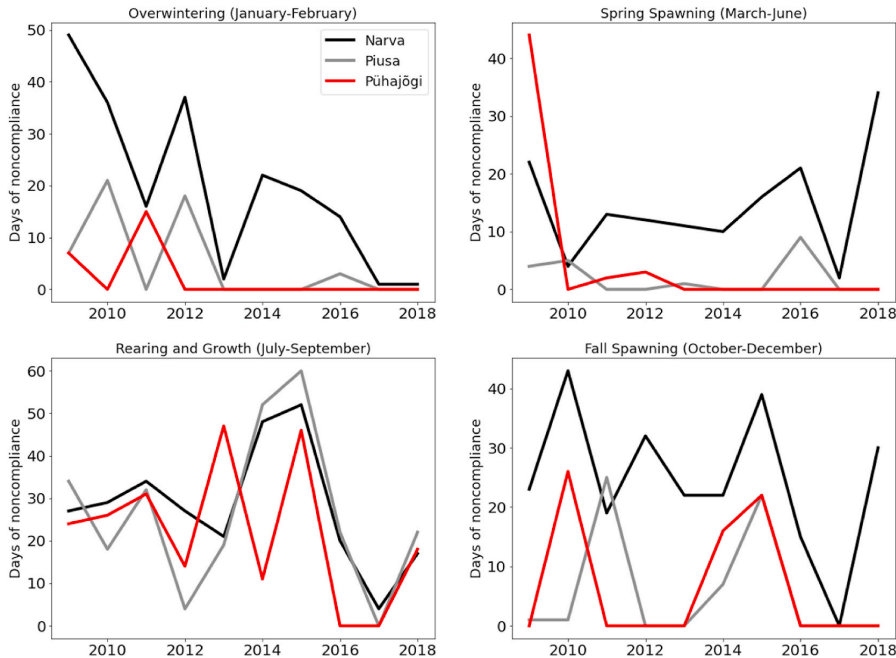


Fig. 6. Compliance estimation per bioperiod results for rivers monitored at three Estonian gauging stations in 2009–2018. Sorted by their mean annual discharges, the “large” Estonian river Narva is shown with black lines, the “medium” river Piusa as grey lines, and the “small” Pühajõgi river as red lines. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

needs of the involved parties. Therefore, the system should serve as an assistant rather than a burden for successful integration.

The major contributions of this work are two-fold. First, we created and implemented an IoOGDT demonstrator automating eflows compliance monitoring and reporting at the national scale. The system provides a scalable platform for the integration of other data-driven methods of eflows compliance estimation, such as the regionally applicable environmental method proposed recently by Parasiewicz et al. (2018). Furthermore, we suggest that the inclusion of compound event scenarios, which are co-occurring weather and climatic combinations of drivers and hazards contributing to societal or environmental risk, can be integrated into these more advanced systems to increase societal climate resilience (Zscheischler et al., 2020).

Second, we introduce the “Internet of Open Government Data and Things” (IoOGDT) concept to encourage others to develop and implement more advanced automated compliance systems using OGD for a wider range of applications. Ultimately, IoT-based systems can provide a largely overlooked source of highly valuable OGD enabling the creation of new and improved systems for decision- and policy-making. We believe that the future of automated monitoring and regulatory compliance lies in linking OGD with the connectivity, speed and ease of use common to IoT technologies. The internet provides the ideal medium for real-time monitoring, thus helping to facilitate secure and reliable online decision-making. In our example, the modular architecture of the eflows system allows for rapid modification and optimisation of various components from functional (e.g. optimisation of routing between nodes, computing, energy use of sensing devices, and others) and non-functional (e.g. security and reliability of data exchange) perspectives. In addition, the integration of modules performing intelligent computing (advanced analytics, machine learning and data-driven learning) can be applied to data pipelines on different network layers (edge, fog or cloud) to detect patterns, extract and interpret the information in an operationally useful way. In future works, we will

investigate the trade-offs in complexity and utility of a large-scale environmental monitoring system with multi-objective, distributed intelligence as a new type of environmental regulatory technology (EnvRegTech).

Since EnvRegTech lags behind RegTech solutions in other sectors, we promote it with the design and development of EI artefact. We suggest designing and implementing EnvRegTech artefacts following the EI framework to ensure that the needs of stakeholders, currently performing the procedures in a semi-automated way, are addressed. With the provision of environmental OGD of proper quality, it should be possible not only to enable automated environmental compliance monitoring and associated with it real-time situational awareness about the ecological state of rivers but also scale this minimum viable assessment and benefit from the proposed centralised solution. Moreover, further insights can be gained after the implementation of scenario and impact modelling and forecasting capabilities, further enhancing evidence-based measures to address river flow alterations.

6. Conclusion

In this paper, we hypothesise and investigate how open government data (OGD), along with IoT-based national river monitoring infrastructure, can be repurposed to automate the estimation of environmental flows (eflows). This estimation is important to establish and monitor regulatory compliance with the EU Water Framework Directive (WFD). In addition, we demonstrate how an environmental intelligence cycle framework can utilise governmental data collection infrastructure to create an automated eflows estimation and compliance estimation web-service. This service can be used by a broad range of stakeholders with an interest in ensuring river environmental regulatory compliance.

We fully acknowledge that it is unrealistic to find a universal eflows method applicable to all rivers. However, we wish to point out that scenario- and impact-modelling are readily achievable based solely on

the available OGD. Our long-term aim is to continue developing the proposed system to automate manual functions (e.g. statistical analysis and reporting), allowing government administrators to focus on other high-priority tasks such as resilience planning (Timashev, 2015) and hazard mapping (Nateghi, 2018).

With the use of Estonian river open government data, we develop an eflows estimation web-service. We showcase the method following current environmental regulations and illustrate its use as a compliance estimation tool via a graphical interpretation of the calculated eflow level exceedance. Through this example, it can be seen that the number of days of non-compliance for each of the four different bioperiods varies depending on the size of the river and can vary widely from year to year. This is important because the variability of flows in rivers is expected to increase due to climate change, indicating that eflows non-compliance rates are therefore subject to increased uncertainty as well. The proposed method is a scalable technology that can be used to further automate eflows estimation and interpretation, aiding policy-makers in assessing how environmental regulations may require further adaptation when faced with increased uncertainty. We critically discuss the associated challenges of the WFD implementation, eflows estimation and the challenges associated with repurposing publicly available river data.

To create a truly automated and scalable system, data standardisation and quality control of the publicly available data must be ensured. Cross-border automation requires consistent supply, the unified format of data and metadata. Data quality is critical since machine processing is sensitive to differences in formatting. The data scattered out in multiple environmental information systems, processed and interpreted differently, complicate capturing the overview on multi-parameter phenomena and scaling of solutions.

This work has multiple future research directions. The system can be improved with alerting, scenario- and impact-modelling, forecasting and early warning capabilities. It could feasibly also assist in decision support and policy-making, by adapting it to reflect additional regulatory requirements (e.g. industrial water withdrawals) and supplemented by stakeholder feedback. To make these changes, it is important to consider that active collaboration and additional interviews with non-governmental stakeholders are required to optimise the workflow.

In conclusion, the current embodiment of an automated eflows compliance monitoring system resolves fundamental inefficiencies of existing river eflows systems, which in their current form hinder the achievement of a good ecological status, as stipulated by the EU WFD. Furthermore, we show that these inefficiencies can be largely addressed by creating an automated eflows estimation service providing a centralised and dynamic overview of the current compliance state of monitored rivers at a national scale.

Credit author statement

Lizaveta Miasayedava: Conceptualisation, Software, Formal analysis, Visualization, Writing – original draft. **Keegan McBride:** Conceptualisation, Validation, Writing – review & editing, Supervision. **Jeffrey Andrew Tuhtan:** Conceptualisation, Data curation, Writing – review & editing, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research has been funded in part by Estonian Research Council grant PRG1243 Multiscale Natural Flow Sensing for Coasts and Rivers, and the Estonian IT Academy IoT programme. We would like to thank

Agne Aruväli from the Estonian Ministry of the Environment and Elina Leiner from the Estonian Environmental Board for their support, guidance and feedback on developing the “Eesti Eflows” system, Isis Vanessa Navarro Rivas from Tallinn University of Technology for conducting interviews and School of IT, Tallinn University of Technology for providing the server. Map data copyrighted OpenStreetMap contributors and available from <https://www.openstreetmap.org>.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jenvman.2021.114283>.

References

- Aguilera, U., Pena, O., Belmonte, O., Lopez-de Ipinza, D., 2017. Citizen-centric data services for smarter cities. *Future Generat. Comput. Syst.* 76, 234–247.
- Ahlgren, B., Hidell, M., Ngai, E.C.H., 2016. Internet of things for smart cities: interoperability and open data. *IEEE Internet Computing* 20, 52–56.
- Borges Neto, J.B., Silva, T.H., Assuncao, R.M., Mini, R.A., Loureiro, A.A., 2015. Sensing in the collaborative internet of things. *Sensors* 15, 6607–6632.
- Calbimonte, J.P., Yan, Z., Jeung, H., Corcho, O., Aberer, K., 2012. Deriving semantic sensor metadata from raw measurements. In: *Proceedings of the 5th International Workshop on Semantic Sensor Networks at ISWC*. CEUR-WS, pp. 33–48.
- Dong, W., Yang, Q., 2020. Data-driven solution for optimal pumping units scheduling of smart water conservancy. *IEEE Internet Things J.* 7, 1919–1926.
- Ekeu-wei, I., Blackburn, G.A., Pedruco, P., 2018. Infilling missing data in hydrology: solutions using satellite radar altimetry and multiple imputation for data-sparse regions. *Water* 10.
- Espinosa, T., Burke, C., Carpenter-Bundhoo, L., Marshall, S., McDougall, A., Roberts, D., Campbell, H., Kennard, M., 2021. Quantifying movement of multiple threatened species to inform adaptive management of environmental flows. *J. Environ. Manag.* 295.
- European Commission, 2015. Ecological flows in the implementation of the water framework directive. *Technical Report Guidance document* 31.
- European Commission, 2017a. Clarification on the application of WFD Article 4(4) time extensions in the 2021 RBMPs and practical considerations regarding the 2027 deadline. In: *Common Implementation Strategy for the Water Framework Directive and the Floods Directive*.
- European Commission, 2017b. From the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions: *Actions to Streamline Environmental Reporting* (Brussels, Belgium).
- European Commission, 2019. *Promotion of Good Practices for National Environmental Information Systems and Tools for Data Harvesting at EU Level* (Technical Report).
- European Parliament, Council of the European Union, 2000. *Directive 2000/60/EC of the European Parliament and of the Council of 23 October 2000: Establishing a Framework for Community Action in the Field of Water Policy*.
- Ferrer-Cid, P., Barcelo-Ordinas, J.M., Garcia-Vidal, J., Ripoll, A., Viana, M., 2020. Multisensor data fusion calibration in IoT air pollution platforms. *IEEE Internet Things J.* 7, 3124–3132.
- Ghamisi, P., Rasti, B., Yokoya, N., Wang, Q., Hofle, B., Bruzzone, L., Bovolo, F., Chi, M., Anders, K., Gloaguen, R., Atkinson, P., Benediktsson, J., 2019. Multisource and multimodal data fusion in remote sensing: a comprehensive review of the state of the art. *IEEE Geosci. Remote Sens. Magazine* 7, 6–39.
- Gopal, B., 2016. A conceptual framework for environmental flows assessment based on ecosystem services and their economic valuation. *Ecosys. Services* 21, 53–58.
- Gupta, P., 2012. User friendly open GIS tool for large scale data assimilation – a case study of hydrological modelling. *ISPRS Int. Archives Photogram. Rem. Sens. Spatial Inform.Sci.* 427–430.
- Janssen, M., Charalabidis, Y., Zuiderwijk, A., 2012. Benefits, adoption barriers and myths of open data and open government. *Inf. Syst. Manag.* 29, 258–268.
- Jetzek, T., Avital, M., Bjorn-Andersen, N., 2014. Data-driven innovation through open government data. *J. Theoretical Appl. Electronic Commerce Res.* 9, 100–120.
- Kallis, A., Sims, A., Tammik, A., Roose, A., Turkson, C.T., Kupri, H.L., Seli, H., Milvee, I., Zaccachaeus, L., Lakson, M., Mols, M., Alber, R., Stokov, S., Jurtom, T., 2017. *Estonia's Seventh National Communication under the United Nations Framework Convention on Climate Change*. Ministry of the Environment.
- Keith-Roach, M., Grundfelt, B., Hoglund, L., Kousa, A., Pohjolainen, E., Magistrati, P., Aggelatou, V., Olivieri, N., Ferrari, A., 2014. *Environmental Legislation and Best Practice in the Emerging European Rare Earth Element Industry*.
- Le Quesne, T., Kendy, E., Weston, D., 2010. *The Implementation Challenge: Taking Stock of Government Policies to Protect and Restore Environmental Flows*. Technical Report. WWF.
- Liu, D., Li, A., Tian, L., Jia, Y., Zou, P., 2010. Study on cost-sensitive communication models on large-scale monitor networks. In: *2010 International Conference on E-Business and E-Government*, pp. 2133–2136.
- Lovett, G.M., Burns, D.A., Driscoll, C.T., Jenkins, J.C., Mitchell, M.J., Rustad, L., Shanley, J.B., Likens, G.E., Haeuber, R., 2007. Who needs environmental monitoring? *Front. Ecol. Environ.* 5, 253–260.
- Manna, P., Moffitt, S., 2019. Traceable tasks and complex policies: when politics matter for policy implementation: traceable tasks and complex policies. *Pol. Stud. J.* 49.

- Mathis, J.T., Osborne, E., Starkweather, S., 2018. Collecting Environmental Intelligence in the New Arctic. <https://arctic.noaa.gov/Report-Card/Report-Card-2017/ArtMID/7798/ArticleID/691/Collecting-Environmental-Intelligence-in-the-New-Arctic>. (Accessed 7 May 2021).
- McBride, K., Aavik, G., Toots, M., Kalvet, T., Krimmer, R., 2019a. How does open government data driven co-creation occur? Six factors and a 'perfect storm'; insights from Chicago's food inspection forecasting model. *Govern. Inf. Q.* 36, 88–97.
- McBride, K., Toots, M., Kalvet, T., Krimmer, R., 2019b. Turning open government data into public value: testing the COPS framework for the Co-creation of OGD-driven public services. In: *Governance Models for Creating Public Value in Open Data Initiatives*. Springer, pp. 3–31.
- Mergel, I., Kattel, R., Lember, V., McBride, K., 2018. Citizen-oriented digital transformation in the public sector. In: *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, pp. 1–3.
- Mezger, G., De Stefano, L., González del Tánago, M., 2019. Assessing the establishment and implementation of environmental flows in Spain. *Environ. Manag.* 64, 721–735.
- Ministry of the Environment, 2014. *Specified Requirements for the Expansion of a Water Body, Environmental Monitoring Related to the Expansion, Protection of Aquatic Life, Dam, Elimination of the Expansion and Lowering of the Water Level, and Methodology for Determining the Minimum Ecological Flow*. <https://www.riigi.teataja.ee/akt/124092014001>. (Accessed 7 May 2021).
- Ministry of the Environment, 2020. *Estonian weather service: historical observation data*. <https://www.ilmateenistus.ee/siseveed/ajaloolised-vaatlusandmed/>. (Accessed 7 May 2021).
- Modafferi, S., Chakravarthy, A., Sabeur, Z., 2013. Multi-level Data Fusion of Environmental Data in Future Internet Applications. *SEBD*.
- Montori, F., Bedogni, L., Bononi, L., 2017. A collaborative internet of things architecture for smart cities and environmental monitoring. *IEEE Internet Things J.* 5, 592–605.
- Nateghi, R., 2018. Multi-dimensional infrastructure resilience modeling: an application to hurricane-prone electric power distribution systems. *IEEE Access* 6, 13478–13489.
- Ocio, D., Beskeen, T., Smart, K., 2019. Fully distributed hydrological modelling for catchment-wide hydrological data verification. *Nord. Hydrol* 50, 1520–1534.
- Opperman, J., Kendy, E., Tharme, R., Warner, A., Barrios, E., Richter, B., 2018. A three-level framework for assessing and implementing environmental flows. *Front. Environ. Sci.* 6.
- Pahl-Wostl, C., Arthington, A., Bogardi, J., Bunn, S., Hoff, H., Lebel, L., Nikitina, E., Palmer, M., Poff, N., Richards, K., Schluter, M., Schulze, R., St-Hilaire, A., Tharme, R., Tockner, K., Tsegai, D., 2013. Environmental flows and water governance: managing sustainable water uses. *Curr. Opin. Environ. Sustain.* 5, 341–351.
- Parasiewicz, P., Prus, P., Suska, K., Marcinkowski, P., 2018. "E = mc2" of environmental flows: a conceptual framework for establishing a fish-biological foundation for a regionally applicable environmental low-flow formula. *Water* 10, 1501.
- Peffer, K., Tuunanen, T., Rothenberger, M., Chatterjee, S., 2007. A design science research methodology for information systems research. *J. Manag. Inf. Syst.* 24, 45–77.
- Poff, N.L., Allan, J., Bain, M., Karr, J.R., Prestegard, K., Richter, B., Sparks, R., Stromberg, J.C., 1997. The natural flow regime. A paradigm for river conservation and restoration. *Bioscience* 47, 769–784.
- Puranik, V., Sharmila, Ranjan, A., Kumari, A., 2019. Automation in agriculture and IoT. In: *2019 4th International Conference on Internet of Things: Smart Innovation and Usages. IoT-SIU*, pp. 1–6.
- Qin, T., 2011. Strategic planning of e-government system for disaster prevention and relief: a case study. In: *2011 International Conference on Computer Science and Service System. C3SS*, pp. 4049–4052.
- Rani, D.S., Jayalakshmi, G.N., Baligar, V.P., 2020. Low-cost IoT based flood monitoring system using machine learning and neural networks: flood alerting and rainfall prediction. In: *2020 2nd International Conference on Innovative Mechanisms for Industry Applications. ICIMIA*, pp. 261–267.
- Shin, D., 2014. A socio-technical framework for Internet-of-Things design: a human-centered design for the Internet of Things. *Telematics Inf.* 31, 519–531.
- Smakhtin, V., Anputhas, M., 2006. An Assessment of Environmental Flow Requirements of Indian River Basins.
- Stamou, A., Polydera, A., Papadonikolaki, G., Martínez-Capel, F., Munoz-Mas, R., Papadaki, C., Zogaris, S., Bui, M.D., Rutschmann, P., Dimitriou, E., 2018. Determination of environmental flows in rivers using an integrated hydrological-hydrodynamic-habitat modelling approach. *J. Environ. Manag.* 209, 273–285.
- Taylor, K., Zarb, S., Jeschke, N., 2021. Ambiguity, uncertainty and implementation. *Int. Rev. Pub. Pol.* 3.
- Tickner, D., Parker, H., Moncrieff, C., Oates, N., Ludi, E., Acreman, M., 2017. Managing rivers for multiple benefits-A coherent approach to research, policy and planning. *Front. Environ. Sci.* 5.
- Timashev, S.A., 2015. Infrastructure resilience: definition, calculation, application. In: *-1078 2015 International Conference on Interactive Collaborative Learning. ICL*, p. 1075.
- Voulvoulis, N., Arpon, K.D., Giakoumis, T., 2017. The EU Water Framework Directive: from great expectations to problems with implementation. *Sci. Total Environ.* 575, 358–366.
- Wineland, S.M., Fovargue, R., York, B., Lynch, A.J., Paukert, C.P., Neeson, T.M., 2021. Is there enough water? How bearish and bullish outlooks are linked to decision-maker perspectives on environmental flows. *J. Environ. Manag.* 280, 111694.
- Young, M., Yan, A., 2017. Civic hackers' user experiences and expectations of Seattle's open municipal data program. In: *Proceedings of the 50th Hawaii International Conference on System Sciences*.
- Zanella, A., Bui, N., Castellani, A., Vangelista, L., Zorzi, M., 2014. Internet of things for smart cities. *IEEE Internet Things J.* 1, 22–32.
- Zeiger, S.J., Hubbard, J.A., 2021. Measuring and modeling event-based environmental flows: an assessment of HEC-RAS 2D rain-on-grid simulations. *J. Environ. Manag.* 285, 112125.
- Zscheischler, J., Martius, O., Westra, S., Bevacqua, E., Raymond, C., Horton, R.M., van den Hurk, B., AghaKouchak, A., Jézéquel, A., Mahecha, M.D., et al., 2020. A typology of compound weather and climate events. *Nat. Rev. Earth Environ.* 1, 333–347.
- Zuiderwijk, A., Janssen, M., Choenni, S., Meijer, R., Alibaks, R.S., Sheikh Alibaks, R., 2012. Socio-technical impediments of open data. *Electron. J. eGovernment* 10, 156–172.

Appendix 2

II

Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty. *Environmental Modeling & Assessment*, June 2023



Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty

Lizaveta Miasayedava^{1,2} · Jaanus Kaugerand¹ · Jeffrey A. Tuhtan²

Received: 10 May 2022 / Accepted: 27 May 2023

© The Author(s), under exclusive licence to Springer Nature Switzerland AG 2023

Abstract

Accurate monitoring systems and numerical models of urban ambient air quality are essential to reduce the risks to public health. The growing quantity of online open data provide new opportunities for assimilation algorithms to improve ambient air quality monitoring, including estimates of their uncertainty. The assimilation of large-scale numerical simulations with observations from urban ambient air quality monitoring stations requires uncertainty estimates from both data sources to cope with unknown events and changing environmental conditions. However, uncertainty estimates from open access numerical models and monitoring stations are frequently unavailable. To address this gap, we propose a lightweight data-driven framework for data assimilation on low-powered embedded hardware suitable for open data without uncertainty estimates, including Internet of Things (IoT) systems. The algorithms are compared and validated using open data from a reference ambient air quality monitoring station during two time periods, the first in fall (October to November) and the second in winter (January to February). Open numerical model data were obtained during these periods from the System for Integrated modeLing of Atmospheric coMposition (SILAM). The algorithms are also demonstrated on two IoT sensors located 60 m and 700 m from the reference station. This work is significant because it offers a computationally lightweight approach to sequentially assimilate station, sensor, and numerical simulation data that do not have prior uncertainty estimates. The proposed method can be applied to impute missing data, to improve the reporting accuracy of air quality observations, and to provide missing uncertainty estimates.

Keywords Ambient air quality · Data assimilation · Environmental monitoring · Internet of things · Open data

1 Introduction

Ambient air pollution poses a persistent and growing threat to human health. In cities, ambient air pollution is monitored using ground station observations of the concentration of atmospheric gases and particulate matter. The most commonly monitored parameters are sulfur dioxide (SO₂),

nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃), and particulate matter (PM_{2.5} and PM₁₀) [1]. These parameters are also often modeled over a wide range of temporal and spatial resolutions, where a detailed classification is provided in [2]. The European Union (EU) requires continuous monitoring of air pollution, and the assessment of air quality is performed using a combined metric of air pollutants [3, 4]. Thus, both modeled and ground station observations are key to assessing health risks and exposure levels and to localize, quantify, and mitigate potentially harmful sources of ambient air pollution [5].

Urban ambient air pollution is challenging to monitor and model, primarily due to the chemical and meteorological interactions that occur on multiple temporal and spatial scales [6, 7]. The modeling of ambient air pollution is most commonly carried out using computational fluid dynamics models that include specialized solvers for chemical transport and near-source dispersion [2]. These models can provide forecasts with spatial resolutions ranging from several

✉ Lizaveta Miasayedava
lizaveta.miasayedava@taltech.ee

Jaanus Kaugerand
jaanus.kaugerand@taltech.ee

Jeffrey A. Tuhtan
jeffrey.tuhtan@taltech.ee

¹ Research Laboratory for Proactive Technologies, Tallinn University of Technology, Ehitajate tee 5, Tallinn 12616, Estonia

² Department of Computer Systems, Tallinn University of Technology, Ehitajate tee 5, Tallinn 12616, Estonia

meters to several hundreds of kilometers and temporal resolutions ranging from minutes to several years [8]. Air pollution changes rapidly over short distances due to micro-meteorological and anthropogenic conditions and includes static and moving sources, including industrial emissions and traffic [9, 10]. As a result, the complexity of building accurate local models with high spatial and temporal resolution remains a substantial challenge.

Air pollution modeling is complemented by in situ ground station monitoring and remote sensing using satellites to retrieve optical aerosol signatures [11, 12]. Calibrated air quality monitoring stations generate data of the highest precision and accuracy, but they are expensive to purchase and maintain. On the contrary, there is a growing number of fixed and mobile low-cost air pollution sensors. However, the performance of low-cost sensors now common for IoT-based platforms varies depending on the specific make and model. A major downside of IoT sensors is that they currently require frequent calibration and maintenance to provide the needed accuracy and completeness of the data for ambient air quality monitoring [10, 13, 14]. Satellites provide gridded data products at resolutions from meters to kilometers in space and from minutes to weeks in time [15]. However, the availability and accuracy of satellite data are often negatively impacted by cloud cover and can vary widely, depending on the post-processing algorithms implemented [16].

To improve the quality of data from single sources, data are combined using data fusion and assimilation methods, referring to both concepts simply as “assimilation” in the remainder of this work. These methods are widely applied in geoscience and engineering for optimal state estimation. This is done to minimize errors between models and measurements by weighting their contributions based on uncertainty estimates [17, 18]. Kalman filters, variational (3DVar, 4DVar), ensemble, and hybrid methods are commonly used assimilation methods, and the choice of methods depends largely on the available observations and model data. The authors of [18] and [19, 20] provide a comprehensive overview of data assimilation methods, including their formulations and limitations, with a focus on air pollution data.

An overview of air pollution assimilation using data from numerical models, satellites, stations, and sensors is provided in [19, 21]. A wide body of literature exists using numerical model simulations (e.g., chemical transport, urban dispersion, microscale models) with in situ measurements from low-cost sensors as well as high-quality monitoring stations for different air pollutants (see, for example, PM_{2.5} [22], NO₂ [10, 23], PM₁₀ [24, 25], O₃ [26]). Where in situ measurements demonstrate a reasonable correlation with satellite aerosol optical depth observations [12, 27], they can be used to create more accurate maps (see, for example, for PM_{2.5} [16, 28], CO₂, O₃, CO, NO_x, SO₂, and HCHO [29]).

To further improve maps, spatial interpolation approaches, including kriging and inverse distance weighting methods [10, 23, 30–32] and regression analysis [12, 33], can also be applied.

A common feature of data assimilation methods is that the contribution of each data source is weighted by its uncertainty estimate [34]. Therefore, current data assimilation methods make use of available uncertainty estimates. The estimates themselves can be evaluated in several different ways and can have a substantial impact on the overall data assimilation results [7, 9, 24].

The main objective of this work is to improve ambient air quality monitoring by assimilating observation/simulation sources without existing uncertainty estimates. The main challenge of this work is to develop a method to provide uncertainty estimates for the open access air quality model and monitoring data. A second challenge is to assimilate these sources considering the large differences in spatial and temporal scales between the modeled and monitored parameters. To address these challenges, we propose a data-driven method to estimate unknown uncertainties. Once the uncertainty estimate has been obtained, we demonstrate that a lightweight least-squares data assimilation algorithm can substantially improve urban ambient air pollutant model estimates at the global to mesoscale. Compared to batch assimilation analogs, recursive methods are more computationally efficient [35], which is key when considering the rapidly growing number of low-cost ambient air quality sensors integrated into the Internet of Things (IoT) [10]. Our approach makes use of open access numerical model results obtained from the System for Integrated modelLing of Atmospheric coMposition (SILAM) [36], ground station monitoring data provided by the Estonian Environmental Research Center, and IoT sensor data from the Tallinn smart city [37].

The major contribution of this work is a computationally lightweight uncertainty estimation method suitable for low-power microcontroller-based IoT systems for the robust assimilation of observational and model time series data at different spatial scales.

2 Background

Given a set of continuous time series observations and numerical simulations which contain errors from multiple sources, data assimilation provides an analysis estimate x_a , of the true state x_{true} at a given point in time based on the error statistics of data sources. The absolute value of the analysis error $|e_a| = |x_a - x_{true}|$ is minimized by solving an optimization problem [38]. The analysis estimate x_a is found by estimating the correction δx of the prior estimate of the true state or by using the background estimate, which is most commonly the numerical simulation data (x_m). Observations

x_{obs} are then used to minimize $|\epsilon_a|$: $x_a = x_{true} + \epsilon_a = x_m + \delta x$. In data assimilation, the following errors are usually considered [38]:

- Background errors $\epsilon_m = x_m - x_{true}$;
- Observation errors $\epsilon_{obs} = x_{obs} - H(x_{true})$ (between measurements and the operator modeling the measuring device);
- Analysis errors $\epsilon_a = x_a - x_{true}$.

Background and observation errors include model simulation errors, the errors which are attributed to the measuring device, discretization, and other representation-related errors. The measuring device is modeled using the operator $H(x_{true})$.

Since x_{true} is unknown, the derivation of the equation for δx depends on several assumptions regarding the optimal state estimate. Most often, each error from ϵ_{obs} , ϵ_m , and ϵ_a is modeled probabilistically as a random variable with a known probability density function (PDF). To reduce the computational cost of calculating PDFs, the most typical error models assume Gaussian distributions described with means, standard deviations, or covariances. Background ϵ_m and observation ϵ_{obs} errors are assumed to be unbiased with zero mean Gaussian distributions and uncorrelated, with zero correlation and covariance [38].

Thus, at each time point t , $\epsilon_{obs} = \epsilon_{obs}[t]$, $\epsilon_m = \epsilon_m[t]$, and $\epsilon_a = \epsilon_a[t]$ are assumed to be fully described by the parameters of a Gaussian distribution: $\sigma_{obs} = \sigma_{obs}[t]$, $\sigma_m = \sigma_m[t]$, and $\sigma_a = \sigma_a[t]$. In this configuration, each $\sigma = \sigma[t]$ can be used to quantify the range of possible values of $x = x[t]$: $[x - \sigma; x + \sigma]$ with different levels of confidence characterizing the uncertainty of x . Therefore, the σ of each data source can be used to estimate the contribution to the analysis estimate needed for data assimilation.

Measurement uncertainties are commonly evaluated as standard deviations of distributions of repeated measurements (type A), or using other information provided in certificates, specifications, and other sources (type B) [39, 40]. The uncertainty of measurements and numerical model simulations can be quantified in a forward way considering each possible source of uncertainty. However, for most practical applications, this remains challenging [41] because there are a large number of sources of uncertainty. The sources include the limitations of measurement devices (noise, systematic errors), discretization, linearization, finite-precision arithmetic, reduction to finite-dimensional problems, incomplete or simplified mathematical models of physical processes and their numerical representation, model parameters, data and metrics used for their calibration and evaluation, interpolation, and human errors. Each modeling assumption, such as the structure of a model and its parameters, introduces an additional source of uncertainty. Furthermore, there are unknown sources of uncertainty and known uncertainties which simply lack data for evaluation.

When there is not enough information about measuring systems, models, and associated uncertainties, the distributions of errors to quantify uncertainty are estimated from the data by building a forecasting model and finding errors as differences between obtained values and their forecasts (e.g., observations and forecasts of observations). Examples of data-driven methods to estimate error using autocovariance and least squares are discussed in [42, 43].

3 Related Work

3.1 Data Sources

To demonstrate the performance of the two proposed lightweight assimilation algorithms, we assimilated CO, NO₂, O₃, SO₂, PM_{2.5}, and PM₁₀ air pollution data from two open data sources: station measurements (observations) and numerical simulations (model data). As an analog to station measurements, we demonstrated the use of two IoT PM₁₀ sensors as observations. The station measurements obtained from the Liivalaia air quality monitoring station located in the city center of Tallinn [24°46'E; 59°26'N], Estonia. The station uses automatic analyzers for air pollutant concentrations: Horiba APNA-360 for NO₂, Horiba APSA-360 for SO₂, Horiba APOA-360 for O₃, Horiba APMA-360 for CO, Met One BAM 1020 for PM_{2.5}, PM₁₀ [44]. The two IoT sensors used for PM₁₀ measurements are located 60 m [24°45'E, 59°25'N] and 700 m [24°44'E, 59°25'N] away from the Liivalaia station. The IoT sensors for air quality use Plantpower PMS7003 particle concentration sensor (cost of ca. 25 USD) for PM₁₀ measurements. These sensors are attached to street lighting poles at a height of 3.5 m and use rechargeable batteries together with a solar panel for powering the sensor.

The model data were obtained from SILAM—System for Integrated modeLing of Atmospheric coMposition, which is a global-to-mesoscale atmospheric dispersion model [36] developed and maintained by the Finnish Meteorological Institute. SILAM provides a 4-day air pollutant forecast with 1-h time intervals for global, European, Northern European, and South-East Asian regions. The model consists of Eulerian and Lagrangian transport routines, 8 chemico-physical transformation modules, and 3- and 4-dimensional variational data assimilation (3DVar, 4DVar) modules [45, 46].

The open model data were exported from the publicly available archives [44–46]. The Liivalaia station provides hourly fixed location [24°46'E; 59°26'N] measurements, whereas SILAM simulations are hourly estimates obtained from a 0.2° grid cell [24°36'E–24°48'E; 59°24'N–59°36'N]. The data used for the validation were exported in the period from 12.10.2021 01:00:00 to 10.11.2021 18:00:00 (“Fall”) and 27.01.2022 01:00:00 to 25.02.2022 15:00:00 (“Winter”). The IoT data were

retrieved from the owner [37]. The PM10 sensors from both fixed locations provide measurements every 10 min, and the hourly averages for the “Fall” period were calculated from them and used to demonstrate the assimilation algorithms.

3.2 Procedures

There are two basic approaches to data assimilation: sequential, which considers only past observations and is most commonly used for real-time systems, and non-sequential, which performs reanalysis of the observations from future time steps. Sequential and non-sequential assimilation can be performed in batches or continuously. They differ in their numerical costs, performance, and suitability for real-time data assimilation [38]. The focus of this work is on low-cost sensor networks which must be capable of processing data in real-time [35], and for this reason, we will use sequential, continuous univariate data assimilation.

Given the probabilistic representation of errors, we further assume that the correction δx is linearly dependent on $x_{obs} - H(x_m)$ and that the operator $H(x_{true})$ is linear. The optimal state is found by minimizing the variance σ_a^2 , which can be implemented using the least-squares analysis. Since background and observation errors are assumed to be of Gaussian distribution, this method is equivalent to maximum likelihood analysis (an excellent derivation is provided in [34, 38]).

Least-squares data assimilation [38] can be considered as one of the simplest and most computationally lightweight data assimilation algorithms. In this work, we will consider three cases of least-squares data assimilation of sources of the same temporal scales: when the prior uncertainties are known (DA1, Fig. 1a), when the prior uncertainties are unknown, the data sources are of the same spatial scales and do not require calibration (DA2, Fig. 1b), and when the prior uncertainties are unknown, the data sources are of different spatial scales and require calibration to the spatial scale of interest (DA3, Fig. 1c).

DA1 is the standard least-squares data assimilation algorithm which requires uncertainty estimates (errors or their distribution parameters). Direct estimation of error distribution parameters requires repeated measurements at each time point t , which is often not possible when using open data. Instead, depending on the data, estimating the errors from the mean within windows, rolling windows by batches or recursively [38], which frequently results in non-Gaussian distributions and consequently suboptimal error estimation using least-squares data assimilation.

To address the challenge, we propose algorithms DA2 and DA3 for two univariate open data sources with missing uncertainty estimates. Both DA2 and DA3 algorithms do not require prior uncertainty estimates; they are sequentially estimated from the provided data streams. Both algorithms are suitable for the data of the same temporal scales.

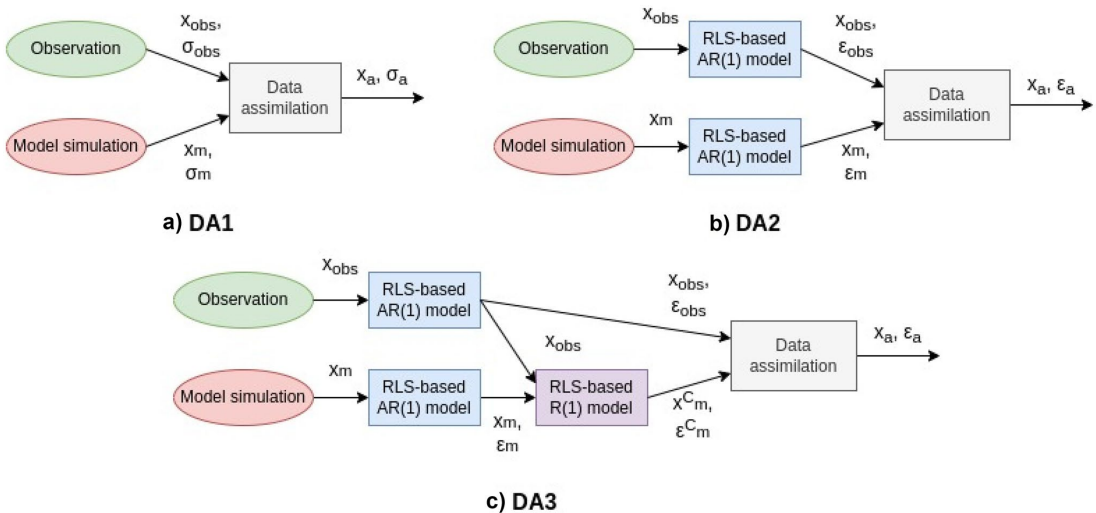


Fig. 1 Data assimilation algorithms: **a)** DA1: standard least-squares data assimilation of a measurement x_{obs} with uncertainty σ_{obs} and model simulation x_m with uncertainty σ_m , both data sources have the same temporal and spatial scales; **b)** DA2: least-squares data assimilation of data sources x_{obs} and x_m of the same temporal and spatial scales with missing uncertainty estimates. The uncertainties are esti-

mated as 1-order autoregression AR(1) modeling errors ϵ_{obs} and ϵ_m ; **c)** DA3: proposed least-squares data assimilation of data sources x_{obs} and x_m of the same temporal and different spatial scales with missing uncertainty estimates. In contrast to DA2, in DA3, x_m is calibrated to the spatial scale x_{obs} with a 1-order linear regression model R(1)

In this work, we demonstrate the performance using hourly data sources at the same site during the fall and winter periods. The difference between the algorithms is that DA2 implements least-squares data assimilation for data sources of the same spatial scales, whereas DA3 can be applied at different spatial scales.

When data sources are of different spatial scales, before applying the least-squares data assimilation, the data should be translated to the same scales. The proposed algorithms allow the translation to the scale of one of the data sources. Thus, when observations from ground stations and numerical models are assimilated, the output of the data assimilation can be obtained at the spatial scale of the observations (from a fixed location) or at the spatial scale of the model. To implement the translation, we suggest calibrating one data source to the other. For example, the DA3 algorithm in Fig. 1c demonstrates a scheme to calibrate a model simulation to the spatial scale of the observations. The choice of the scheme depends on the user's interest, where the more trusted data source should be preferred.

The calibration is implemented using an RLS-based regression R(1) model that serves as a calibration linear operator. The R(1) model is fitted using the RLS algorithm, a data value to be calibrated as input and a data value of a spatial scale of interest as output. The regression relationships obtained in the previous step are applied for the prediction; the predictions are used instead of the input values of a calibrated data source.

Similarly to AR(1) model predictions, R(1) model predictions are used to estimate uncertainty as errors. However, since the input of the R(1) model already has estimated AR(1) uncertainty, we apply the rules of uncertainty propagation to consider not only the error of R(1) modeling as uncertainty but also AR(1) uncertainty.

4 Methods

4.1 DA1: Least-squares Data Assimilation with Known Uncertainty

The procedures of the "Data Assimilation" block for all DA algorithms used in this work, shown in Fig. 1, can be described with the equations of the continuous sequential least-squares data assimilation. In this section, we begin by assuming that the uncertainties, σ , are known.

When both data sources are of the same scales, at each time step t , the analysis estimate $x_a = x_a[t]$ is equivalent to the weighted average of the data sources $x_{obs} = x_{obs}[t]$ and $x_m = x_m[t]$ [34] and can be written as

$$x_a = k \cdot x_{obs} + (1 - k) \cdot x_m, \tag{1}$$

where x_{obs} - observation (measurement), x_m - background (model) estimate (simulation), x_a - analysis estimate, k - coefficient characterizing the contribution of x_{obs} to x_a .

The coefficient $k = k[t]$ in Eq. (1) is derived from the least-squares algorithm by minimizing $\sigma_a^2 = \sigma_a^2[t]$, which by the rules of uncertainty propagation [34] can be given by

$$\begin{aligned} \sigma_a^2 = & \left(\left| \frac{\partial x_a}{\partial x_{obs}} \right| \cdot \sigma_{obs} \right)^2 + \left(\left| \frac{\partial x_a}{\partial x_m} \right| \cdot \sigma_m \right)^2 + \\ & + 2 \cdot \left| \frac{\partial x_a}{\partial x_{obs}} \right| \cdot \left| \frac{\partial x_a}{\partial x_m} \right| \cdot cov_{obs,m}. \end{aligned} \tag{2}$$

Under the assumption of uncorrelated errors ($cov_{obs,m} = 0$), Eq. (2) can be simplified as

$$\sigma_a^2 = (k \cdot \sigma_{obs})^2 + ((1 - k) \cdot \sigma_m)^2, \tag{3}$$

where $\sigma_{obs} = \sigma_{obs}[t]$ - uncertainty estimate of x_{obs} , $\sigma_m = \sigma_m[t]$ - uncertainty estimate of x_m , $cov_{obs,m} = cov_{obs,m}[t]$ - error covariance of x_{obs} and x_m , $\sigma_a = \sigma_a[t]$ - uncertainty estimate of x_a .

Then, the coefficient k is found from solving $\frac{\partial \sigma_a^2}{\partial k} = 0$ as follows:

$$k = \frac{\sigma_m^2}{\sigma_m^2 + \sigma_{obs}^2}. \tag{4}$$

When data sources are of different scales, they should be calibrated to the scale of interest. In time, this can be the mapping of hourly observations to observations recorded each minute or the averaging of observations recorded each minute to hourly observations. In space, this can be a match between data of a lower and higher spatial resolution (e.g., local observations and numerical model grid cells). A mismatch between scales in the observations and model field results in the representation error between observations and model simulations [47].

For example, when we consider 2 data sources of the same temporal scales but different spatial scales and assume that the scale of interest is the scale of x_{obs} , x_m should be calibrated to the scale of x_{obs} with the operator H . Then, Eq. (1) for $H(x_m) = x_m$ (without calibration) can be generalized as follows:

$$x_a = x_m + \delta x = x_m + k \cdot (x_{obs} - H(x_m)), \tag{5}$$

which can be interpreted as the correction δx of the background (model) estimate x_m .

The operator $H(x_m)$ is used to linearly map model simulations x_m of a lower spatial resolution to the resolution of observations x_{obs} to calibrate their mismatch in spatial scales. In Fig. 1, DA2 corresponds to the data assimilation without calibration and DA3 with calibration, where the operator H is sequentially fitted from the provided data.

4.2 DA2: Least-squares Data Assimilation with Unknown Uncertainty Without Calibration

In this work, we estimate the unknown uncertainty from the data in an inverse way. To do this, we apply a simple linear autoregressive model structure of order (lag) 1 (AR(1)), fitting the model coefficients sequentially from the data in each step, and estimate the errors by taking a difference between the AR(1) prediction (using the model fitted in a previous step) and the observation value.

The approach results in a simple and lightweight strategy to quantify changes in the data through the obtained AR(1) modeling errors (corresponds to the “RLS-based AR(1) model” in Fig. 1). In the probabilistic representation, the AR(1) modeling errors would correspond to standard deviations σ of zero-mean Gaussian distributions. We will use a first-order AR(1) filter as follows:

$$\mathbf{x}[t] = w_1 \cdot \mathbf{x}[t - 1] + w_0 + \epsilon[t], \quad (6)$$

where $\mathbf{x}[t]$ and $\mathbf{x}[t - 1]$ - state estimates at time t and $t - 1$ correspondingly, w_0 and w_1 - coefficients of the AR(1) model (fitted at time $t - 1$: $w_0 = \mathbf{w}_0[t - 1]$ and $w_1 = \mathbf{w}_1[t - 1]$), $\epsilon[t]$ - AR(1) modeling error.

AR models are usually used to predict future values using only the past values. In this case, we use the assumption that the state can be modeled as a general linear process and as a hidden Markov chain. Each state estimate at time step t is assumed to be conditioned on the most recent state estimate at time $t - 1$ (lag 1) and quantifies how well a linear model fitted at time $t - 1$ is capable of predicting the state at time t (by finding $\epsilon[t]$). Since w_0 and w_1 are fitted based on the error ϵ , the larger the error, the more model coefficients are modified. The changes of regression relationships over time are often used to quantify the stability of a process [48].

The coefficients of the AR(1) filter are the linear regression model coefficients estimated recursively using the recursive least squares (RLS) algorithm commonly used for optimal state estimation in adaptive filters for sequential data assimilation. In real-time implementation, recursive methods have advantages over stage-wise methods in more rapid convergence and no requirement for direct matrix inversion [35, 49]. The implementation of the 1-order RLS-based regression model is presented in Algorithm 1.

Algorithm 1 RLS algorithm

P - covariance matrix, w - vector of the 1-order linear regression model coefficients, ϵ - regression modeling error.

procedure INIT()

$$P = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$w = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\epsilon = 0$$

end procedure

procedure PREDICT(x)

$$X = (1 \ x)$$

return $X \cdot w$

end procedure

procedure UPDATE(x, y)

$$X = (1 \ x)$$

$$\alpha = y - X \cdot w$$

$$g = P \cdot X^T / (1 + X \cdot P \cdot X^T)$$

$$\epsilon = |\alpha|$$

$$w = w + g \cdot \alpha$$

$$P = P - g \cdot X \cdot P$$

end procedure

procedure RLS()

INIT()

end procedure

Algorithm 1 begins by initializing *RLS()* (with a constructor *INIT()*) with the following algorithm parameters: a state matrix P (2x2) and a vector of coefficients w (2x1). For the AR(1) model, with each new observation at time t , the parameters P and w are updated with *RLS.UPDATE*(x, y), where $x = \mathbf{x}[t - 1]$ and $y = \mathbf{x}[t]$. The error $\epsilon = \epsilon[t]$ is found by taking the difference between the prediction and the observation.

The AR(1) model can also be used for imputation in the case of missing values. In this case, the procedures are as follows in Algorithm 2.

Algorithm 2 RLS-based AR(1) model

x_{past} - past (previous) value, RLS_x - RLS-based model (see Algorithm 1), ϵ - AR(1) modeling error, t - number of acquired data points.

```

procedure INIT( )
     $x_{past} = None$ 
     $RLS_x = RLS()$ 
     $\epsilon = 0$ 
     $t = 0$ 
end procedure
procedure IMPUTE( )
    if  $RLS_x$  is not updated, only initialised then
        if  $x_{past}$  is None then
            return 0
        else
            return  $x_{past}$ 
        end if
    else
        return  $RLS_x.PREDICT(x_{past})$ 
    end if
end procedure
procedure ESTIMATE( $x_{new}$ )
     $t = t + 1$ 
    if  $x_{new}$  is missing (is None) then
         $x_{corr} = IMPUTE()$ 
    else
         $x_{corr} = x_{new}$ 
        if  $x_{past}$  is not missing (is not None) then
             $RLS_x.UPDATE(x_{past}, x_{corr})$ 
        end if
    end if
     $\epsilon = RLS_x.\epsilon$ 
     $x_{past} = x_{corr}$  // saving for the next iteration
    return  $x_{corr}, \epsilon$ 
end procedure
procedure AR( )
    INIT()
end procedure

```

Algorithm 2 starts with the initialization $INIT()$ of the parameters of the RLS algorithm. Afterwards, we model (6) using the $ESTIMATE()$ procedure. This requires finding the coefficients $w = (w_0, w_1)^T$ using the RLS algorithm using $x[t - 2]$ as input and $x[t - 1]$ as output (perform $RLS.UPDATE()$ with $x[t - 2]$ and $x[t - 1]$). Next, we use w and $x[t - 1] = x_{past}$ to model $x[t]$, and the error of the modeling $\epsilon[t] = \epsilon$ between the modeled $w_1 \cdot x[t - 1] + w_0$ and actual x_{new} is returned and applied to complete the assimilation.

We also use the AR(1) model to fill in the missing values (when x_{new} is missing). For this, we use the modeled $w_1 \cdot x[t - 1] + w_0$ instead of x_{new} and use the error

from the previous modeling $\epsilon[t - 1]$ without updating the model. For several missing values in a row, the last fitted (updated) model is used.

4.3 DA3: Least-squares Data Assimilation with Unknown Uncertainty with Calibration

Compared to DA2, algorithm DA3 adds a procedure “RLS-based R(1) model” to calibrate model simulations to the spatial scale of observations (see Algorithm 3).

Algorithm 3 RLS-based R(1) model

RLS_H - RLS-based model (see Algorithm 1), t - number of acquired data points.

```

procedure INIT( )
     $RLS_H = RLS()$ 
     $t = 0$ 
end procedure
procedure CALIBRATE( $x_m, \epsilon_m, x_{obs}$ )
     $t = t + 1$ 
    if  $RLS_H$  hasn't been updated yet then
         $x_m^C = x_m$ 
         $\epsilon_m^C = \epsilon_m$ 
    else
         $x_m^C = RLS_H.PREDICT(x_m)$ 
         $\epsilon_m^C = |RLS_H.w_1| \cdot \epsilon_m + RLS_H.\epsilon$ 
    end if
     $RLS_H.UPDATE(x_m, x_{obs})$ 
    return  $x_m^C, \epsilon_m^C$ 
end procedure
procedure R( )
    INIT()
end procedure

```

In Algorithm 3, the mismatch between the scales is calibrated using the operator H ($H(x) = h_1 \cdot x + h_2$) within an RLS-based linear regression model R(1).

Algorithm 3 uses the data (or imputations if needed) and errors obtained from both AR(1) models for x_{obs} and x_m to linearly calibrate the mismatch using coefficients w fitted with the RLS algorithm ($H = w$) where x_m is the input and x_{obs} is the output. Thus, at step t , the calibrated prediction is $x_m^C = w_1 \cdot x_m + w_0$, where coefficients $w = (w_0, w_1)^T$ were fitted at step $t - 1$.

To calculate the errors of the calibrated predictions, we apply the rules of the propagation of uncertainty. In Algorithm 3, we represent ϵ_m^C as a sum of the calibration error $RLS_H.\epsilon$ and the scaled $|\frac{\partial H(x_m)}{\partial x_m}| \cdot \epsilon_m = |\frac{\partial x_m^C}{\partial x_m}| \cdot \epsilon_m = |RLS_H.w_1| \cdot \epsilon_m$. Thus, ϵ_m^C can be found as follows:

$$\epsilon_m^C = |RLS_H.w_1| \cdot \epsilon_m + RLS_H.\epsilon. \tag{7}$$

Finally, in Algorithm 3, the prediction x_m^C and error ϵ_m^C are further used instead of x_m and ϵ_m for assimilation.

4.4 Main Algorithm

The overall procedures of the data assimilation algorithms are presented in Algorithm 4.

Algorithm 4 Data assimilation algorithm

AR_{obs} - 1-order RLS-based autoregression model AR(1) for observations x_{obs} , AR_m - 1-order RLS-based autoregression model AR(1) for model simulations x_m , R_m - 1-order RLS-based regression model R(1) for model simulations x_m calibrating them to x_{obs} .

procedure INIT()

$AR_{obs} = AR()$

$AR_m = AR()$

$R_m = R()$

end procedure

procedure ASSIMILATE(x_{obs} , x_m)

$x_{obs}, \epsilon_{obs} = AR_{obs}.ESTIMATE(x_{obs})$

$x_m, \epsilon_m = AR_m.ESTIMATE(x_m)$

if do calibration **then**

$x_m, \epsilon_m = R_m.CALIBRATE(x_m, \epsilon_m, x_{obs})$

end if

$k = \frac{(\epsilon_m)^2}{(\epsilon_m)^2 + (\epsilon_{obs})^2}$

$x_a = k \cdot x_{obs} + (1 - k) \cdot x_m$

$\epsilon_a = \sqrt{(k \cdot \epsilon_{obs})^2 + ((1 - k) \cdot \epsilon_m)^2}$

return x_a, ϵ_a

end procedure

Algorithm 4 includes the steps to estimate the missing uncertainties using the AR(1) models AR_{obs} and AR_m for observations and model simulations (see Algorithm 2 for the implementation details). The R(1) model R_m is used to calibrate model simulations to the spatial scale of the observations (see Algorithm 3 for the implementation details). After the uncertainties are estimated, the standard procedures of least-squares data assimilation are applied. In Algorithm 4, the data assimilation scheme DA3 includes calibration. Otherwise, Algorithm 4 is identical to the assimilation scheme DA2.

5 Results

To demonstrate the performance of the developed algorithms, we assimilated the hourly simulations x_m from SILAM (time series from the grid cell [24°36'E–24°48'E; 59°24'N–59°36'N]) and hourly observations x_{obs} from the Liivalaia air quality monitoring station (time series from the location [24°46'E; 59°26'N]) in the period from 12.10.2021 01:00:00 to 10.11.2021 18:00:00 (“Fall”) and

from 27.01.2022 01:00:00 to 25.02.2022 15:00:00 (“Winter”). The air quality variables used for demonstration are CO, SO₂, PM_{2.5}, NO₂, O₃, and PM₁₀.

x_{obs} and x_m have missing uncertainty estimates at the same hourly temporal scale, but different spatial scales (x_{obs} are obtained from a single-point location, while x_m from the 0.2° grid). Therefore, the spatial scales can be directly calculated. In order to compare the performance of DA2 and DA3 (included in Algorithm 4), we chose to apply both data assimilation schemes.

Figure 2 demonstrates the experiment for DA2 visualizing time series plots of SO₂, NO₂, and PM₁₀ station observations x_{obs} (“Station”), SILAM simulations x_m (“Model”), and assimilated values x_a without calibration (“DA2”) for the Liivalaia air quality monitoring station (Tallinn, Estonia) in fall, October–November 2021 and in winter, January–February 2022.

Considering DA3, we applied Algorithm 3 for the calibration of model simulations to station observations. DA3 time series plots of SO₂, NO₂, and PM₁₀ are shown in Fig. 3 including the station observations x_{obs} (“Station”), SILAM simulations x_m (“Model”), and assimilated values x_a with the calibration of “Model” to “Station” (“DA3 (Model → Station)”).

In both cases in Figs. 2 and 3, the assimilated values are constrained to input values of “Station” and “Model” as their weighted sums. Without calibration, “DA2” values in Fig. 2 depend only on the dynamics of a process encoded in the data of each data source independently, whereas “DA3” values in Fig. 3 use reference values for the calibration. In other words, being calibrated to the spatial scale of interest, the error between the reference values used for calibration and the assimilated values decreases. Thus, when calibrating model simulations to observations, the error between the assimilated values of “DA3 (Model → Station)” and “Station” decreases.

The data assimilation results for SO₂, NO₂, and PM₁₀ are provided in Figs. 2 and 3. A summary of the root mean squared errors (Table 1) and mean absolute uncertainties (Table 2) is provided for both data sources for the fall and winter time periods.

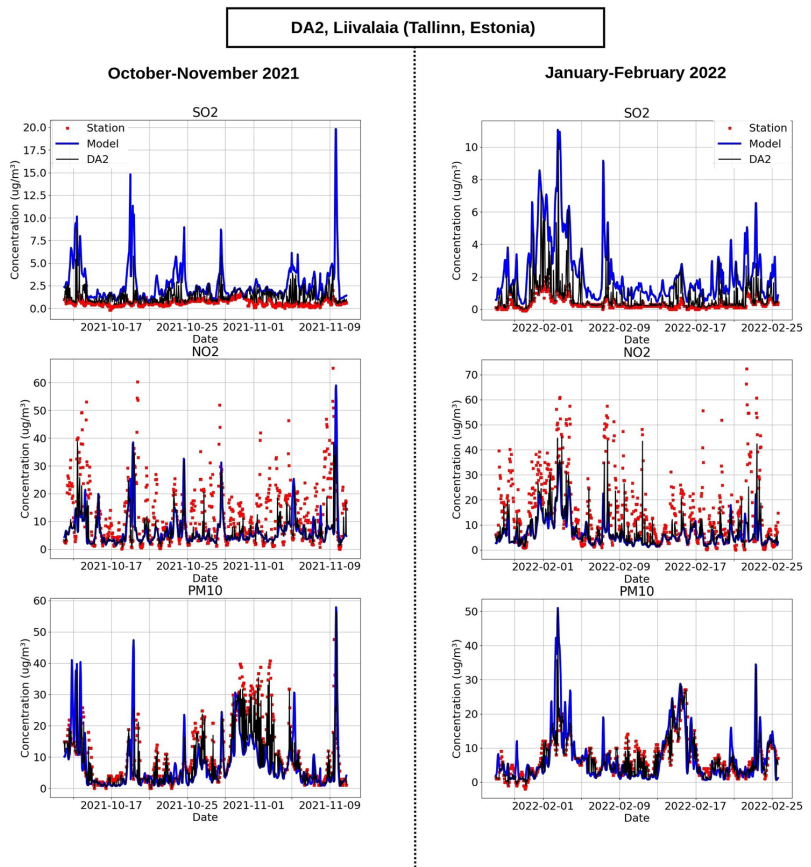
The root mean squared error (RMSE) is calculated pairwise and given by

$$RMSE(\mathbf{x}_1; \mathbf{x}_2) = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_1[i] - x_2[i])^2}, \quad (8)$$

where \mathbf{x}_1 and \mathbf{x}_2 are vectors of data values of length n from 2 data sources. The RMSE values were calculated using Eq. (8) for each of “Station,” “Model,” “DA2,” and “DA3 (Model → Station).”

Table 1 provides the RMSE of station observations and assimilated values without calibration (“Station - DA2”).

Fig. 2 Time series of observations and assimilated values for SO₂, NO₂, and PM₁₀ air quality variables in fall, October–November 2021 and in winter, January–February 2022. “Station” corresponds to observations made by the Liivalaia air quality monitoring station (Tallinn, Estonia), “Model” - SILAM simulations, “DA2” - least-squares data assimilation without calibration



Model simulations and assimilated values without calibration (“Model - DA2”) were found to be lower than the RMSE of station observations as well as the model simulations (“Station - Model”). Unsurprisingly, the calibration of model simulations to station observations resulted in the lowest RMSE values when comparing the assimilated data to the station (“Station - DA3”). Similarly, after calibration of the station observations to the model simulations, the RMSE of model simulations and calibrated assimilated values (“Model - DA3”) was lower than “Station - Model” and “Model - DA2.” However, the RMSE of station observations and calibrated assimilated values (“Station - DA3”) was found to be higher than “Station - DA2.” In summary, calibration results in the minimum error between the reference and assimilated values, but leads to a higher error between the values being calibrated and assimilated values when compared to the uncalibrated data assimilation. This general finding was found to be true for both time periods.

To compare the regression-based uncertainties, we calculated the mean absolute uncertainties (MAU) for each of the station observations, model simulations, and assimilated values as follows:

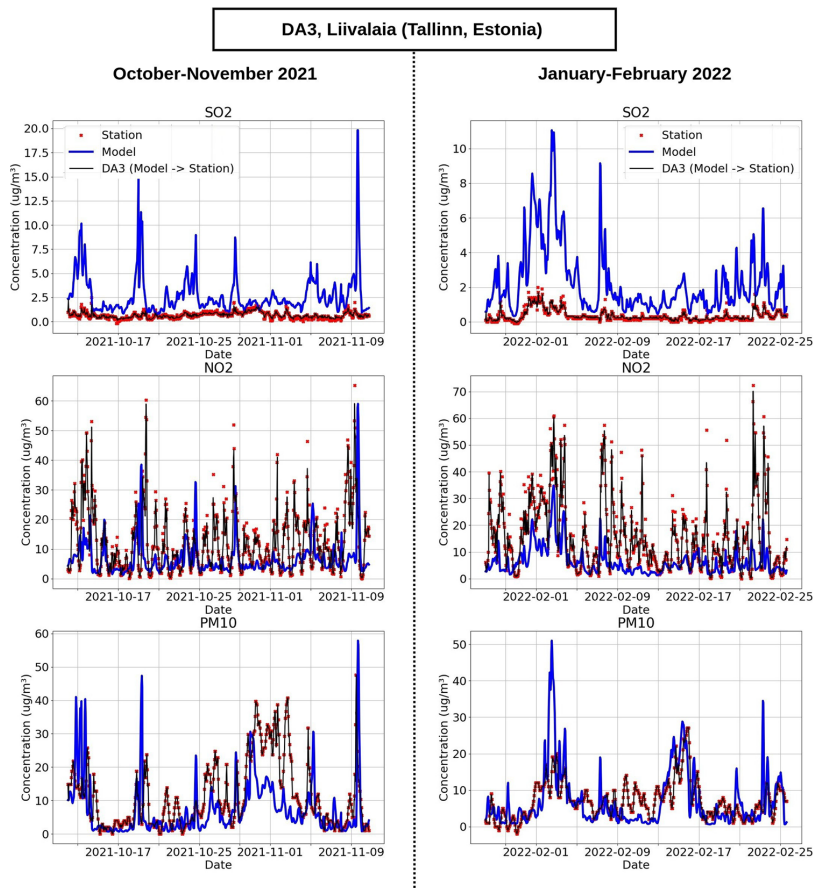
$$MAU(\epsilon) = \frac{1}{n} \sum_{i=1}^n |\epsilon[i]|, \quad (9)$$

where ϵ is a vector of uncertainties of length n .

The MAU values calculated using Eq. (9) for each of the air quality variables for both time periods and are presented in Table 2. “Station” and “Model” are AR(1) uncertainties of station and model data values, and “Model (DA3)” are the uncertainties of calibrated model values. Equation (7) was used in “DA3,” and “DA2” and “DA3” are the uncertainties of the assimilated values with and without calibration.

In Table 2, the uncertainties of the calibrated values “Model (DA3)” are higher than the uncertainties of uncalibrated values for both time periods. After data

Fig. 3 Time series of observations and assimilated values for SO₂, NO₂, and PM₁₀ air quality variables for SO₂, NO₂, and PM₁₀ air quality variables in fall, October–November 2021 and in winter, January–February 2022. “Station” corresponds to observations made by the Liivalaia air quality monitoring station (Tallinn, Estonia), “Model” - SILAM simulations, “DA3 (Model → Station)” - least-squares data assimilation with calibration of “Model” to “Station” data



assimilation, the uncertainty is reduced compared to the uncertainties of single input sources. However, due to the larger uncertainties of the calibrated values, “DA3” uncertainties are also higher than the “DA2” uncertainties. In

general, the lowest uncertainties were obtained using “DA2.”

Similarly to the station data, the data from the IoT sensors can be used as observations. During the “Fall” period,

Table 1 Root mean squared errors (RMSE) between station observations (“Station”), model simulations (“Model”), and obtained assimilated values for both uncalibrated (“DA2”) and calibrated (“DA3”)

data assimilation. “Fall” - fall, October–November 2021, “Wint.” - winter, January–February 2022

Variable	RMSE between station observations (“Station”), model simulations (“Model”), and assimilated values (“DA2” and “DA3”) [$\mu\text{g}/\text{m}^3$] in fall 2021 (“Fall”) and winter 2022 (“Wint.”) at the Liivalaia station (Tallinn, Estonia)				
	Station-Model (Fall/Wint.)	Station-DA2 (Fall/Wint.)	Station-DA3 (Fall/Wint.)	Model-DA2 (Fall/Wint.)	Model-DA3 (Fall/Wint.)
CO	145.41/103.61	75.2/67.71	15.06/21.54	110.39/64.15	145.44/101.05
SO ₂	2.98/2.58	1.01/0.96	0.16/0.11	2.6/2.2	2.97/2.59
PM _{2.5}	6.6/5.04	3.35/2.05	1.03/0.31	5.07/4.17	6.38/5.01
NO ₂	13.23/15.26	11.64/13.64	3.13/3.92	4.77/4.49	11.86/13.76
O ₃	19.41/12.87	13.66/8.67	3.12/3.13	10.3/7.55	18.56/11.9
PM ₁₀	11.04/6.68	7.52/3.37	1.28/0.51	6.75/5.17	10.66/6.64

Table 2 Mean absolute uncertainties (MAU) of station observations (“Station”), model simulations (“Model”), model simulations obtained in the DA2 (“Model in DA2”) and DA3 (“Model in DA3”) algorithms, and of obtained assimilated values for both uncalibrated (“DA2”) and calibrated (“DA3”) data assimilation. “Fall” - fall, October–November 2021, “Wint.” - winter, January–February 2022

Variable	MAU of station observations (“Station”), model simulations (“Model”), and assimilated values (“DA2” and “DA3”) [$\mu\text{g}/\text{m}^3$] in fall 2021 (“Fall”) and winter 2022 (“Wint.”) at the Liivalaia station (Tallinn, Estonia)				
	Station (Fall/Wint.)	Model in DA2 (Fall/Wint.)	Model in DA3 (Fall/Wint.)	DA2 (Fall/Wint.)	DA3 (Fall/Wint.)
CO	15.59/19.65	12.99/10.25	54.88/45.04	4.62/5.48	11.65/14.04
SO2	0.18/0.11	0.37/0.38	0.4/0.21	0.1/0.08	0.13/0.07
PM2.5	0.96/0.39	0.98/0.82	4.31/2.48	0.32/0.21	0.75/0.33
NO2	3.23/3.84	1.07/0.9	9.23/9.99	0.6/0.69	2.43/2.76
O3	4.14/4.01	3.18/2.32	15.02/11.8	1.53/1.36	3.27/3.04
PM10	1.23/0.67	1.31/1.05	7.49/3.54	0.43/0.31	1.03/0.56

the sensor, station, and model time series data are depicted in Fig. 4.

The hourly PM10 data from sensors corresponds to the data from the same SILAM grid cell and as of the station, and therefore, the same data assimilation schemes DA2 and DA3 were applied as shown in Fig. 5.

Similarly to Tables 1 and 2, the summary results of the experiments with sensor data are provided in Tables 3 and 4.

When comparing the station and sensor data sources and the results of their assimilation with the same model data source, the results primarily depend on the difference between the assimilated data sources. Thus, the difference between the model and observation data is lowest for the Liivalaia station, and the errors between the input data sources and the assimilated values for both DA2 and DA3 algorithms are lowest for the station data. Comparing two sensors, the highest error between the model and observations was found to be for the IoT sensor located furthest from the station (700 m). Thus, the errors between the observations and assimilation results are lower for the closer sensor, whereas the error between the model and DA2 results is lower for the farther sensor. This follows intuitively, since DA3 calibrated the model data to the observation data, the errors between the model and DA3 results are reasonably

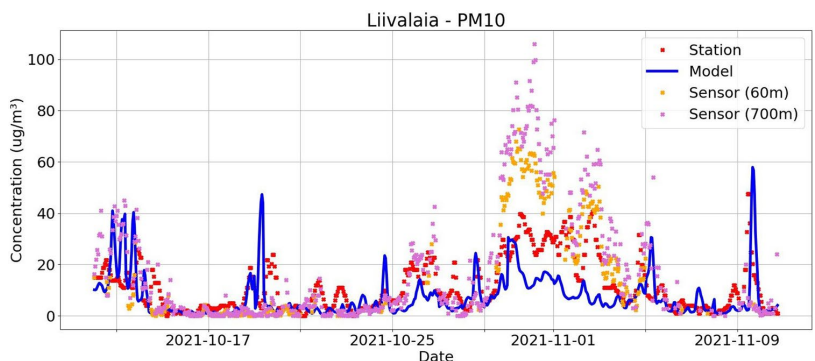
expected to be lowest for the station and highest for the farther sensor.

Using the developed methods to estimate uncertainty, the station observations and the numerical model data calibrated to the station observations were found to have the lowest uncertainty when compared to the sensor observations and the model data calibrated to the sensor observations. As a result, both DA2 and DA3 uncertainties are lowest when using the station data for data assimilation.

At the same time, the closer IoT sensor had a higher uncertainty than the more distant IoT sensor, but since the closer IoT sensor had a lower error from the model data, it allowed calibration of the model data with lower overall uncertainty. Thus, when comparing the assimilation uncertainties of the two sensors, the closer to the station sensor has a higher uncertainty in DA2 but a lower uncertainty in DA3.

The data and results of this work can be reproduced using the open-source Python software developed by the authors and are accessible on GitHub by <https://github.com/effie-ms/rls-assimilation> distributed under the MIT license. The repository includes the scripts of the described algorithms, exported data sources (SILAM time series from the grid cell [24°36'E–24°48'E; 59°24'N–59°36'N] and Liivalaia station (time series from the location [24°46'E; 59°26'N]) SO2, NO2,

Fig. 4 Time series plots of input PM10 data sources in fall, October–November 2021. “Station” corresponds to observations made by the Liivalaia air quality monitoring station (Tallinn, Estonia), “Model” - SILAM simulations, “Sensor (60 m)” - IoT sensor data located 60 m from the Liivalaia station, and “Sensor (700 m)” - IoT sensor data located 700 m from the Liivalaia station



**PM10 in October–November 2021,
Liivalaia (Tallinn, Estonia)**

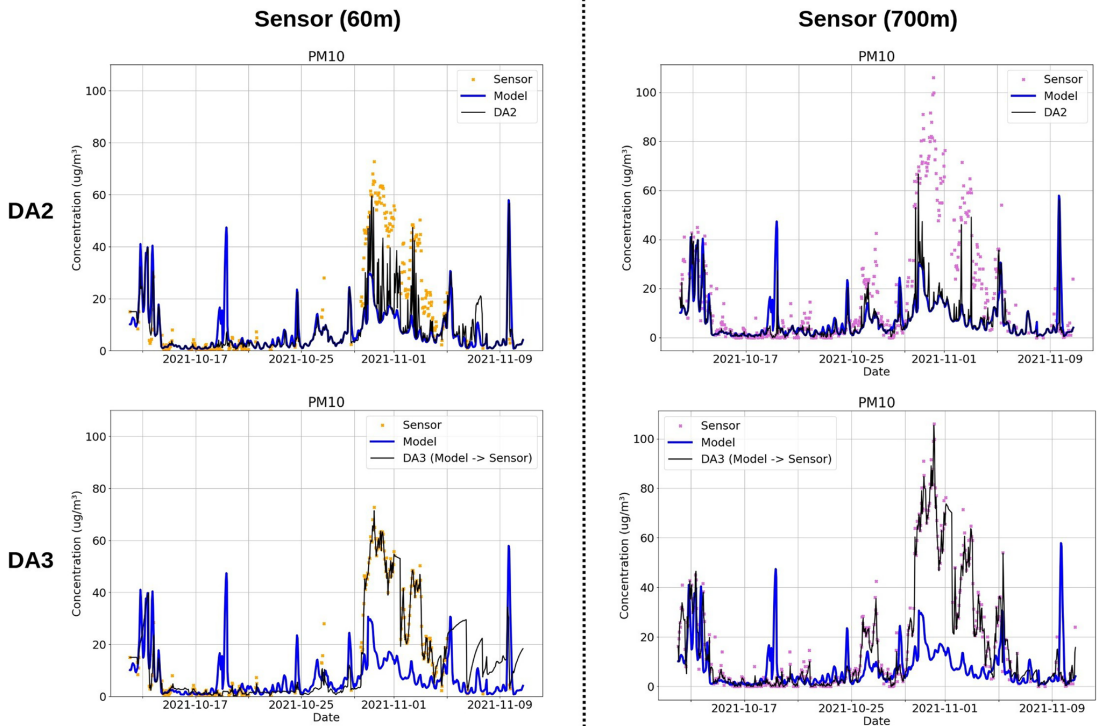


Fig. 5 Time series plots of input sensor data sources and assimilated values for PM10 air quality variable in fall, October–November 2021. “Sensor” corresponds to observations made by one of the IoT sensors located 60 meters or 700 meters away from the Liivalaia air quality

monitoring station (Tallinn, Estonia), “Model” - SILAM simulations, “DA2” - least-squares data assimilation without calibration, “DA3 (Model → Sensor)” - least-squares data assimilation with calibration of “Model” to “Sensor” data.

CO, O3, PM10, PM2.5 air quality data from [44–46] in the period from 12.10.2021 01:00:00 to 10.11.2021 18:00:00 and from 27.01.2022 01:00:00 to 25.02.2022 15:00:00). The IoT PM10 sensor data from the [24°45'E, 59°25'N] and [24°44'E,

59°25'N] locations in the period from 12.10.2021 01:00:00 to 10.11.2021 18:00:00 can be also found in the repository. Please refer to the repository’s README for the description of the repository content and installation instructions.

Table 3 Root mean squared errors (RMSE) between PM10 observations (“Observ.”), model simulations (“Model”), and obtained assimilated values for both uncalibrated (“DA2”) and calibrated (“DA3”) data assimilation in fall, October–November 2021. The observations

are represented by “Station” -Liivalaia station (Tallinn, Estonia) and 2 IoT sensors: “Sensor (60 m)” and “Sensor (700 m)” from the nearby locations

Variable	RMSE between PM10 observations (“Observ.”), model simulations (“Model”), and assimilated values (“DA2” and “DA3”) [$\mu\text{g}/\text{m}^3$] in October–November 2021 (“Fall”) at and nearby the Liivalaia station (Tallinn, Estonia)				
	Observ.-Model	Observ.-DA2	Observ.-DA3	Model-DA2	Model-DA3
Station	11.04	7.52	1.28	6.75	10.66
Sensor (60 m)	16.11	14.25	11.87	6.86	16.16
Sensor (700 m)	20.17	18.24	12.29	6.5	21.28

Table 4 Mean absolute uncertainties (MAU) of PM10 observations (“Observ.”), model simulations (“Model”), and assimilated values (“DA2” and “DA3”) in October–November 2021 (“Fall”) at and nearby the Liivalaia station (Tallinn, Estonia) and 2 IoT sensors: “Sensor (60 m)” and “Sensor (700 m)” from the nearby locations

Variable	MAU of PM10 observations (“Observ.”), model simulations (“Model”), and assimilated values (“DA2” and “DA3”) [$\mu\text{g}/\text{m}^3$] in October–November 2021 (“Fall”) at and nearby the Liivalaia station (Tallinn, Estonia)				
	Observ.	Model in DA2	Model in DA3	DA2	DA3
Station	1.23	1.31	7.49	0.43	1.03
Sensor (60 m)	5.61	1.31	10.14	0.75	2.61
Sensor (700 m)	4.07	1.31	14.09	0.67	3.14

6 Discussion

In order to implement a lightweight data-driven sequential estimation of uncertainty, we used first-order regression modeling. In this work, the uncertainties were estimated using an inverse approach based on the prediction errors of the regression models. The regression models were then fitted sequentially using the recursive least squares (RLS) algorithm. A recursive vector–matrix implementation of the least squares method was applied, in which the parameters of the linear regression model are updated at each step with new observations [35, 49]. Thus, when the data sources have the same spatial scales, we suggest estimating the uncertainty using errors from the RLS-based autoregression AR(1) model predictions. Each data source then has an AR(1) model that recursively models the incoming data and makes predictions based on one previous value. The predictions can be used for imputation if the input value is missing; otherwise, the predictions are used to estimate the uncertainty, and the actual input value is used for data assimilation. The algorithm DA2 uses the same temporal and spatial scales and AR(1) uncertainties as the input for the least-squares data assimilation.

DA2 estimates the uncertainty for each data source independently using only Algorithm 2 to provide a 1-order autoregressive (AR(1)) error. Compared to DA2, DA3 calibrates one of the data sources (e.g., “Model simulation,” “Model”) to the other (“Observation,” “Station,” “Sensor,” also called as “reference”) using Algorithm 3 to perform a 1-order regressive (R(1)) calibration and uncertainty propagation. The reference data source uses the AR(1) uncertainty as input for the final assimilation step in both algorithms.

The calibration suggested in DA3 may be especially useful when data sources have different spatial scales. The calibration

Algorithm 3 offers a fully data-driven approach and does not require any knowledge of the relation between varying scales. Thus, the calibration step in DA3 improves the accuracy at the cost of lower precision. The accuracy of DA3 depends on the accuracy of the ground station because the model data is calibrated using the station data as the ground truth. The precision depends on the dynamics of the data sources for both DA2 and DA3, as well as the changes in the relationship between them (DA3 only). The estimated uncertainties can be potentially helpful to relative comparison and change detection, as changes in time and relationship between data sources result in higher uncertainties. It is worth noting that DA2 generally results in a lower uncertainty than DA3. This is expected as the calibration step biases one data source to the reference source to achieve higher accuracy, at the cost of lower precision.

7 Conclusion

In this work, we propose two lightweight data assimilation methods suitable for real-time execution that are able to provide uncertainty of estimates which are frequently unavailable when using open data. The proposed methods sequentially estimate unknown uncertainty of air pollution data sources of the same temporal and different spatial scales using least-squares data assimilation.

The proposed uncertainty estimation methods are based on sequential 1-order linear autoregression AR(1) and regression R(1) methods and are fitted using the recursive least squares (RLS) algorithm. Specifically, the AR(1) and R(1) models estimate and propagate uncertainty over time as an error from the sequential prediction. Thus, AR(1) prediction errors characterize the uncertainties associated with changes in the observed air quality parameters with respect to the previous values modeled sequentially. The R(1) prediction errors account for differences in the regression relationships between the numerical simulation and ground monitoring station source data sources. To encourage their use by others, we have provided an open-source repository including scripts and test data to reproduce the results obtained in this study. Our hope is that researchers explore the use of these lightweight algorithms, extending their range of application to other urban monitoring networks, using data from a wider range of temporal and spatial scales.

The proposed assimilation algorithms DA2 and DA3 perform sequential data assimilation using data sources of the same and different spatial scales, without (DA2) and after (DA3) calibration. The calibration of one data source to another is derived from the propagation of uncertainty equations. Considering the Tallinn test case, it was found that the minimum uncertainty was achieved using DA2, whereas DA3 provided the minimum error between the ground station measurements after assimilation.

It is worth noting that sequential recursive calculations based on first-order models, linear operators, and single variables were found to substantially reduce the computational effort of the proposed methods. Furthermore, we wish to point out that DA2 and DA3 do not fully quantify the uncertainty but rather provide a standardized method that can be efficiently implemented when simulation and observational data are present.

The proposed assimilation methods were developed to be used by IoT-based devices with limited communication bandwidth and computational power. Air quality IoT sensors are known to vary in agreement with the reference monitors (such as stations) depending on the measured variable, specific make and model of the sensor, and their state in terms of calibration and maintenance [14]. However, they have the potential to increase spatial and temporal coverage of existing air quality monitoring stations and can be used to optimize the station placement of additional locations.

The accuracy of the data assimilation results largely depends on the accuracy of the input data sources. Without calibration, the assimilation results depend only on the uncertainties of the input data sources: the higher the uncertainty of sources, the higher the uncertainty of the assimilated result (DA2). Calibration forces the assimilated result to become dependent on the difference (assumed error) between the input data sources. The higher this difference is, the higher the resulting uncertainty of the calibrated assimilation (DA3).

This study showcases how the growing amount of online open data can be effectively used for ambient air quality monitoring in a typical urban European city environment. This is especially promising, as the cost of building, maintaining, and processing new IoT-based systems for ambient air quality may be substantially reduced at locations where reliable open data already exist. Further research will evaluate the methods at a wider range of European monitoring locations. The methods presented in this work can be implemented in networks of hundreds of distributed low-cost IoT ambient air quality monitoring stations to reduce costs and improve the reliability and accuracy of IoT urban sensors. The results of the data assimilation algorithms proposed in this work depend only on the available data, which makes the algorithms applicable to a broad range of data types. Finally, future research can evaluate spatial maps of the estimated uncertainties to identify locations for optimal sensor placement, reduce the total number of sensors, and optimize the accuracy of the sensor network.

Acknowledgements We would like to thank Thinnect OÜ (<https://thinnect.com/> - the Estonian IoT edge network service provider) for providing the IoT data for experiments.

Author Contribution All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by L.M. The first draft of the manuscript was written by L.M. All authors read and approved the final manuscript.

Funding Lizaveta Miasayedava's contribution to this work was supported by the European Union through European Social Fund Project "ICT programme." The contributions of Jaanus Kaugerand and Jeffrey A. Tuhtan were funded by the project ISC2PT II (Intelligent Smart City and Critical Infrastructure Protection Technologies), funded by the European Regional Development Fund within the framework of the EU Smart Specialisation programme. Jeffrey A. Tuhtan's contribution was also funded in part by the Estonian Research Council Grant PRG 1243.

Data Availability The data used in the study and Python scripts of the described algorithms are accessible via GitHub by <https://github.com/effie-ms/rls-assimilation> and distributed under the MIT license.

Declarations

Ethics Approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflict of Interest The authors declare no competing interests.

References

1. World Health Organization. (2021). *WHO global air quality guidelines: particulate matter (PM_{2.5} and PM₁₀), ozone, nitrogen dioxide, sulfur dioxide and carbon monoxide*.
2. Fallah Shorshani, M., André, M., Bonhomme, C., & Seigneur, C. (2015). Modelling chain for the effect of road traffic on air and water quality: techniques, current status and future prospects. *Environmental Modelling and Software*, 64, 102–123. <https://doi.org/10.1016/j.envsoft.2014.11.020>
3. European Parliament and Council of European Union. (2008). *Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe*.
4. Kotsev, A., Peeters, O., Smits, P., & Grothe, M. (2014). Building bridges: experiences and lessons learned from the implementation of inspire and e-reporting of air quality data in Europe. *Earth Science Informatics*, 8, 353–365.
5. Lee, P., Saylor, R. D., & McQueen, J. T. (2018). Air quality monitoring and forecasting. *Atmosphere*, 9(3), 89.
6. Borrego, C., et al. (2015). Challenges for a new air quality directive: the role of monitoring and modelling techniques. *Urban Climate*, 14, 328–341.
7. Holnicki, P., & Nahorski, Z. (2015). Emission data uncertainty in urban air quality modeling – case study. *Environmental Modeling & Assessment*, 20, 583–597.
8. Weidinger, T., Baranka, G., Makra, L., & Gyongyosi, A. Z. (2010). Urban air quality, long term trends and road traffic air pollution modeling of Szeged. *Urban transport and hybrid vehicles*. IntechOpen.
9. Evans, R. J. (2004). *GEMS: an airborne system for urban environmental monitoring*.
10. Weissert, L., et al. (2019). Low-cost sensors and microscale land use regression: data fusion to resolve air quality variations with high spatial and temporal resolution. *Atmospheric Environment*, 213, 285–295.
11. Cotta, H. H. A., Reisen, V. A., Bondon, P., & Filho, P. R. P. (2020). Identification of redundant air quality monitoring stations using robust principal component analysis. *Environmental Modeling & Assessment*, 25, 521–530.
12. Ben Youssef, K., et al. (2016). Estimation of aerosols dispersion and urban air quality evaluation over Malaysia using MODIS

- satellite. *International Journal of Advanced Scientific and Technical Research*, 3, 229–238.
13. Bartonova, A. et al. (2019). *Low cost sensor systems for air quality assessment*. Tech. Rep. <https://publications.jrc.ec.europa.eu/repository/handle/JRC115379>
 14. Khreis, H., Johnson, J., Jack, K., Dadashova, B., & Park, E. S. (2022). Evaluating the performance of low-cost air quality monitors in Dallas, Texas. *International Journal of Environmental Research and Public Health*, 19(3), 1647. <https://doi.org/10.3390/ijerph19031647>; <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8835131/>
 15. Kleissl, J., Hong, S.-H., & Hendrickx, J. (2009). New Mexico scintillometer network: supporting remote sensing and hydrologic and meteorological models. *Bulletin of The American Meteorological Society*, 90, 207–218. <https://doi.org/10.1175/2008BAMS2480.1>
 16. Shin, M., et al. (2020). Estimating ground-level particulate matter concentrations using satellite-based data: a review. *GIScience and Remote Sensing*, 57, 174–189.
 17. Khaleghi, B., Khamis, A., Karray, F., & Razavi, S. (2013). Multisensor data fusion: a review of the state-of-the-art. *Information Fusion*, 14, 28–44.
 18. Carrassi, A., Bocquet, M., Bertino, L., & Evensen, G. (2018). Data assimilation in the geosciences: an overview of methods, issues, and perspectives. *WIREs Climate Change*, 9(5), e535. <https://doi.org/10.1002/wcc.535>
 19. Hamer, P., Walker, S.-E. & Schneider, P. (2021). *Appropriate assimilation methods for air quality prediction and pollutant emission inversion: an urban data assimilation systems report*. <https://www.nilu.com/pub/1890445/>
 20. Monteiro, A., et al. (2012). Ensemble techniques to improve air quality assessment: focus on O3 and PM. *Environmental Modelling and Assessment*, 18, 249–257.
 21. Handschuh, J., Baier, F., Erbetseder, T., & Schaap, M. (2020). Deriving ground-level PM2.5 concentrations over Germany from satellite column AOD for implementation in a regional air quality model. In A. Comerón, et al. (Eds.), *Remote sensing of clouds and the atmosphere XXV* (Vol. 11531, pp. 5–16). US: SPIE. International Society for Optics and Photonics.
 22. Lopez-Restrepo, S., et al. (2021). Urban air quality modeling using low-cost sensor network and data assimilation in the Aburra Valley, Colombia. *Atmosphere*, 12(1), 91. <https://doi.org/10.3390/atmos12010091>
 23. Schneider, P., et al. (2017). Mapping urban air quality in near real-time using observations from low-cost sensors and model information. *Environment International*, 106, 234–247.
 24. Gressent, A., Malherbe, L., Colette, A., Rollin, H., & Scimia, R. (2020). Data fusion for air quality mapping using low-cost sensor observations: feasibility and added-value. *Environment International*, 143, 105965.
 25. Castell, N., et al. (2018). Localized real-time information on outdoor air quality at kindergartens in Oslo, Norway using low-cost sensor nodes. *Environmental Research*, 165, 410–419.
 26. Sicardi, V., et al. (2011). Ground-level ozone concentration over Spain: an application of Kalman Filter postprocessing to reduce model uncertainties. *Geoscientific Model Development Discussions*, 4, 343–384.
 27. Liu, Y., Sarnat, J., Kilaru, V., Jacob, D., & Koutrakis, P. (2005). Estimating ground-level PM2.5 in the eastern United States using satellite remote sensing. *Environmental Science and Technology*, 39(9), 3269–78.
 28. Ha, S., Liu, Z., Sun, W., Lee, Y., & Chang, L. (2020). Improving air quality forecasting with the assimilation of GOCI aerosol optical depth (AOD) retrievals during the KORUS-AQ period. *Atmospheric Chemistry and Physics*, 20, 6015–6036.
 29. Engelen, R., et al. (2006). *Environmental monitoring of the atmosphere using a 4-dimensional variational (4DVAR) data assimilation system at ECMWF*.
 30. Lin, Y.-C., Chi, W.-J., & Lin, Y.-Q. (2020). The improvement of spatial-temporal resolution of PM2.5 estimation based on micro-air quality sensors by using data fusion technique. *Environment International*, 134, 105305. <https://doi.org/10.1016/j.envint.2019.105305>
 31. Zhong, X., Kealy, A., & Duckham, M. (2016). Stream Kriging: incremental and recursive ordinary Kriging over spatiotemporal data streams. *Computers and Geosciences*, 90, 134–143.
 32. Janssen, S., Viana, P., Fierens, F., Dumont, G., & Mensink, C. (2008). *MERIS AOD and PM 10 in-situ measurements: data fusion in an operational air quality forecast model*. European Space Agency - Special Publication (ESA SP).
 33. Lon, L. (2015). Data fusion of MODIS AOD and OMI AOD over East China using Universal Kriging. *Journal of Geo-information Science*, 10, 1224–1233.
 34. Taylor, J. R. (1982). *An introduction to error analysis*.
 35. Islam, S. A. U., & Bernstein, D. S. (2019). Recursive least squares for real-time implementation. *IEEE Control Systems Magazine*, 39(3), 82–85. <https://doi.org/10.1109/MCS.2019.2900788>. Lecture Notes.
 36. Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., & Kukkonen, J. (2006). A dispersion modelling system SILAM and its evaluation against ETEX data. *Atmospheric Environment*, 40, 674–685.
 37. Thinect. (2019). *Smart city overview*. Retrieved March 27, 2023, from <https://thinect.com/smart-city-overview/>
 38. Bouët, F., & Courtier, P. (1999). *Data assimilation concepts and methods*.
 39. Joint Committee for Guides in Metrology. (2008). Evaluation of measurement data – guide to the expression of uncertainty in measurement. *JCGM*, 100, 1–116.
 40. Damasceno, J. C., & Couto, P. R. (2018). Methods for evaluation of measurement uncertainty. In Anil (Ed.), *Metrology* (Ch. 2). Rijeka: IntechOpen.
 41. Cofta, P., Karatzas, K., & Orlowski, C. (2021). A conceptual model of measurement uncertainty in IoT sensor networks. *Sensors (Basel, Switzerland)*, 21(5), 1827.
 42. Odelson, B. J., Lutz, A., & Rawlings, J. B. (2006). The autocovariance least-squares method for estimating covariances: application to model-based control of chemical reactors. *IEEE Transactions on Control Systems Technology*, 14, 532–540.
 43. Bania, P., & Baranowski, J. (2016). Field Kalman filter and its approximation. *2016 IEEE 55th Conference on Decision and Control (CDC)* (pp. 2875–2880).
 44. Estonian Environmental Research Centre. (2021). *Estonian air quality*. <http://airviro.klab.ee/>
 45. Finnish Meteorological Institute. (2021). *Air quality forecasts*. <https://en.ilmatiiteenlaitos.fi/airquality-forecasts>
 46. Finnish Meteorological Institute. (2021). *SILAM v.5.7: System for integrated modelling of atmospheric composition*. <http://silam.fmi.fi/>
 47. Janjić, T., et al. (2018). On the representation error in data assimilation. *Quarterly Journal of the Royal Meteorological Society*, 144, 1257–1278.
 48. Brown, R. L., Durbin, J. E., & Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37, 149–163.
 49. Young, P. (1974). Recursive approaches to time series analysis. *Bulletin of Mathematical Analysis and Applications*, 10, 209–224.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.

Appendix 3

III

Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Open Data Assimilation of Pan-European Urban Air Quality. *IEEE Access*, 11:84670–84688, August 2023

Received 12 June 2023, accepted 31 July 2023, date of publication 7 August 2023, date of current version 15 August 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3302348

RESEARCH ARTICLE

Lightweight Open Data Assimilation of Pan-European Urban Air Quality

LIZAVETA MIASAYEDAVA^{1,2}, (Graduate Student Member, IEEE),
JAANUS KAUGERAND¹, (Member, IEEE), AND JEFFREY A. TUHTAN^{1,2}, (Member, IEEE)

¹Research Laboratory for Proactive Technologies, Tallinn University of Technology, 12616 Tallinn, Estonia

²Department of Computer Systems, Tallinn University of Technology, 12616 Tallinn, Estonia

Corresponding author: Lizaveta Miasayedava (lizaveta.miasayedava@taltech.ee)

This work was supported by the European Union through the European Social Fund as part of the "Information and Communication Technology (ICT) Programme."

ABSTRACT The number of ambient air quality monitoring stations is growing globally, driven by the need to quantify potential health risks posed by air pollution on urban populations. Reliable, robust and interoperable air quality monitoring requires observations with consistent accuracy and low amounts of missing data. In practice, this is challenging to achieve due to the measurement limitations and complexity of the physical phenomena. Data assimilation methods are widely used to fill missing or faulty observations and improve data quality by combining observations from fixed air quality monitoring ground stations with large-scale numerical models. A further advantage of data assimilation is that it can decrease costs by reusing existing open government data. A key requirement for assimilation is that uncertainty estimates are available for both measurements and model data. However, this poses a major bottleneck for widespread data assimilation with open data because uncertainty estimates are frequently unavailable. Additional challenges addressed in this work include the needs to impute missing data and process observations and model simulation results at different temporal and spatial scales. To address these challenges, we have developed novel, lightweight data assimilation algorithms based on recursive least-squares. The algorithms provide a fully data-driven way to estimate unknown uncertainties by defining the weights of the input data sources using least-squares data assimilation. The lightweight data assimilation algorithms can be executed to update the current state estimate in near real-time scenarios to improve the accuracy, completeness, and precision of the analysis estimate. A sensitivity analysis is conducted using synthetic data based on logistic maps with increasing noise levels. In addition, the proposed assimilation algorithms are applied to large-scale open pan-European air quality monitoring station data. The data were obtained from 86 stations for CO, 593 stations for NO₂, 462 stations for O₃, 137 stations for SO₂, 254 stations for PM_{2.5}, and 445 stations for PM₁₀ in the period from 2022-01-27 01:00:00 to 2022-02-25 15:00:00 from the European Environmental Agency (EEA) and corresponding simulation results from the System for Integrated modeLLing of Atmospheric composition (SILAM, global, version 5.7, FRC forecasts at the surface). The proposed lightweight data assimilation methods were found suitable to improve the completeness (filling in all missing data), accuracy (taken as the RMSE between the assimilation results and ground station observations) and precision for all of the open air quality parameters evaluated in this work. Furthermore, the proposed lightweight assimilation algorithms may also provide new and cost-effective methods to improve the data quality of the growing number of Internet of Things (IoT) urban air quality sensors.

INDEX TERMS Ambient air quality, data assimilation, environmental monitoring, open data, uncertainty quantification.

The associate editor coordinating the review of this manuscript and approving it for publication was Geng-Ming Jiang¹.

I. INTRODUCTION

Cities across the globe rely on urban air quality (AQ) data to develop strategies to reduce emissions, lower the population's

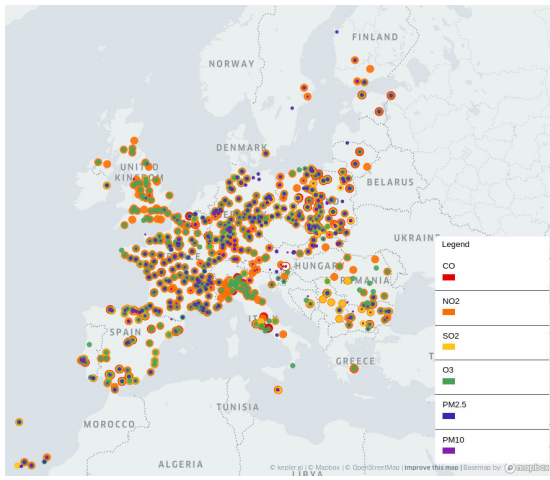


FIGURE 1. Map showing the ground station locations of the open access air quality monitoring network of the European Environment Agency (EEA) [11] used in this work for testing and validation. The map is generated using [12], [13] and [14].

exposure to air pollution and assist in emergency response [1], [2], [3], [4]. Urban ambient AQ monitoring can be performed using ground station observations and numerical simulations. Data assimilation (DA) algorithms combine both sources and can substantially improve the accuracy and spatial coverage of urban AQ monitoring. However, in practice, the widespread use of DA remains highly challenging due to the non-linear dynamics and spatio-temporal complexity of the underlying physical phenomena [5].

The European Environment Agency (EEA) provides an open access, pan-European database of urban AQ monitoring station data (see Fig 1). These data can be combined with numerical models of large-scale systems, including atmospheric, oceanic, and land surface interactions using DA methods [6], [7], [8], [9]. The choice of a particular DA method depends largely on the case-specific observations and models available. Considering AQ data, 3- and 4-dimensional variational assimilation, Kalman and particle filters are the most common. These DA methods solve inverse problems and are thus mathematically similar to machine learning (ML) optimization problems. The main difference between the DA and ML methods is that DA considers observation and model uncertainties [10]. When uncertainties are well characterized, they can be used to reduce the overall uncertainty of the system's state when compared with only observation or model data on their own [5].

DA methods also improve the quality of single-source estimates by imputing the missing values and can increase the accuracy and precision of predictions [5], [9], [10]. High-density AQ monitoring networks are costly to purchase, install and maintain, and therefore they remain scarce [15]. Recent advances in low-cost sensing now allow for the possibility of creating high-density AQ monitoring networks

based on the Internet of fixed and mobile Things (IoT) [15]. Currently, a variety of authors have proposed, developed and tested low-cost AQ sensors [16], [17], [18].

In contrast to previous works focusing on the development and implementation of new IoT-based sensor networks [4], [15], we propose to reuse open government AQ data sources for DA. Our concept has several benefits: it decreases the costs of providing AQ monitoring by applying DA to openly available large-scale numerical models of air pollution transport and dispersion. Moreover, we demonstrate that large-scale numerical model data from open numerical models such as SILAM can be reused without explicit knowledge about the model. In contrast to research performed in [19] and [20], our work takes SILAM numerical model results as a source of continuous spatial and temporal data, which can be used to address a large amount of missing data without uncertainty estimates from the EEA ground station observations. Since numerical models provide globally complete spatial and temporal coverage, they can provide estimates at locations where observations from fixed or mobile AQ monitoring stations are sparse or completely absent. Our proposed lightweight DA methods are tested and validated on a pan-European scale, making them suitable for large-scale mapping and decision-making [1], [21].

The reuse of open data sources for DA is significantly complicated by the missing uncertainty estimates, which are required parameters for all the DA algorithms. In work [22], we elaborate on why it is hardly possible to fully estimate the uncertainty parameters and suggest a workaround by estimating the uncertainty parameters recursively over time from the input data values as regression errors ("regression-based uncertainties") and develop methods for their estimation. The uncertainties are estimated using chained 1-order recursive least squares (RLS) filters representing a 1-order linear regression model the parameters of which are estimated by the RLS algorithm from the observed data. The filters are chained using the rules of error (uncertainty) propagation (as described in [22]).

The reasoning behind this type of uncertainty estimation is as follows: if the behaviour of the system changed at the moment when a prediction should be made, then the model fitted with the RLS algorithm may not give an accurate result. Rather, it would give a result that would be accurate for the system that didn't change. Therefore, for single sources, the RLS filter is not used to predict the current value if it is provided. Instead, the predictions under the assumption of a steady state are used only if the current value is missing. In other words, the imputed values are predicted for a system the behaviour of which didn't change since the last RLS filter update. Instead, we use the errors from the steady-state prediction as an uncertainty estimate. The more a system changed at a certain moment of time, the higher the error from its steady state prediction is. And we claim that this property - the error from the steady state prediction - can be used as a data-driven uncertainty estimate for DA algorithms that use uncertainties to determine the weights of input data sources.

For the fair estimation of weights, the steady state modelling, relative to which the errors are calculated, is suggested to be executed uniformly for all the data sources. This means that the parameters of the RLS filters are suggested to be the same, which would standardize the procedure on a large scale.

We acknowledge that “uncertainty” and “error” are two distinct concepts, and this work does not intend to conflate them but rather demonstrate how the suggested regression-based uncertainties could be used for the assimilation procedures to improve the data quality (accuracy, completeness, precision) of single data sources. We intend to perform DA for univariate streams of air pollution data by applying DA to the most recent (current) value. We do not take into account any other data or information such as information about the SILAM model, distribution patterns of air pollutants, or weather data assuming that it is unknown or unavailable. In this study, we demonstrate that with our algorithms and reuse of open data, it is still possible to improve the data quality of single open data sources only from the timestamped air pollutant data values with their location coordinates, without uncertainty estimates or additional information provided.

In this work, we do not intend to analyze the state of AQ in Europe or validate the reported data. Instead, we provide a method which reuses existing AQ monitoring station data and numerical model data of Europe. As much of the existing ground station data has large amounts of missing data and is without uncertainty estimates, our proposed methods improve the quality of existing European AQ ground station monitoring data by using DA with open numerical simulations. In addition, our proposed methods can be applied on a large number of low-power low-quality IoT-based AQ monitoring stations. This allows for the reuse of open data and may provide higher quality data to local and regional decision-makers to improve the enforcement of European environmental policy objectives.

This work is an extension of our previous work [22], compared to which we add new algorithms demonstrating how DA without known uncertainty estimates can be applied when not only the spatial resolution is different (DA3), but also the temporal resolution of assimilated data sources is different (on the example of hourly and daily values, DA4). We also demonstrate the effect of reusing the previous analysis values on the suggested DA (S-DA and S-DA4), perform sensitivity analysis for a logistic map in different modes and different noise (uncertainty) levels for all the algorithms and validate all the algorithms using the data from urban background stations throughout Europe.

The paper is structured as follows: Chapter II provides an overview of previous works on urban AQ DA. Chapter III describes the methods, the sources of observations and numerical simulation data, the performance evaluation criteria and sensitivity analysis using synthetic logistic map data. Chapter IV presents the results and compares the performance of the DA algorithms, and Chapter V discusses the obtained results with a focus on the effects of spatial and temporal scaling. Finally, Chapter VI provides concrete

suggestions for further applications, improvements, limitations and a future outlook of the proposed lightweight DA methods for open urban AQ monitoring systems.

II. RELATED WORK

Monitoring urban air quality (AQ) commonly involves regression, interpolation and when numerical models are available, data assimilation (DA) of the available data [23]. Within the European Union (EU), AQ time series and maps are frequently generated by assimilation of observation and model data using linear regression models followed by residual kriging [2]. However, the real-time estimation of urban AQ data has substantial computational constraints. Due to these constraints, conceptually and computationally simple, or “lightweight” methods suitable for large spatially distributed data sets as well as for IoT sensors in smart cities have become a focus of AQ assimilation research [23], [24].

Open AQ data often do not include uncertainty estimates which are required inputs in most DA methods [5]. Neglecting uncertainty has led to fallacious risk assessments and incorrect environmental policy decisions [1], [25], [26]. To address the lack of uncertainty data, we set out to create a way to estimate the uncertainty. This poses a substantial challenge, as AQ parameters vary widely over space and time, and the mathematical methods used to quantify uncertainty typically rely on long-term observations from calibrated fixed-station observations [1], [3], [10], [22], [25]. In addition, the classical formulation of the propagation of uncertainty requires sub-models for each system component for bias correction and to account for the underlying variability of the physical measurement processes themselves [27].

To obtain uncertainty estimates, previous works have applied computationally-expensive ensemble methods which perturb model parameters and input data within their uncertainty ranges [1], [3]. A major drawback of ensemble methods is that due to their high computational costs, they remain unsuitable for the generation of real-time air pollution forecasts for large open data sets of varying data quality as well as for low-power IoT devices with limited communication bandwidth and computational power. In general, most DA methods require comprehensive uncertainty models [25], which remain largely unsuitable for computationally constrained IoT devices. To address this, lightweight uncertainty estimation methods using sequential inverse modelling have been proposed to obtain the simple difference between the regressed estimates and the actual values of the ground station observations or numerical models [3], [16], [22].

In our previous work [22], the authors have proposed lightweight least-squares DA (LSDA) regression-based methods to assimilate open observation and numerical model data for a single ground observation station in the Tallinn metropolitan region. The methods impute missing values, estimate uncertainties and provide a linear observation operator to calibrate observation and model data to the same spatial scale. Our previous methods also provide a standardized uncertainty estimate for open ground station

observations and numerical model results which do not include uncertainty data. The current work presents a major advancement in the use of DA for open large-scale AQ monitoring data and includes new algorithms which can cope with multiple temporal and spatial scales [26].

In contrast to our previous work, which made use of a single observation station, we have substantially expanded and improved our previous methods by including hourly pan-European openly available urban background observations obtained from the European Environment Agency (EEA). In addition to increasing the background data to the pan-European scale, we test and validate three new lightweight DA algorithms; sequential single-source DA with unknown uncertainty (S-DA), non-sequential and sequential DA for two data sources of different spatial and temporal scales (DA4 and S-DA4). The AQ monitoring stations used for testing and validation in this work include 86 stations for CO, 593 stations for NO₂, 462 stations for O₃, 137 stations for SO₂, 254 stations for PM_{2.5}, and 445 stations for PM₁₀. The locations are shown in Fig 1), and the observations from the location were assimilated with hourly and daily 0.2° numerical model simulation results obtained from the openly available System for Integrated modeling of Atmospheric composition (SILAM, global, version 5.7, FRC forecasts at the surface).

Contributions: The major contributions of this work are as follows:

- We provide three new algorithms S-DA, DA4 and S-DA4 for the lightweight assimilation of urban AQ data with unknown uncertainty.
- We investigate how sequential estimation affects DA performance at the pan-European scale using openly available EEA and SILAM AQ data.
- We demonstrate how the proposed lightweight algorithms can utilize data sources of higher temporal resolution, using hourly observations, to improve the estimates from lower-resolution data sources based on daily model simulations.
- We validate and illustrate the scalability of the three proposed methods using several hundred AQ monitoring stations of open government observations provided by the EEA and the numerical model, SILAM.

III. METHODS

A. OVERVIEW AND ABBREVIATIONS OF DATA ASSIMILATION METHODS

The algorithms proposed in this work are based on the least-squares data assimilation (LSDA) algorithm. Our primary contributions are to automatically impute missing data, to calibrate the analysis between observations and numerical models with different temporal and spatial scales and to provide uncertainty estimates for datasets with unknown uncertainties. In our previous work [22], the authors have proposed the following algorithms for lightweight DA:

- DA1: LSDA of 2 sources with known uncertainties. This method corresponds to the classic LSDA approach and serves as the basis for the proposed algorithms presented in these works.
- DA2: LSDA with unknown uncertainties using data from two sources. This algorithm requires that both data sources have the same temporal and spatial scales.
- DA3: LSDA with unknown uncertainties using data from two sources. Here, the requirement is that the same temporal scales are used for the two sources, and spatial calibration is applied using an observation operator to assimilate the two sources at different spatial scales.

In the current work, we present three new LSDA-based methods providing substantial improvements over our previous DA2 and DA3 methods:

- S-DA: Sequential LSDA of a single source and its predictions with unknown uncertainty. Compared to DA2 and DA3, S-DA does not require another data source.
- DA4: LSDA with unknown uncertainties using data from two sources of different temporal and spatial scales. Compared to DA2, which requires data of the same temporal and spatial scales and compared to DA3, which requires data of the same temporal and different spatial scales, DA4 allows for the use of data sources with both different temporal and spatial scales.
- S-DA4: Sequential LSDA with unknown uncertainties using data from two sources of different temporal and spatial scales. Compared to S-DA using the source data, S-DA4 uses the assimilation results of DA4 and their predictions.

B. DATA ASSIMILATION WITH UNKNOWN UNCERTAINTY AND DIFFERENT SPATIAL SCALES

The methods developed in work [22] are designed to preprocess data before applying the LSDA algorithm. All the developed preprocessing methods are based on the first-order recursive least squares (RLS) algorithm shown in Fig. 2 (a). For each new data point, RLS sequentially fits the coefficients of a first-order linear regression model w using inputs x_{in} and outputs x_{out} by correcting an initial prediction x_{pred} based on the error ϵ from the actual value x_{out} . The RLS outputs x'_{out} and ϵ' depend on the regression model it was used for.

We suggest applying two RLS-based first-order regression models, as each model is well-suited for different purposes. The first model is an RLS-based first-order autoregression AR(1) model (see Fig. 2 (b)) to estimate initial uncertainties at the given temporal and spatial scales. At time step t , the AR(1) model fits a past value $x[t - 1]$ to a current value $x[t]$. If $x[t]$ is missing, the RLS prediction x_{pred} is used to impute the missing value, otherwise $x[t]$ is used as-is, and the error ϵ from the prediction is taken as the regression-based uncertainty estimate. AR(1) models are applied to each data source.

The second model is an RLS-based first-order regression R(1) model for spatial calibration (see Fig. 2 (c)) of two data

sources that takes the outputs of AR(1) models as inputs and calibrates one data source x_1 to the other data source with different spatial scale, x_2 by RLS-based fitting. The calibrated input value, x_{pred} is used instead of x_1 LSDA, and the AR(1) error ϵ_1 is scaled by the rules of uncertainty propagation and augmented with the R(1) error ϵ .

The DA2 algorithm was based solely on AR(1) models, whereas the DA3 algorithm uses both AR(1) and R(1) models. We found that both models are required when implementing DA3 in order to provide one data source with an additional spatial calibration step. Detailed explanations of the DA1, DA2 and DA3 algorithms including their pseudocode can be found in [22].

C. SEQUENTIAL DATA ASSIMILATION WITH MISSING UNCERTAINTY, DIFFERENT SPATIAL AND TEMPORAL SCALES

In this work, we extend the previously developed methods with temporal calibration and reuse of the previously estimated (at time step $t - 1$) analysis values for DA (see Fig. 3).

Here, we restrict the temporal calibration to hourly and daily values. However, the same approach can be applied to other temporal scales without a loss of generality. As an example, when hourly outputs are desired, at least one of the data sources must be hourly. We also wish to point out that similar procedures can be applied to other temporal scales. If the input temporal scales are hourly and monthly, then the number of hours in a month should be used instead of 24 hours in the recursive average estimator. If the input temporal scales are monthly and daily, then the number of days in a month should be used instead, or both should be transformed into hourly supplied data. In Fig. 3 (a, left), the hourly data are transformed to daily data by recursively obtaining a full-day (24-hour) daily average of values and errors, which is reset every 24 hours. The algorithm for recursive daily averages is further outlined in Algorithm 1.

To transform daily data x_1^d into hourly data x_1^h (see Fig. 3 (a, right)), we suggest fitting an RLS-based first-order model for the hourly data source x_2^h . The input of the model is daily values x_2^d obtained with the recursive daily average estimator $RD()$, and the output is hourly values x_2^h . Afterwards, the coefficients of the model for x_2 are used to predict x_1^h from x_1^d . Similarly to the spatial alignment model, the output uncertainty ϵ_1^h is taken as the simple sum of the scaled input uncertainty ϵ_1^d and the model prediction error ϵ .

The DA outputs (also commonly referred to as analysis values) can be fitted autoregressively to assimilate the analysis predictions with the obtained data. In Fig. 3 (b), we demonstrate the use of an RLS-based AR(1) model for sequential estimation using the outputs of LSDA $x_a[t - 1]$ at the previous time step, $t - 1$ as the input and the analysis value, $x_a[t]$ as the output to predict the next analysis value.

The RLS-based AR(1) model for sequential estimation enables a sequential single-source LSDA as shown in Fig. 4 (a) S-DA algorithm. Furthermore, in this work the

Algorithm 1 Recursive Daily Average Estimator

\bar{x}^h - current average data value, $\bar{\epsilon}^h$ - current average error (uncertainty), x^d - last full-day daily average data value, ϵ^d - last full-day daily average error (uncertainty), N - counter of previous hours (reset after each 24 hour interval).

procedure INIT()

$\bar{x}^h, \bar{\epsilon}^h, x^d, \epsilon^d = 0, 0, 0, 0$

$N = 0$

end procedure

procedure RESET()

$\bar{x}^h, \bar{\epsilon}^h = 0, 0$

$N = 0$

end procedure

procedure UPDATE($x_{new}^h, \epsilon_{new}^h$)

if $N == 24$ **then**

$x^d = \bar{x}^h$

$\epsilon^d = \bar{\epsilon}^h$

RESET()

end if

$N = N + 1$

$\bar{x}^h = \frac{1}{N} \cdot (x^h \cdot (N - 1) + x_{new}^h)$

$\bar{\epsilon}^h = \frac{1}{N} \cdot (\epsilon^h \cdot (N - 1) + \epsilon_{new}^h)$

end procedure

procedure RD()

INIT()

end procedure

S-DA algorithm is compared with the previously suggested DA3 algorithm that uses 2 data sources (see Fig. 4 (b)) for LSDA with respect to the reference data source (data source of spatial scale S).

Models for temporal alignment are integrated into the DA4 algorithm, enabling LSDA with unknown uncertainties including both temporal and spatial calibration. Overall, DA4 is similar to DA3 but adds the temporal calibration (alignment) step after the spatial calibration, as shown in Fig. 5 (a). When used the output of DA4 instead of AR(1)-preprocessed data taken directly from a source, S-DA shown in Fig. 4 (a) is transformed into S-DA4, as shown in Fig. 5 (b).

The performance of DA4 and S-DA4 algorithms is compared by transforming one of the hourly data sources to daily intervals by averaging over a 24-hour period. Afterwards, the original hourly values are assimilated and used as a reference. The daily data are assimilated with the hourly data from the other data source and compared to the hourly assimilation results.

D. PARAMETERS AND SENSITIVITY ANALYSIS

DA algorithms often require continuous data without missing values, uncertainties (error and noise covariance matrices), state transition and observations operators, as well as additional algorithm-specific parameters (e.g. the number of particles for particle filters, number of ensemble members for ensemble filters, among others) [28], [29]. Unfortunately,

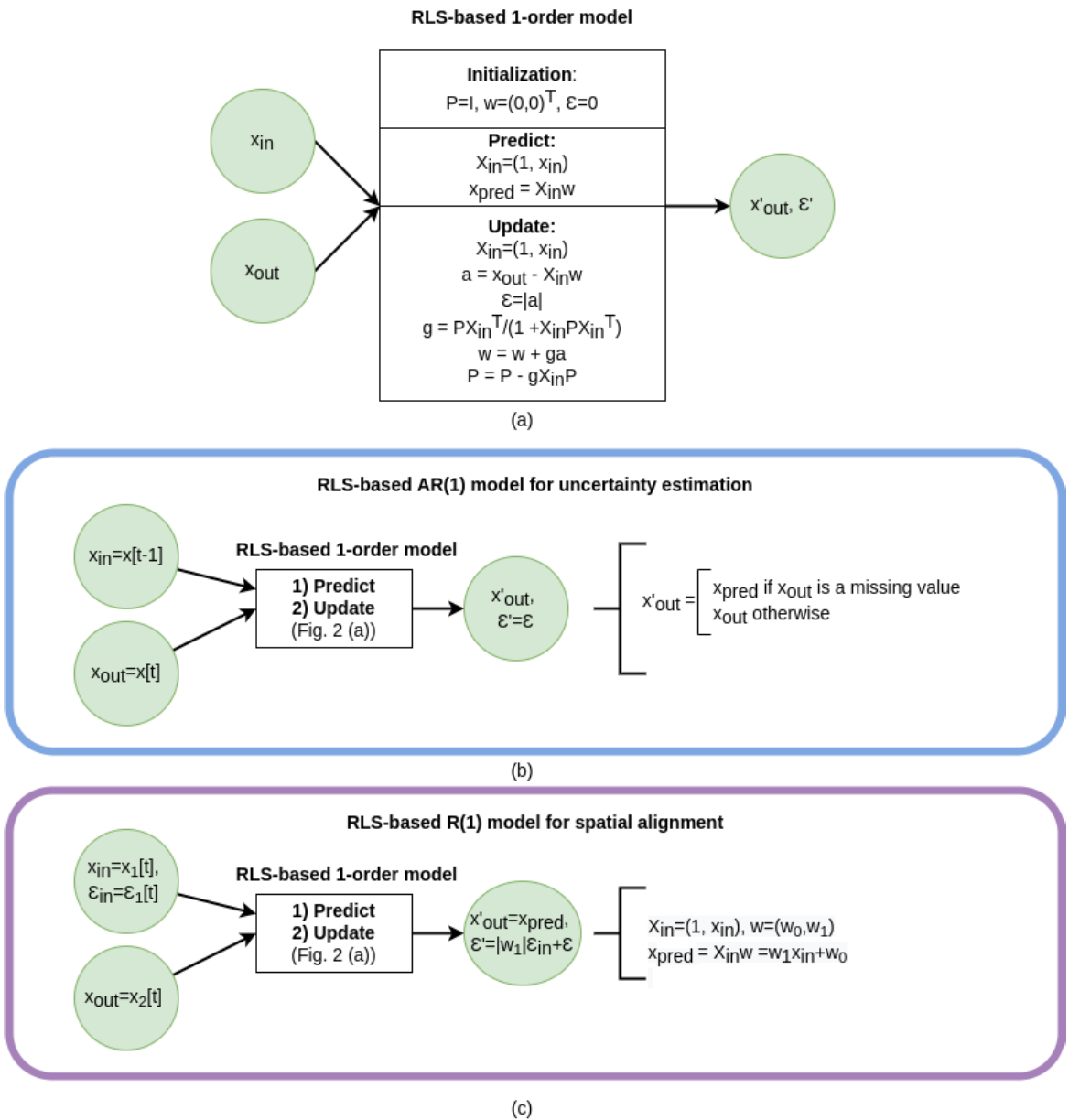


FIGURE 2. Recursive algorithms for least-squares data assimilation (LSDA) from [22]. (a) recursive least squares (RLS)-based first-order model used as a core for data-driven uncertainty estimation and spatio-temporal alignment (calibration). (b) RLS-based first-order autoregression AR(1) model for sequential imputation and uncertainty estimation using the AR(1) model. (c) RLS-based first-order regression R(1) model for spatial alignment (calibration) of x_{in} with the scales of x_{out} , considering the errors x_{in} and ϵ from the AR(1) and R(1) models.

open AQ datasets and IoT sensors provide only the physical parameter values without sufficient information to quickly and efficiently determine the necessary additional parameters to carry out DA [17], [21].

With the goal to enable the use of DA to improve the accuracy, completeness and precision of the single input data sources, we propose methods estimating the uncertainties recursively over time from the input data values as regression

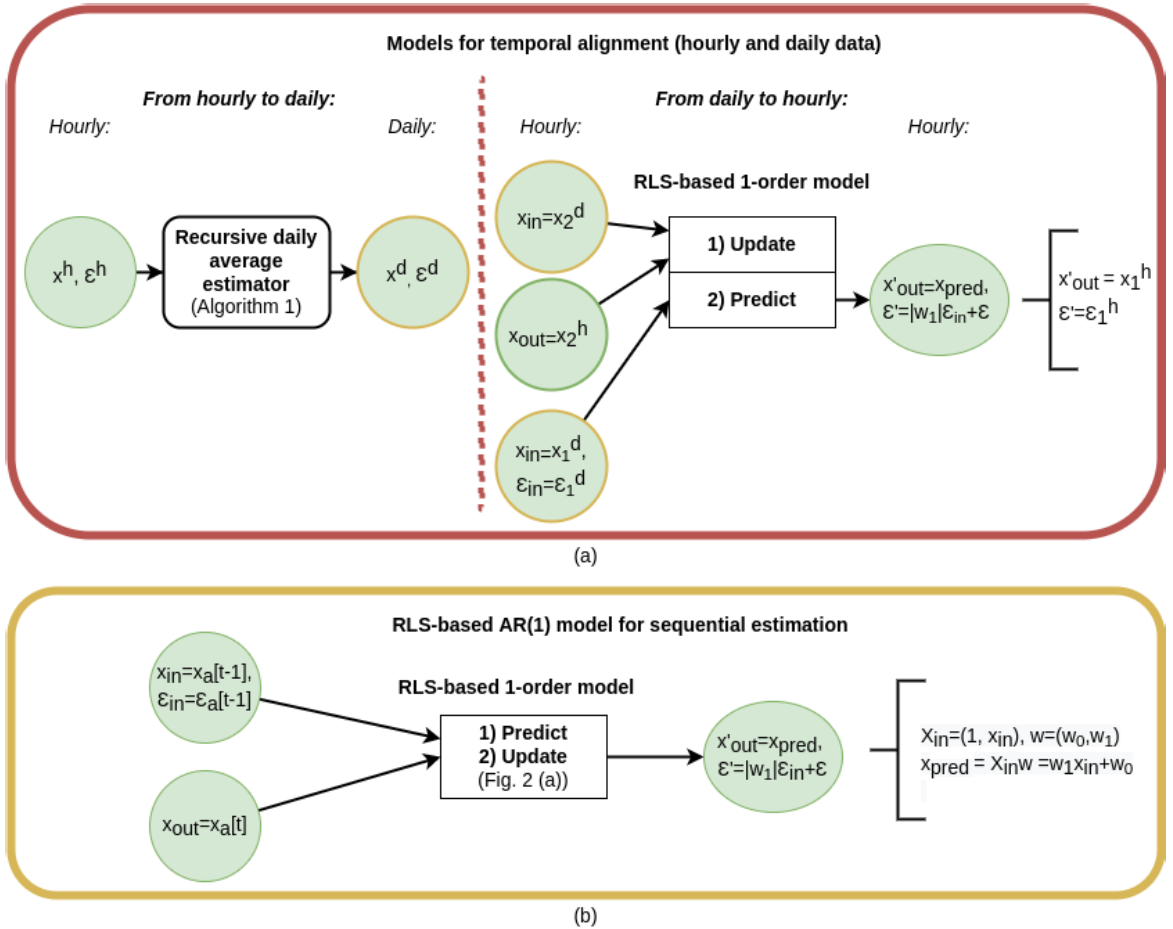


FIGURE 3. The recursive data-driven preprocessing algorithms for least-squares data assimilation (LSDA) proposed in this work. (a) models of temporal alignment (calibration) of hourly x^h and daily x^d data. (b) RLS-based first-order regression model for sequential estimation using the previously estimated analysis values, $x_a[t - 1]$ as input and the newly obtained data, x_t as output.

errors (“regression-based uncertainties”). We do not intend to conflate the two distinct concepts of “uncertainty” and “error”. Instead, we suggest an alternative to theoretical uncertainty estimates specifically for the cases of DA and demonstrate that the suggested parameters in conjunction with DA algorithms are capable of improving the data quality (accuracy, completeness, precision) of single data sources.

The uncertainties are estimated using chained 1-order RLS filters, creating a 1-order linear regression model whose parameters are estimated by the RLS algorithm using ground station observations. The filters are chained using the rules of the propagation of uncertainty as described in [22]. To minimize the number of parameters, we use a classical RLS algorithm for univariate data sources. The parameters are filter coefficients, w which consist of a 2×1 vector of the linear model coefficients estimated by the algorithm as well

as an inverse covariance matrix, $P (2 \times 2)$ which weights the previous contributions. Since there is no prior information available, the classical implementation of RLS initializes the weights to zero to avoid any bias in the estimate of the filter coefficients and the matrix P to the identity matrix that performs a linear transformation and makes all past observations weighted equally regardless of their time index. We wish to point out that this common practice may result in a slower convergence compared to the initialization with other parameters based on the prior knowledge or sensitivity analysis of each particular input signal. However, their identification and optimization are not the objective of the current work. Our approach is in line with the classical RLS algorithm without a forgetting factor, meaning that all the past observations are weighted equally in the estimate of the filter coefficients and that the regularization parameter is set to zero.

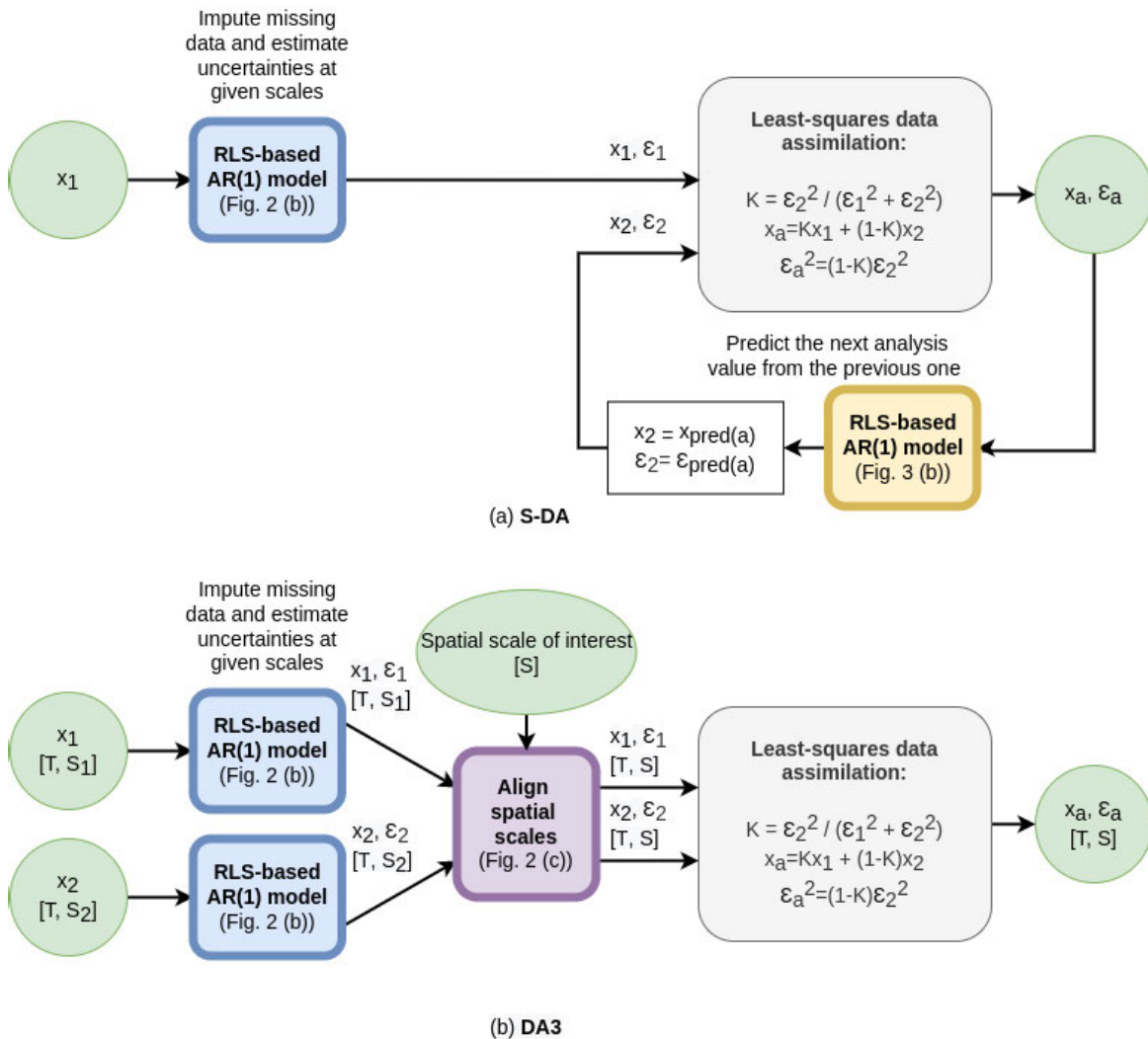


FIGURE 4. Validated data assimilation (DA) algorithms: (a) 1-source sequential least-squares DA (S-DA) using AR(1) model from Fig. 3 (b) and (b) 2-source least-squares DA with unknown uncertainties and different spatial scales (DA3, previously suggested in work [22]).

The application of the algorithms varies depending on the spatial and/or temporal scales (need to calibrate the data in space and/or time). Thus, each of the developed algorithms corresponds to a scenario of matching or non-matching scales, as described in Chapter IIIA. Each of the scenarios varies in the estimation of uncertainty, and after the uncertainties are estimated, the best-performing DA algorithm should be applied. We have chosen LSDA since it requires only the uncertainties as parameters and it is lightweight enough to perform DA in real-time and on low-powered IoT devices in the future. Nevertheless, if the parameters for other algorithms are known, the estimated uncertainties can be used as input for the other DA algorithms

such as Kalman or particle filters with low numbers of particles (e.g. 100).

The performance of DA algorithms (accuracy of the analysis results) largely depends on the optimality of provided parameters, but as mentioned above, the parameters are not always known in advance for real-world real-time implementation.

Nevertheless, the tests can also be carried out using synthetically generated datasets. For this, we perform a sensitivity analysis using one-dimensional datasets of a logistic map [30] $x_{n+1} = r \cdot x_n \cdot (1 - x_n)$ in 3 modes: periodic ($r = 3.5, x_0 = 0.5$), transient ($r = 3, x_0 = 0.75$) and chaotic ($r = 4, x_0 = 0.1$). Since the proposed algorithms

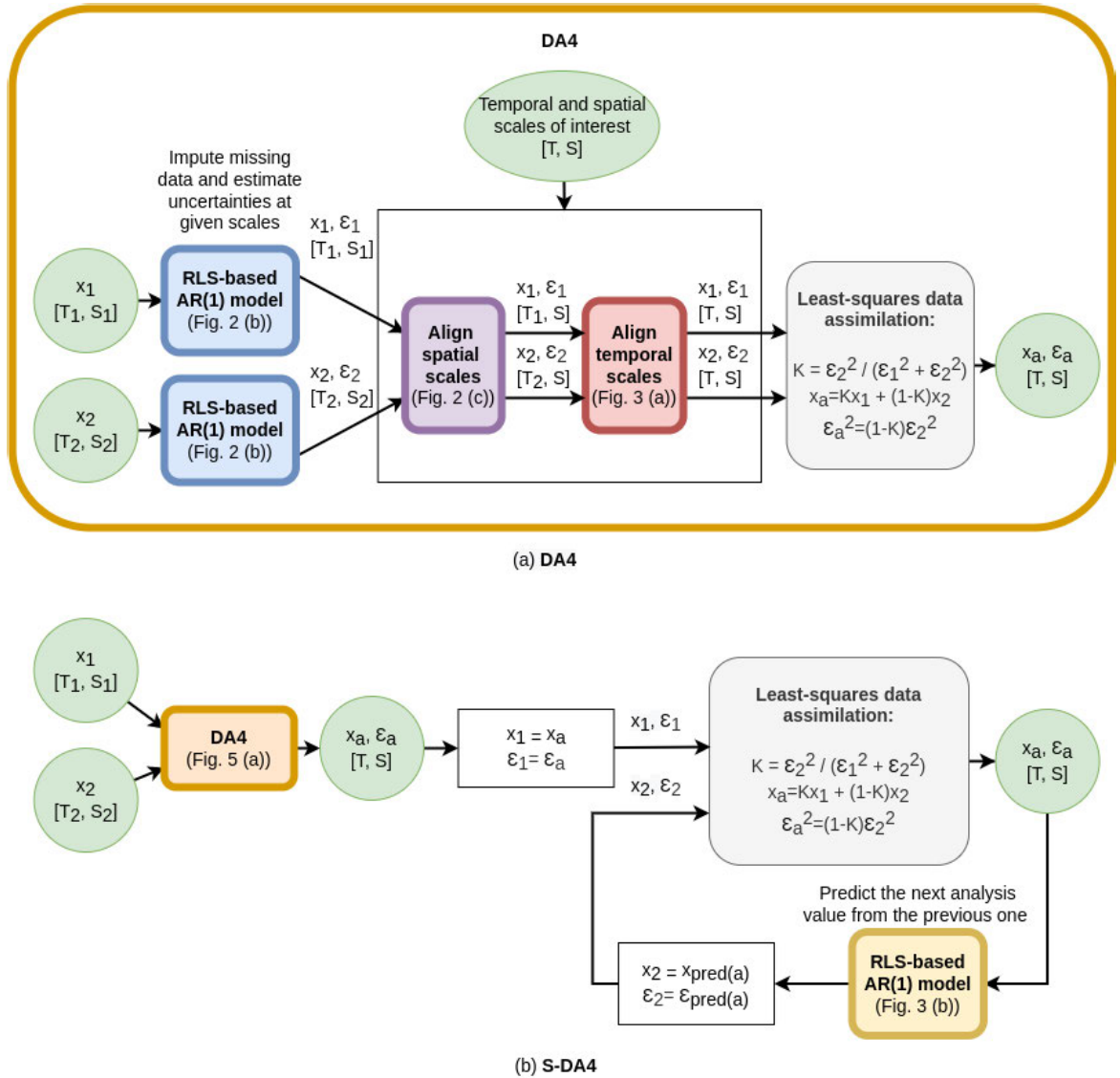


FIGURE 5. Validated data assimilation (DA) algorithms: (a) 2-source least-squares DA with unknown uncertainties, different temporal and spatial scales (DA4) and (b) 2-source sequential DA4 (S-DA4).

do not require the provision of any parameters (except the data values), we examine the performance using the data of different uncertainty (noise) levels. To generate the data sources for the DA algorithms, we apply Gaussian noise of zero mean and standard deviation σ to a clean signal of 100 iterations. The first data source is generated with a fixed amount of noise $\sigma = 0.1$ and the second data source with an increasing amount of noise from $\sigma = 0.1$ to $\sigma = 1$.

The plots for all the scenarios (DA2, DA3, S-DA, DA4 and S-DA4) are presented in Supplementary material (see Fig. 10 for DA2, DA3 and S-DA, Fig. 11 for DA4 and Fig. 12 for S-DA4). The results include the plots of increases in accuracy with respect to the amount of noise (uncertainty) in the second data source. The results in the plots are arranged in columns, each column corresponds to the same mode of a logistic map, and each row to the same scenario (DA2, DA3, S-DA, DA4, and S-DA4) of different window sizes ($M=2, M=5, M=10$).

The window size corresponds to the temporal resolution: when assimilating daily and hourly data, the window size is $M=24$ (the number of hours). Fig. 6 demonstrates an example for DA3.

The scenarios vary in purpose: S-DA is suitable when the second data source is not provided, DA2: when both data sources match in scales, represent the same variables and do not require calibration, DA3: when any of the data sources requires calibration (e.g. spatial calibration) to match the second data source, DA4 or S-DA4: when the data sources have a different temporal resolution (e.g. hourly and daily) and need alignment to produce the analysis result. No matter what the input uncertainties are, if any calibration or mapping procedure is performed, the rules of uncertainty propagation should be applied to update the final uncertainty estimate correspondingly, which creates the technical differences in the procedures of any DA algorithms (LSDA or any other DA algorithm) performed in DA2, DA3, S-DA, DA4 or S-DA4 scenarios.

The proposed by the authors algorithms use the LSDA procedure for DA and are named after the name of a scenario: DA2, DA3, S-DA, DA4 or S-DA4. For the logistic map test cases, the applied noise levels (standard deviations σ of Gaussian distributions) are known and can serve as uncertainty parameters. Therefore, we can perform LSDA with known uncertainties (in our notation in Supplementary material, LSDA for DA2 (also DA1), LSDA for DA3, LSDA for S-DA, LSDA for DA4, and LSDA for S-DA4). The difference between “LSDA for DA3” (LSDA with known uncertainties, the second data source is calibrated to the first) and “DA3” (LSDA with unknown uncertainties, the second data source is calibrated to the first) is in the provision of input uncertainties: we use standard deviations σ of Gaussian distributions of the applied noise as known uncertainties, whereas our algorithms estimate uncertainties not knowing about the σ uncertainties using the regression procedures described above.

At the same time, instead of LSDA, other lightweight methods can be used, e.g. ensemble Kalman filter (EnKF) [28] or particle filter (PF) [29]. EnKF updates the state estimate by propagating a set of model state vectors (ensemble members) through time and using the observations to correct the ensemble’s mean and covariance. PF uses a set of weighted particles to represent the probability distribution of the state variables. The particles are sampled from the prior distribution and are propagated through the transition function to obtain a posterior distribution. The particles are then resampled based on their weights, which are computed using the likelihood function to account for the observation uncertainty. Both filters provide a flexible DA framework, but the quality of the estimates depends on the number of ensemble members or particles used and the choice of the weighting scheme. In general, larger numbers of ensemble members and particles increase the accuracy at the cost of more calculations, limiting the use of these methods for computationally limited applications such as IoT sensors.

To demonstrate the performance of lightweight versions of EnKF and PF assimilations, we use an EnKF with 10 ensemble members and a PF with 100 particles. Models of this size are feasible to run on IoT devices and thus these models provide a realistic comparison of the two established DA methods (EnKF, PF) against those proposed in this work (S-DA, DA4 and S-DA4).

For each of the DA cases, the source is corrupted with noise of increasing amounts to generate progressively less accurate sources. The performance is assessed using the root mean squared error (RMSE) in relation to the ideal, zero-noise signal. The change in accuracy after assimilation is estimated as a percentage: $(1 - \frac{RMSE(\text{true, assimilated})}{RMSE(\text{true, less accurate source})}) \cdot 100\%$.

For each of the logistic map test cases, we assimilate with a data source of the fixed lowest amount of noise, we expect the sources of the lowest uncertainty to result in lower increases in accuracy and the sources of the highest uncertainty to have the largest increases in accuracy after assimilation. The goal of the analysis is to compare LSDA with known uncertainties to the author’s proposed LSDA methods with unknown uncertainties. For each of the test cases, 4 algorithms were compared against each other: LSDA for one of the scenarios (DA2, DA3, S-DA, DA4, or S-DA4) with known uncertainties σ , LSDA with unknown uncertainties. The scenarios are named based on the classical filter type (EnKF or PF), both of which are run using unknown uncertainties. For each noise level, the test is repeated 100 times, and the mean increase in accuracy is plotted as the ensemble average of these 100 repetitions.

The results show that for all the modes of the logistic map, the algorithms using a single source (S-DA) scenario perform worse than in scenarios with 2 data sources. For periodic and chaotic modes of the logistic map test cases, LSDA with known uncertainties outperforms the suggested LSDA with unknown uncertainties by around 20%, EnKF: 25%, and PF: 40% of increase in accuracy. For the transient mode, the results of LSDA with known uncertainties, LSDA with unknown uncertainties and EnKF provide similar results, varying within 5-8% with LSDA using unknown uncertainties. Without calibration (scenario DA2), PF performance decreases by nearly a factor of two when compared to the LSDA and EnKF algorithms. With calibration (scenario DA3), the performance of PF becomes closer to the other 3 algorithms, and consistently under-performs with a margin of around 5%. Considering the S-DA scenario, PF and LSDA with unknown uncertainties were found to be the two best performing algorithms.

Since DA4 and S-DA4 are designed to handle data of different temporal scales, their performance is tested for different data resolutions, defined by the window size M : the lower the window size, the higher the resolution of the data. The averaging mechanism is used only to generate the data and does not affect the execution of algorithms. For all the algorithms, the lowering of the resolution of data slightly drops the increase in accuracy within 15% from $M=2$

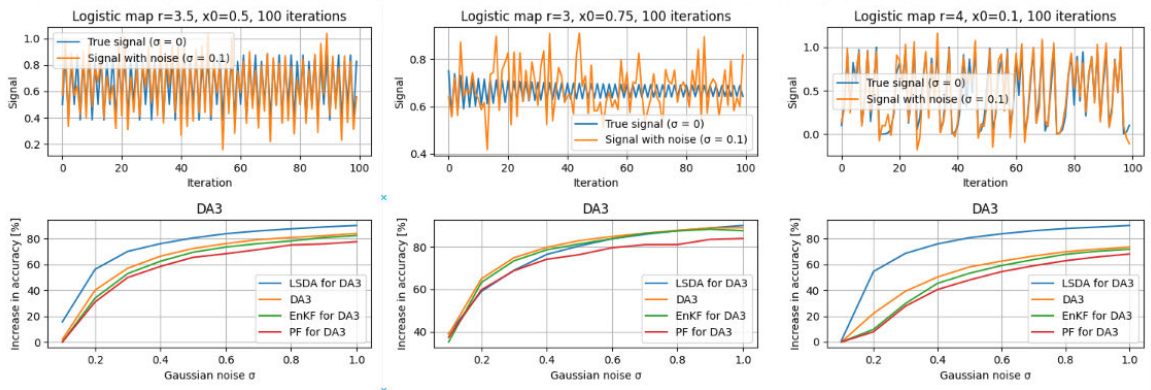


FIGURE 6. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors’ methods, labeled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors’ methods, number of ensemble members is 10), particle filter (PF, uncertainties are estimated using the authors’ methods, number of particles is 100) in scenario DA3 (with calibration). Assimilation is performed using 2 data sources corrupted with Gaussian noise of zero mean and standard deviation σ . For the assimilation of 2 data sources, the first source has a fixed amount of noise $\sigma = 0.1$, whereas the second data source has an increasing amount of noise from $\sigma = 0.1$ to $\sigma = 1$. Each experiment is performed 100 times, and the mean value of the increase in accuracy compared to the accuracy of the second data source is plotted.

to $M=10$. Nevertheless, the ranking of the algorithms per mode in DA4 is as follows: in the periodic mode, LSDA with known uncertainties outperformed LSDA with unknown uncertainties by around 10%, EnKF by 20%, and PF by 30%; in the chaotic mode, LSDA with known uncertainties also outperforms the rest by 20%, 30%, 40% correspondingly in the same order, but in the transient mode, LSDA with known uncertainties demonstrated the worst performance when compared to the other three algorithms.

Compared to the DA4 scenario, S-DA4 does not introduce a significant increase in accuracy for LSDA with unknown uncertainties. The observed reduction in accuracy increases by around 10% in the periodic mode and by around 20% in the chaotic mode compared to DA4. In the transient mode, all the other algorithms demonstrate a similar performance for both DA4 and S-DA4 with LSDA with known uncertainties being closer to the rest of the algorithms in performance. It is worth noting that the implementation of both the EnKF and PF methods require knowledge of optimal parameters and therefore the most accurate results using ensemble algorithms (EnKF or PF) may not be achievable when they are applied as lightweight DA algorithms.

Overall, the results show that the introduction of the sequential loop for S-DA4 did not provide a substantial gain in performance when compared to the DA4 algorithm for the logistic map test cases. Since DA4 has a lower computational complexity, it should therefore be chosen over S-DA4 in this example. The comparison of algorithms’ performance in DA4 or S-DA4 scenarios to DA2, DA3, or S-DA scenarios was not conducted because each algorithm is designed to handle different types of data sources. Thus, there is no need to apply DA4 to the data of the same temporal resolution, as the mapping between data sources is already handled by the calibration operator in DA3. When both data sources

measure or model the state in the same manner (e.g. 2 sensors measuring the concentration of an air pollutant, 2 accurate logistic map signals corrupted by the noise resulting in different precision), DA3 is not expected to provide a significant boost in accuracy compared to DA2, as is illustrated when comparing the DA2 and DA3 logistic map test cases.

The sensitivity analysis based on the logistic map scenarios shows that LSDA, EnKF and PF are suitable for lightweight assimilation. In general, the methods were able to cope with increasing level of random noise. We wish to point out that, in general, the results obtained by assimilating two data sources of $\sigma_1 = 0.1$ and $\sigma_2 = 0.1$ are less accurate than those obtained by assimilating two data sources of $\sigma_1 = 0.1$ and $\sigma_2 = 1$. The assimilation of 2 data sources with overall lower uncertainty would typically result in a more accurate estimate than the assimilation of data sources of higher uncertainty. This point is crucial when applying lightweight DA to cases where the data source quality is mixed: for example, one of the sources provides more accurate data, but the second source has less missing data, or when the quality of any of the data sources changes over time. In order to further investigate the performance of the proposed lightweight DA methods for sources with unknown uncertainty, open air quality data are taken from pan-European sources and assimilated with a global numerical model at a large scale.

E. DATA SOURCES

In this work, we have assimilated AQ data from the following open data sources: System for Integrated modelLing of Atmospheric coMposition (SILAM, global, version 5.7, FRC forecasts at the surface, hourly 0.2° model grid) and European Environment Agency (EEA) Air Quality data (European AQ data, hourly fixed point surface observations) in the period

from 2022-01-27 01:00:00 to 2022-02-25 15:00:00. When generating the daily values from the hourly data, we retrieve the arithmetic averages of hourly values within the same day. The AQ data include the concentrations of the following air pollutants: sulfur dioxide (SO₂), nitrogen dioxide (NO₂), carbon monoxide (CO) and ozone (O₃), and particulate matter (PM_{2.5} and PM₁₀).

SILAM generates global 4-day forecasts of AQ data including SO₂, NO, NO₂, O₃, PM_{2.5}, and PM₁₀. The results are updated daily and stored in a 30-day publicly available archive [31]. The same model was used for our previous work, albeit for a single ground observation station [22].

The European AQ dataset used in this work includes AQ data reported by the European Union (EU) member states, meta-information on the monitoring networks, stations and measurements, and assessment settings [11]. Stations were filtered by the AQ station type (“background”) and station area (“urban”). For validation purposes, we have also chosen stations that have less than 20% of missing data. The filter criteria resulted in 86 stations for CO, 593 stations for NO₂, 462 stations for O₃, 137 stations for SO₂, 254 stations for PM_{2.5}, and 445 stations for PM₁₀. The individual station locations and corresponding AQ variables are shown in Fig. 1. The data from each of the stations are assimilated with simulation results obtained from the corresponding SILAM numerical model grid cell.

The data used for the experiments as well as the source code of the algorithms are available via GitHub by <https://github.com/effie-ms/rls-assimilation> and distributed under the MIT license.

IV. RESULTS

A. COMPUTATIONAL COMPLEXITY AND PERFORMANCE

The proposed algorithms are based on a conventional first-order RLS filter with $O(L^2)$ computational complexity per iteration, where L is the filter length ($L = 2$). The complexity can be further reduced using other versions of RLS filters, see [32] for a more detailed overview. DA2 and S-DA use 2 RLS filters, DA3: 3, DA4: 4, S-DA4: 5.

The computational performance on a standard desktop PC was assessed using an Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz x 8 with 16Gb RAM. The execution times per single iteration of the algorithm are provided in Table 1. Note that this baseline is only used to provide a rough estimate of the computational performance of the lightweight assimilation methods.

B. VALIDATION

To compare the developed algorithms, the results obtained at each of the individual European AQ monitoring stations were pooled and averaged across all sites. Two different scenarios were compared: S-DA and DA3 and DA4 and S-DA4. Examples of results for selected EEA AQ monitoring stations are presented in Fig. 7 (DA3 and S-DA, calibration to station

TABLE 1. Execution time of 1 iteration of algorithms. The tests were performed on a computer with Intel(R) Core(TM) i7-8565U CPU @ 1.80GHz x 8 and 16Gb RAM.

Algorithm	Time (ms)
	mean ± sd [min; max]
DA2	0.056 ± 0.013 [0.045; 0.145]
DA3	0.077 ± 0.006 [0.073; 0.135]
S-DA	0.058 ± 0.006 [0.055; 0.120]
DA4	0.100 ± 0.008 [0.076; 0.166]
S-DA4	0.126 ± 0.009 [0.103; 0.206]

observations) and Fig. 8 (DA4 and S-DA4, calibration to model simulations).

First, we compare the performance of algorithms S-DA (sequential 1-source LSDA) and DA3 (non-sequential 2-source LSDA) as illustrated in Fig. 4. Hourly observations were taken from the EEA AQ dataset and assimilated with hourly SILAM simulation data. To compare performance, the root mean squared error (RMSE, see Equation (1)) and mean absolute uncertainty (MAU, see Equation (2)) were used.

$$RMSE(\mathbf{x}_1; \mathbf{x}_2) = \sqrt{\frac{1}{n} \sum_{i=1}^n (\mathbf{x}_1[i] - \mathbf{x}_2[i])^2}, \quad (1)$$

where $\mathbf{x}_1, \mathbf{x}_2$ are vectors of data values of length n from 2 data sources.

$$MAU(\epsilon) = \frac{1}{n} \sum_{i=1}^n |\epsilon[i]|, \quad (2)$$

where ϵ is a vector of regression-based uncertainties of length n .

As a reference data source for S-DA and DA3, we chose station observations (x_{obs}). Here, the S-DA assimilated station observations and analysis predictions of station observations and RMSE were calculated between the analysis values $x_{a(S-DA)}$ and input station observations x_{obs} . For DA3, the spatial scale of interest S is the scale of station observations, meaning that model estimates are calibrated to the scale of station observations and RMSE is also calculated between the analysis values $x_{a(DA3)}$ and input station observations x_{obs} . After calculating RMSE and MAU for each station using the S-DA and DA3 algorithm, we obtained ratios for RMSE (see (3)) and MAU (see (4)) for each station.

$$r_{RMSE} = \frac{RMSE(x_{a(S-DA)}; x_{obs})}{RMSE(x_{a(DA3)}; x_{obs})}, \quad (3)$$

$$r_{MAU} = \frac{MAU(\epsilon_{a(S-DA)})}{MAU(\epsilon_{a(DA3)})}, \quad (4)$$

where ϵ is a vector of uncertainties of length n .

When dividing the calculated RMSE and MAU metrics of S-DA by the metrics of DA3, if r_{RMSE} is 1 or larger, then the performance is the same, or $S - DA$ results in a higher error as compared to DA3. Otherwise, DA3 had the larger error. If r_{MAU} is 1 or larger, then the uncertainties are the same for both algorithms or $S - DA$ has a higher uncertainty than DA3,

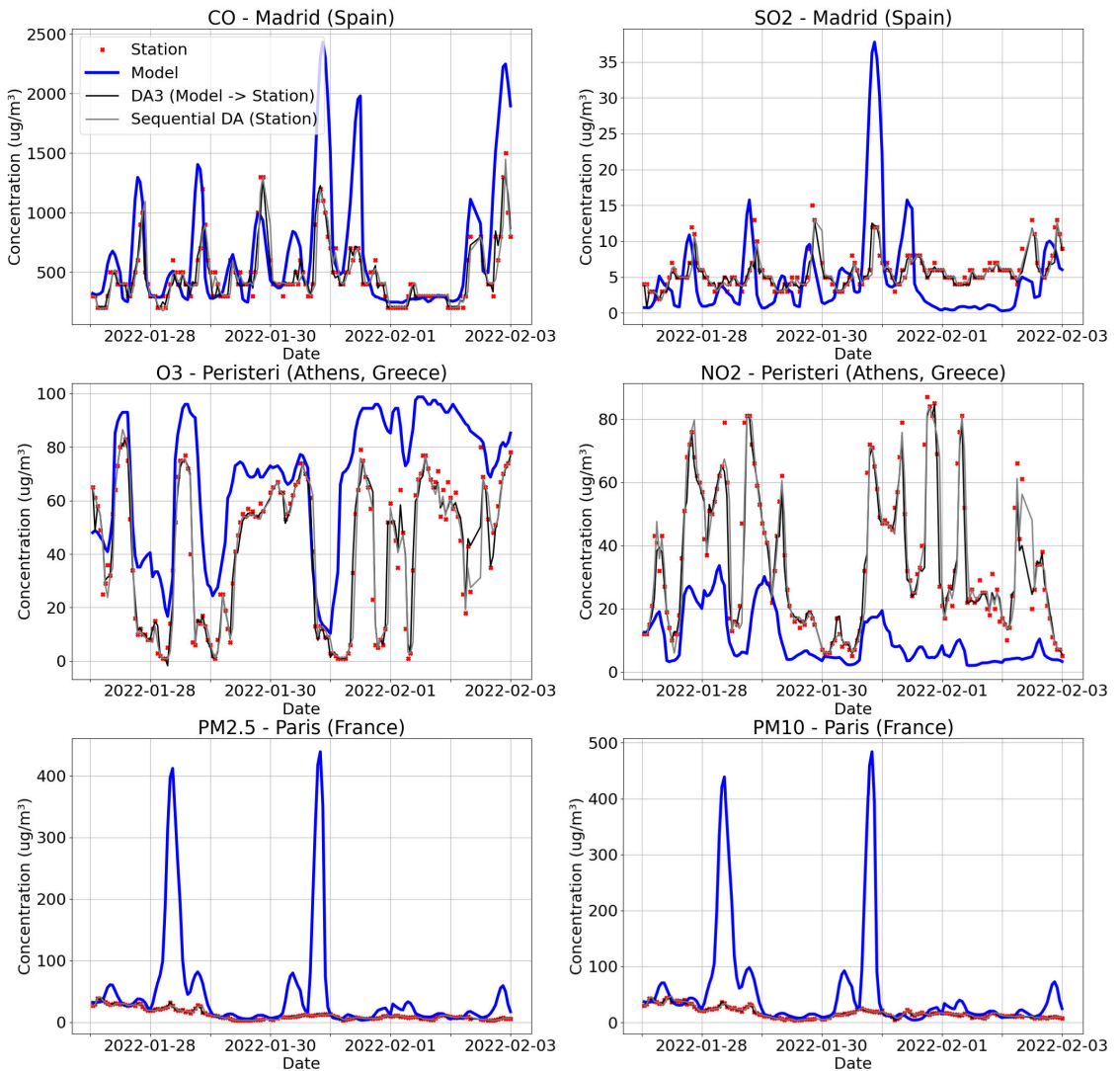


FIGURE 7. Time series plots of input data sources and assimilated values for CO, SO₂, PM_{2.5}, NO₂, O₃, PM₁₀ AQ variables. “Station” corresponds to observations made by the AQ monitoring stations in Madrid (Spain, CO, SO₂), Peristeri (Athens, Greece, O₃, NO₂), Paris (France, PM_{2.5}, PM₁₀). “Model” refers to the SILAM simulations, “DA3 (Model → Station)”, applied DA3 using a calibration of hourly model simulations to hourly station observations. “Sequential DA”, used algorithm S-DA for hourly station observations. The shown time interval is the first week of the interval used for experiments: from 2022-01-27 01:00:00 to 2022-02-03 00:00:00.

otherwise DA3 results in a higher uncertainty. The results of the comparison of S-DA and DA3 are presented in Table 2.

Overall, the RMSE ratios show that S-DA results in a slightly higher error from the reference compared to DA3. However, the MAU ratios demonstrate that S-DA can provide a lower uncertainty than DA3. Thus, the use of two sources results in a slightly lower error from the reference, whereas sequential estimation resulted in an overall lower uncertainty.

Secondly, we compared the DA4 and S-DA4 algorithms as illustrated in Fig. 5. Here, two data sources (station observations and SILAM model estimations) were used to see whether sequential estimation for 2 data sources can improve the results of DA4. The DA4 algorithm assimilates data of both different temporal and spatial scales. For this test, we replaced hourly SILAM estimations, x_m^h with the last available daily averages from the previous day, x_m^d . We define hourly as the temporal scale of interest, T and the spatial

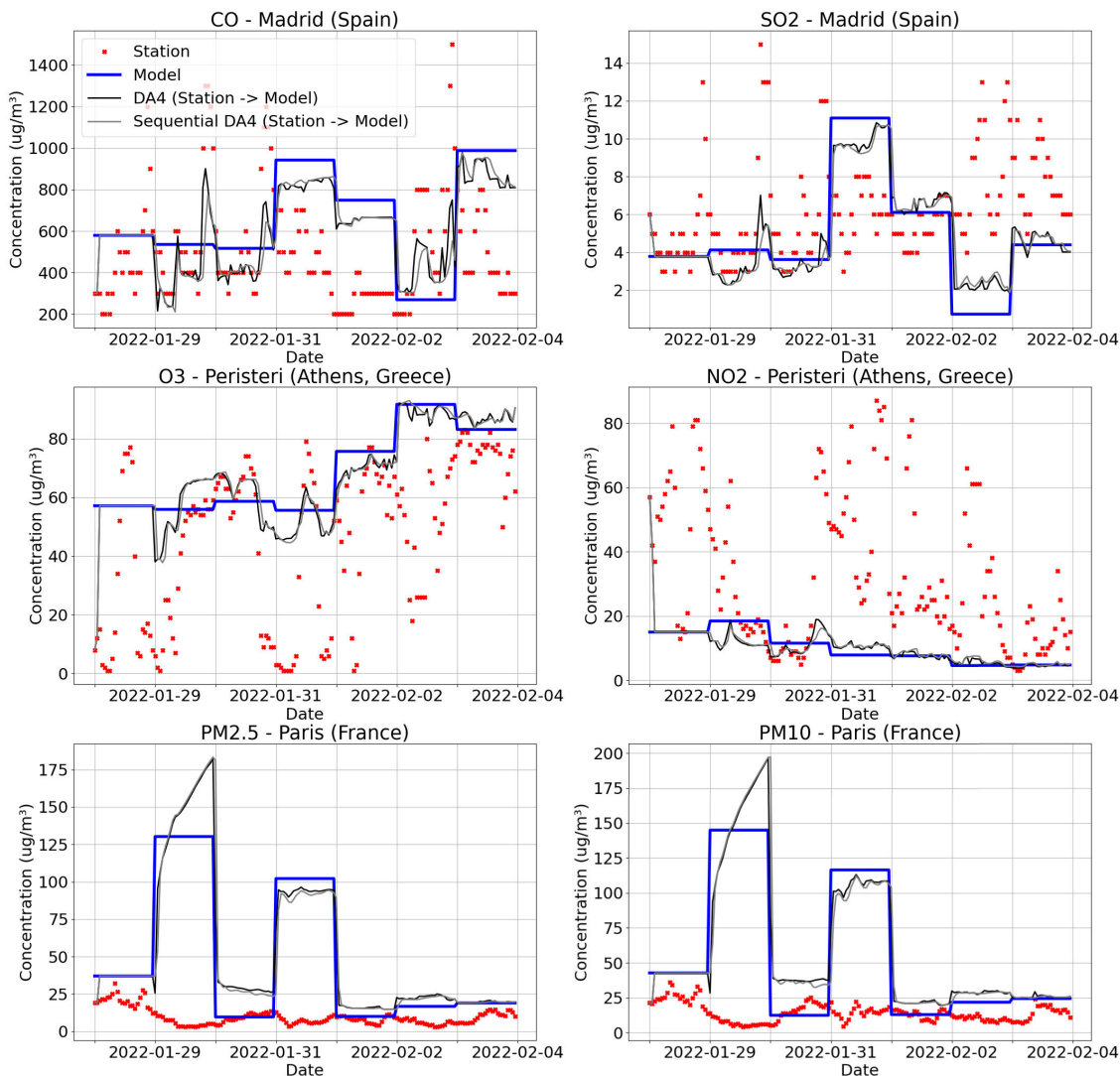


FIGURE 8. Time series plots of the input data and assimilated values for CO, SO₂, PM_{2.5}, NO₂, O₃, PM₁₀ AQ variables. “Station” corresponds to observations made by the AQ monitoring stations in Madrid (Spain, CO, SO₂), Peristeri (Athens, Greece, O₃, NO₂), Paris (France, PM_{2.5}, PM₁₀). “Model” - SILAM simulations, “DA4 (Station → Model)” refers to the DA4 algorithm with calibration of hourly station observations to daily model simulations, “Sequential DA4 (Station → Model)” shows results from S-DA4 based on the calibration of hourly station observations to daily model simulations. The time interval is the first week of the interval used for experiments: from 2022-01-27 01:00:00 to 2022-02-03 00:00:00.

scale of SILAM as the spatial scale of interest, S . In this case, using DA4, station observations were spatially calibrated to the scale of SILAM (as in DA3) and included the temporal alignment of daily SILAM to hourly station observations to obtain hourly SILAM values. The recursive daily average estimator based on RLS are shown in Fig. 3 (a). The motivation of this experiment was to test the suggested DA algorithms to improve the accuracy of hourly SILAM

results given daily SILAM values and hourly ground station observations.

The tests were performed for each of the European AQ stations, and the RMSE was calculated with respect to the reference hourly model values following Equation (5). In this case, when the ratios are higher than 1, the errors between the hourly assimilated and hourly reference values are larger than the errors obtained between the daily averages and hourly

TABLE 2. Comparison of RMSE and MAU for S-DA and DA3 algorithms by S-DA/DA3 ratios for hourly station observations as reference. The mean ratio value (mean), standard deviation (sd), minimum and maximum values of ratios (min and max) and the number of stations (N).

Variable	RMSE and MAU comparison for S-DA/DA3 (hourly, reference: station observations)	
	r_{RMSE} mean \pm sd [min; max]	r_{MAU} mean \pm sd [min; max]
CO (N=86)	1.116 \pm 0.149 [0.660; 1.621]	0.925 \pm 0.056 [0.797; 1.034]
SO ₂ (N=137)	1.043 \pm 0.172 [0.286; 2.000]	0.914 \pm 0.083 [0.689; 1.128]
PM _{2.5} (N=254)	1.158 \pm 0.181 [0.792; 1.865]	0.902 \pm 0.064 [0.575; 1.372]
NO ₂ (N=593)	1.257 \pm 0.182 [0.892; 2.036]	0.919 \pm 0.044 [0.771; 1.262]
O ₃ (N=462)	1.367 \pm 0.176 [0.651; 2.068]	0.902 \pm 0.040 [0.801; 1.273]
PM ₁₀ (N=445)	1.136 \pm 0.156 [0.697; 1.923]	0.899 \pm 0.066 [0.455; 1.079]

TABLE 3. Comparison of RMSE and MAU for the S-DA4 and DA4 algorithms using ratios based on the hourly SILAM simulations as reference. The mean ratio value (mean), standard deviation (sd), minimum and maximum values of ratios (min and max and the number of stations (N).

Variable	RMSE and MAU comparison for S-DA4 and DA4 (daily to hourly, reference: SILAM estimations)		
	$r_{RMSE}^{d \rightarrow h}$ for DA4 mean \pm sd [min; max]	$r_{RMSE}^{d \rightarrow h}$ for S-DA4 mean \pm sd [min; max]	$r_{MAU}^{d \rightarrow h}$ mean \pm sd [min; max]
CO (N=86)	0.980 \pm 0.125 [0.708; 1.242]	0.933 \pm 0.102 [0.684; 1.215]	0.593 \pm 0.024 [0.536; 0.639]
SO ₂ (N=137)	0.972 \pm 0.064 [0.863; 1.472]	0.967 \pm 0.043 [0.849; 1.116]	0.638 \pm 0.074 [0.480; 0.869]
PM _{2.5} (N=254)	0.959 \pm 0.074 [0.799; 1.410]	0.969 \pm 0.072 [0.819; 1.403]	0.601 \pm 0.046 [0.336; 0.818]
NO ₂ (N=593)	0.911 \pm 0.062 [0.770; 1.335]	0.926 \pm 0.057 [0.795; 1.339]	0.592 \pm 0.037 [0.415; 0.823]
O ₃ (N=462)	0.868 \pm 0.084 [0.675; 1.449]	0.885 \pm 0.080 [0.697; 1.466]	0.614 \pm 0.033 [0.479; 0.770]
PM ₁₀ (N=445)	0.940 \pm 0.075 [0.787; 1.611]	0.949 \pm 0.075 [0.793; 1.665]	0.594 \pm 0.038 [0.509; 0.790]

reference values. This indicates that the calibration did not substantially improve the assimilation results.

$$r_{RMSE}^{d \rightarrow h} = \frac{RMSE(x_a; x_m^h)}{RMSE(x_m^d; x_m^h)}, \quad (5)$$

where x_a are analysis values for the DA4 and S-DA4 algorithms.

The MAU ratios $r_{MAU}^{d \rightarrow h}$ are calculated similarly to Equation (4), but by dividing $MAU(\epsilon_a(S-DA4))$ by $MAU(\epsilon_a(DA4))$.

The results of the comparison of S-DA4 and DA4 are presented in Table 3.

Table 3 indicates that both DA4 and S-DA4 can result in higher accuracy (lower overall error) than the daily reference when compared to the hourly reference. However, the algorithms with the lowest RMSE ratio vary depending on the AQ variable. In particular, DA4 was found most suitable

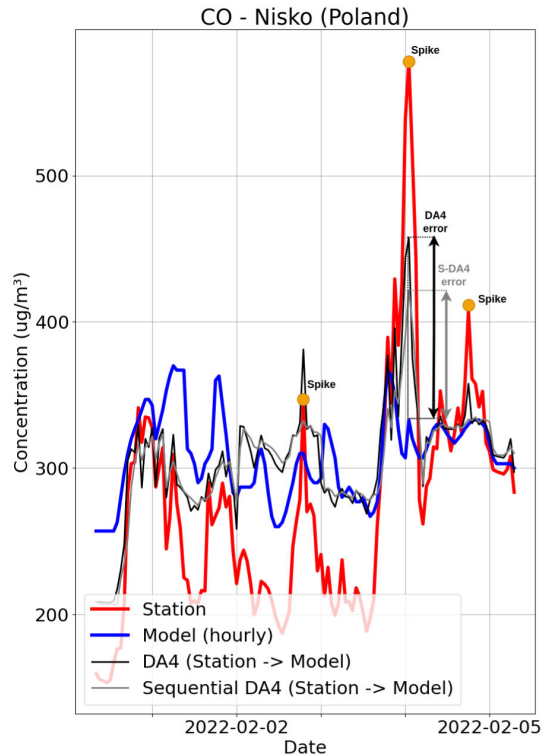


FIGURE 9. Demonstration of situations when S-DA4 can outperform DA4. “Station” corresponds to observations made by the Nisko AQ monitoring station (Nisko, Poland), “Model (hourly)”, hourly SILAM simulations, “DA4 (Station \rightarrow Model)”, algorithm DA4 with calibration of hourly station observations to daily model simulations, “Sequential DA4 (Station \rightarrow Model)”, algorithm S-DA4 with calibration of hourly station observations to daily model simulations. “Model (hourly)” are target values used for validation when performing DA with calibration of hourly station observations to daily model simulations. When spikes occur in “Station”, but not in “Model” data, S-DA4 smooths the analysis value more than DA4 resulting in a lower error from the target value (“Model (hourly)”) and consequently higher accuracy.

for PM_{2.5}, NO₂, O₃ and PM₁₀ and S-DA4 for CO and SO₂. It should also be noted that the uncertainties of S-DA4 were found to be significantly lower than the uncertainties of DA4. Similar tests with additional observations and numerical models can be obtained using the open code repository provided in this work.

V. DISCUSSION

Algorithm performance was found to correspond to the specific temporal and spatial scales of the assimilation output. In particular, if only one data source is available, S-DA is recommended for use. In cases where the temporal and spatial scales of the data sources are the same, DA2 can be applied. If the spatial scales are different, DA3 was applied for data of the same temporal scales and DA4 (or S-DA) for data of different temporal scales. The current implementation of the algorithms serves as a demonstration of how to assimilate

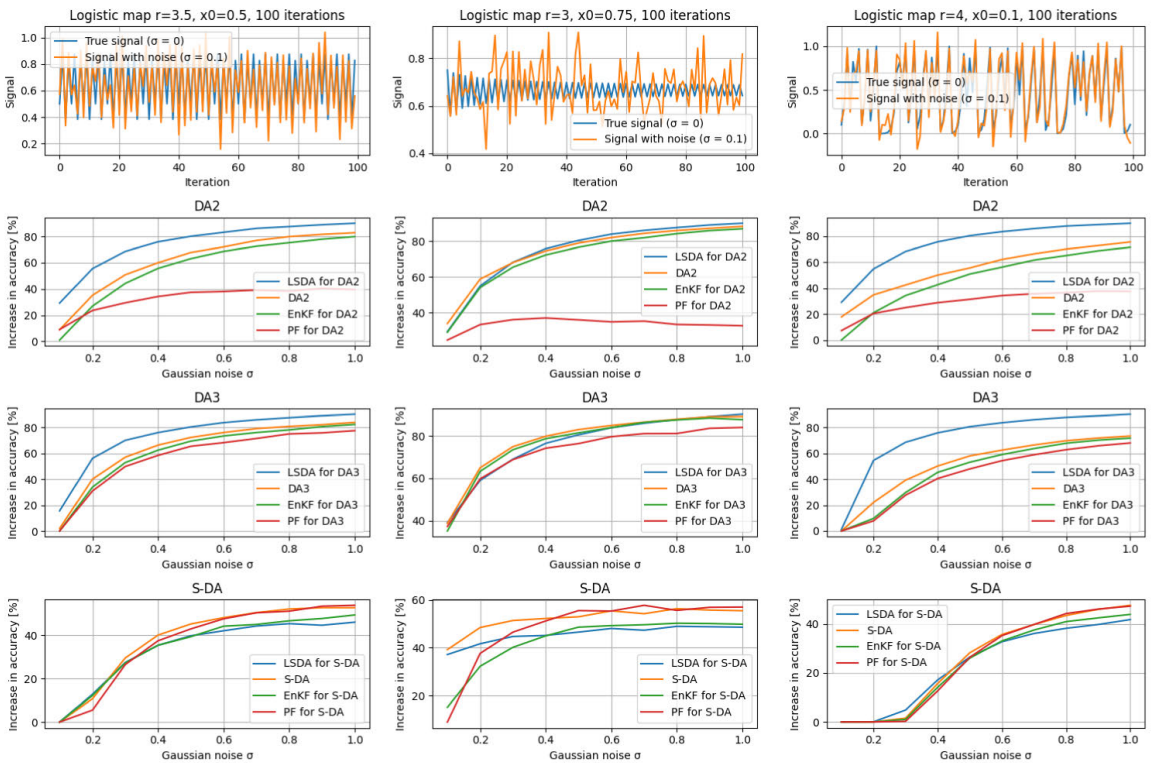


FIGURE 10. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members is 10), particle filter (PF, uncertainties are estimated using the authors' methods, number of particles is 100) in scenarios DA2 (without calibration), DA3 (with calibration) and S-DA (sequential assimilation for a single data source) for the logistic map in 3 different modes. Assimilation is performed using 2 data sources corrupted with Gaussian noise of zero mean and standard deviation σ . For the assimilation of 2 data sources, the first source has a fixed amount of noise $\sigma = 0.1$, whereas the second data source has an increasing amount of noise from $\sigma = 0.1$ to $\sigma = 1$. S-DA performs assimilation for a single data source of an increasing amount of noise. Each experiment is performed 100 times, and the mean value of the increase in accuracy compared to the accuracy of the second data source is plotted.

data of two data sources, leaving the extension of more than two sources for future research. The current code implementation of the algorithms covers only hourly and daily temporal scales, however, additional scales could be implemented and tested as needed by users after modification of the provided open source code.

When assimilating data from 2 data sources, the required temporal and spatial scales (resolution) must be represented by one of the data sources, especially when obtaining analyses of a higher resolution. For example, when assimilating grids of $0.2\hat{A}^\circ$ and $0.4\hat{A}^\circ$ spatial resolution, the algorithms allow only for retrieving analyses of $0.2\hat{A}^\circ$ or $0.4\hat{A}^\circ$ spatial resolution, unless explicitly coding a translation operator to other resolutions. The same applies to both temporal and spatial scales.

When choosing between algorithms DA4 and S-DA4, both demonstrated similar overall performance, but DA4 is computationally more lightweight compared to S-DA4. Nevertheless, S-DA4 can provide a higher accuracy compared to

DA4 when a calibrated data source has rapid changes with high magnitudes which are not captured by the reference data source. In this case, when assimilating with a previous analysis value after applying DA4 (S-DA4) the analysis was found to frequently generate short-duration peaks at lower amplitudes. As an example, in Fig. 9, the station observations are found to produce rapid changes of a high magnitude, but these events are not well-resolved by the numerical model simulations. Since model simulations serve as a reference data source for these analyses and station observations are calibrated to model simulations, the analysis peaks from both DA4 and S-DA4 exhibit a lower magnitude. The magnitude of S-DA4 was lower than that of DA4, making the result closer to the reference source and consequently of higher accuracy.

VI. CONCLUSION

The growing number of openly available AQ data require improved and standardized methods for uncertainty estimation as well as spatio-temporal calibration to

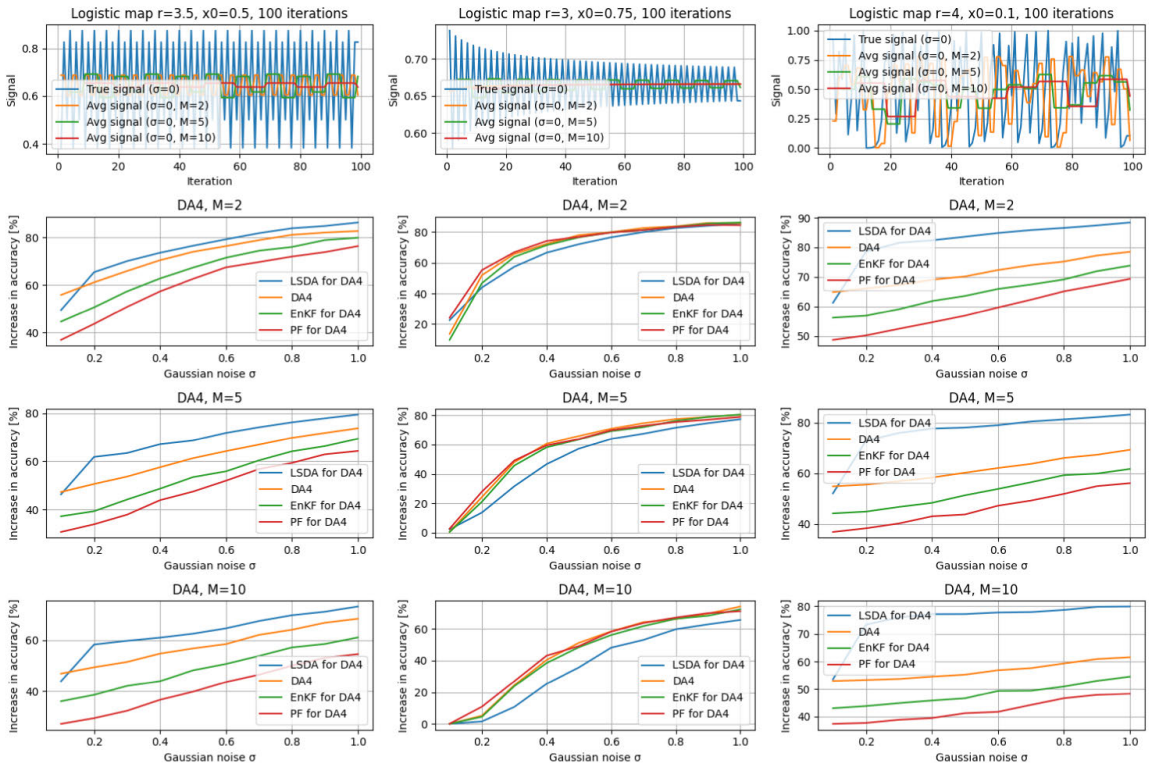


FIGURE 11. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors’ methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors’ methods, number of ensemble members is 10), particle filter (PF, uncertainties are estimated using the authors’ methods, number of particles is 100) in scenarios DA4 for the logistic map in 3 different modes. Assimilation is performed using 2 data sources corrupted with Gaussian noise of zero mean and standard deviation σ . The first source has a fixed amount of noise $\sigma = 0.1$, whereas the second data source has an increasing amount of noise from $\sigma = 0.1$ to $\sigma = 1$. The temporal resolution of the second data source is also decreased within windows of size $M=2$, $M=5$ and $M=10$. Each experiment is performed 100 times, and the mean value of the increase in accuracy compared to the accuracy of the second data source is plotted.

enable data assimilation. In our work, we have developed a lightweight method to pre-process data for least-squares data assimilation in a fully data-driven way. Compared to our previous work on a single station [22], we extend lightweight assimilation methods to include temporal calibration and sequential estimation and validate the proposed methods using the data from urban AQ monitoring stations throughout Europe.

To evaluate algorithmic performance, we assessed the errors of the assimilated values from the ground station reference sources as well as their corresponding uncertainties. First, we compared a single-source sequential (S-DA) algorithm against a two-source non-sequential with different spatial scales (DA3) algorithm. This error comparison indicated that DA3 can reduce the error from the ground station reference value, but exhibited higher uncertainties when compared with the canonical S-DA algorithm. Secondly, we compared two-source non-sequential (DA4) and sequential (S-DA4) algorithms with different temporal and spatial scales. The comparison showed that both DA4 and

S-DA4 results were more accurate with respect to the hourly reference as compared to daily reference values.

Using the openly available EEA AQ ground station observations and SILAM numerical simulation results, the proposed lightweight assimilation methods were shown to improve the overall quality of single-source estimates. In particular, the proposed methods were found to improve the completeness, accuracy and precision of the AQ observations. This study also demonstrates that the reuse of open data without uncertainty could become a cost-efficient alternative to the deployment of additional urban AQ monitoring stations.

In Fig. 7, the differences between the model and observations are expected due to the significant scale differences between the values from the SILAM grid forecasts and fixed-point observations. As a result, local sources of pollution such as traffic congestion or industrial emissions observed locally might not be included in the model forecasts. Additionally, the observations themselves may not be perfectly accurate due to instrument errors or meteorological conditions. We do not intend to draw definitive conclusions about the validity

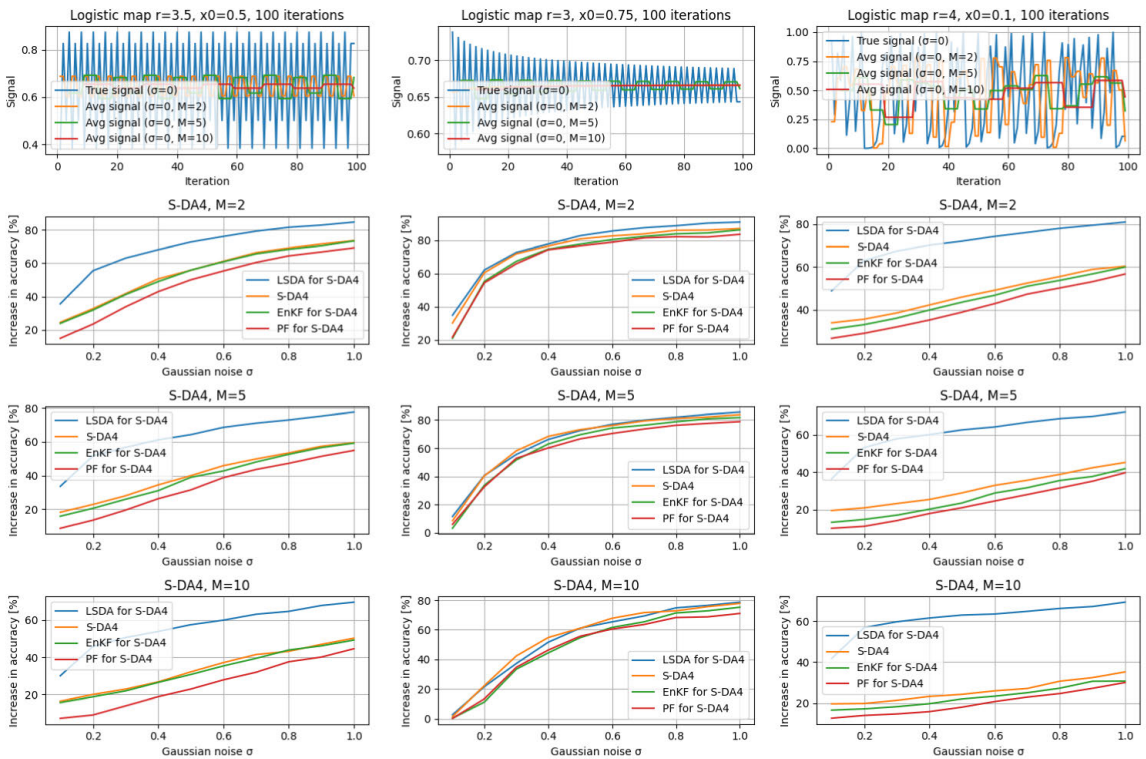


FIGURE 12. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members is 10), particle filter (PF, uncertainties are estimated using the authors' methods, number of particles is 100) in scenarios S-DA4 for the logistic map in 3 different modes. Assimilation is performed using 2 data sources corrupted with Gaussian noise of zero mean and standard deviation σ . The first source has a fixed amount of noise $\sigma = 0.1$, whereas the second data source has an increasing amount of noise from $\sigma = 0.1$ to $\sigma = 1$. The temporal resolution of the second data source is also decreased within windows of size $M=2$, $M=5$ and $M=10$. Each experiment is performed 100 times, and the mean value of the increase in accuracy compared to the accuracy of the second data source is plotted.

of the data from any of the data sources where there are significant differences, as we are reusing open data collected or generated by external sources. Moreover, we do not have detailed information on the true reasons for the drastic differences observed.

A univariate state-space model was applied to create a dynamic linear model of a system or process in which a single variable (e.g. air pollutants) is observed over time. The observation function specifies the relationship between the observed variable and the state variable, and the state transition function specifies how the state variable evolves over time. The LSDA methods applied in this work implicitly assume that the state transitions are time-invariant. Thus, the observed variables are used “as-is” for the state estimation and provide a weighted average. Additional variables can also be included to account for weather-related parameters and nonlinear transition operators could be applied to improve the final accuracy. However, multivariate cases are beyond the scope of the paper.

Future research will focus on creating time-varying maps based on the interpolation of the data assimilation outputs to at hourly and daily temporal resolutions. In addition, we intend on exploring the use of the proposed lightweight data assimilation methods to develop algorithms for the optimal placement of urban air quality monitoring stations to reduce AQ forecast uncertainty. We hope that other researchers make use of the open repository provided in this work, as the lightweight algorithms provided can be tested, calibrated and validated on monitoring data of various types and can be feasibly extended to assimilate additional data sources such as satellite observations or mobile sensors.

SUPPLEMENTARY MATERIAL

See Figures 10–12.

REFERENCES

- [1] P. Holnicki and Z. Nahorski, “Emission data uncertainty in urban air quality modeling—Case study,” *Environ. Model. Assessment*, vol. 20, no. 6, pp. 583–597, Dec. 2015.

- [2] J. Horálek, M. Schreiberová, L. Vlasáková, J. Marková, F. Tognet, P. Schneider, P. Kurfürst, and J. Schovánková, "European air quality maps for 2018. PM₁₀, PM_{2.5}, Ozone, NO₂ and NO_x spatial estimates and their uncertainties," Eur. Topic Centre Air Pollut., Transp., Noise Ind. Pollut. (ETC/ATNI), Norwegian Inst. Air Res. (NILU), Kjeller, Norway, Tech. Rep. 10/2020, 2021.
- [3] B. Denby, A. V. Dudek, S. E. Walker, A. P. A. Costa, A. Monteiro, S. van den Elshout, and B. E. A. Fisher, "Towards uncertainty mapping in air-quality modelling and assessment," *Int. J. Environ. Pollut.*, vol. 44, pp. 14–23, Jan. 2011.
- [4] B. Crawford, D. H. Hagan, I. Grossman, E. Cole, L. Holland, C. L. Heald, and J. H. Kroll, "Mapping pollution exposure and chemistry during an extreme air quality event (the 2018 Kilauea eruption) using a low-cost sensor network," *Proc. Nat. Acad. Sci. USA*, vol. 118, no. 27, 2021, Art. no. e2025540118.
- [5] I. Mokhtari, W. Bechkit, H. Rivano, and M. R. Yaici, "Uncertainty-aware deep learning architectures for highly dynamic air quality prediction," *IEEE Access*, vol. 9, pp. 14765–14778, 2021.
- [6] A. Sanpei, T. Okamoto, S. Masamune, and Y. Kuroe, "A data-assimilation based method for equilibrium reconstruction of magnetic fusion plasma and its application to reversed field pinch," *IEEE Access*, vol. 9, pp. 74739–74751, 2021.
- [7] M. Eltahan and S. Alahmadi, "Numerical dust storm simulation using modified geographical domain and data assimilation: 3DVAR and 4DVAR (WRF-Chem/WRFDA)," *IEEE Access*, vol. 7, pp. 128980–128989, 2019.
- [8] J. Y. Seo and S.-I. Lee, "Predicting changes in spatiotemporal groundwater storage through the integration of multi-satellite data and deep learning models," *IEEE Access*, vol. 9, pp. 157571–157583, 2021.
- [9] Z. H. Ismail and N. A. Jalaludin, "Robust data assimilation in river flow and stage estimation based on multiple imputation particle filter," *IEEE Access*, vol. 7, pp. 159226–159238, 2019.
- [10] M. Fan, Y. Bai, L. Wang, and L. Ding, "Combining a fully connected neural network with an ensemble Kalman filter to emulate a dynamic model in data assimilation," *IEEE Access*, vol. 9, pp. 144952–144964, 2021.
- [11] European Environment Agency. (2022). *Download of Air Quality Data*. [Online]. Available: <https://discomap.eea.europa.eu/map/fme/AirQualityExport.htm>
- [12] Kepler.gl Contributors. (2022). *Kepler.gl Geospatial Analysis Tool for Large-Scale Data Sets*. [Online]. Available: <https://github.com/keplergl/kepler.gl>
- [13] (2022). *OpenStreetMap Contributors*. [Online]. Available: <https://planet.osm.org> and <https://www.openstreetmap.org>
- [14] Mapbox. (2022). *Mapbox Mapping Platform*. [Online]. Available: <https://www.mapbox.com/about/maps/>
- [15] D. Zhang and S. S. Woo, "Real time localized air quality monitoring and prediction through mobile and fixed IoT sensing network," *IEEE Access*, vol. 8, pp. 89584–89594, 2020.
- [16] B. Nathan, S. Kremser, S. Mikaloff-Fletcher, G. E. Bodeker, L. J. Bird, E. R. Dale, D. Lin, G. Olivares, and E. Somervell, "The MAPM (Mapping Air Pollution eMissions) method for inferring particulate matter emissions maps at city scale from in situ concentration measurements: Description and demonstration of capability," *Atmos. Chem. Phys.*, vol. 21, pp. 14089–14108, 2021, doi: 10.5194/acp-21-14089-2021.
- [17] A. Gressent, L. Malherbe, A. Colette, H. Rollin, and R. Scimia, "Data fusion for air quality mapping using low-cost sensor observations: Feasibility and added-value," *Environ. Int.*, vol. 143, Oct. 2020, Art. no. 105965.
- [18] Y. Yu, J. J. Q. Yu, V. O. K. Li, and J. C. K. Lam, "A novel interpolation-SVT approach for recovering missing low-rank air quality data," *IEEE Access*, vol. 8, pp. 74291–74305, 2020.
- [19] J. Vira and M. Sofiev, "Assimilation of surface NO₂ and O₃ observations into the SILAM chemistry transport model," *Geosci. Model Develop.*, vol. 8, pp. 191–203, Feb. 2015.
- [20] M. Sofiev, "On possibilities of assimilation of near-real-time pollen data by atmospheric composition models," *Aerobiologia*, vol. 35, pp. 1–9, Apr. 2019.
- [21] P. Schneider, N. Castell, M. Vogt, F. R. Dauge, W. A. Lahoz, and A. Bartonova, "Mapping urban air quality in near real-time using observations from low-cost sensors and model information," *Environ. Int.*, vol. 106, pp. 234–247, Sep. 2017.
- [22] L. Miasayedava, J. Kaugerand, and J. A. Tuhtan, "Lightweight assimilation of open urban ambient air quality monitoring data and numerical simulations with unknown uncertainty," *Environ. Model. Assessment*, pp. 1–15, Jun. 2023.
- [23] P. Hamer, S.-E. Walker, and P. Schneider. (2021). *Appropriate Assimilation Methods for Air Quality Prediction and Pollutant Emission Inversion: An Urban Data Assimilation Systems Report*. [Online]. Available: <https://www.nilu.com/pub/1890445/>
- [24] S. Ameer, M. A. Shah, A. Khan, H. Song, C. Maple, S. U. Islam, and M. N. Asghar, "Comparative analysis of machine learning techniques for predicting air quality in smart cities," *IEEE Access*, vol. 7, pp. 128325–128338, 2019.
- [25] Y. Yang, G. Christakos, W. Huang, C. Lin, P. Fu, and Y. Mei, "Uncertainty assessment of PM_{2.5} contamination mapping using spatiotemporal sequential indicator simulations and multi-temporal monitoring data," *Sci. Rep.*, vol. 6, no. 1, Apr. 2016, Art. no. 24335.
- [26] A. Preston and K.-L. Ma, "Communicating uncertainty and risk in air quality maps," *IEEE Trans. Vis. Comput. Graphics*, vol. 29, no. 9, pp. 3746–3757, Sep. 2023.
- [27] J. D. Fine, "The ends of uncertainty: Air quality science and planning in Central California," Lawrence Berkeley Nat. Lab., Berkeley, CA, USA, Tech. Rep. 54222, 2003.
- [28] G. Evensen, "The ensemble Kalman filter: Theoretical formulation and practical implementation," *Ocean Dyn.*, vol. 53, no. 4, pp. 343–367, Nov. 2003.
- [29] A. Doucet and A. M. Johansen, "A tutorial on particle filtering and smoothing: Fifteen years later," in *The Oxford Handbook of Nonlinear Filtering*, D. Crisan and B. Rozovskii, Eds. New York, NY, USA: Oxford Univ. Press, 2011, pp. 656–704.
- [30] H. L. Mitchell and P. L. Houtekamer, "Ensemble Kalman filter configurations and their performance with the logistic map," *Monthly Weather Rev.*, vol. 137, no. 12, pp. 4325–4343, Dec. 2009.
- [31] Finnish Meteorological Institute. (2022). *SILAM V5.7: System for Integrated Modelling of Atmospheric Composition. Model and Data Access*. [Online]. Available: <http://silam.fmi.fi/thredds/catalog.html>
- [32] S. Yao, H. Qian, K. Kang, and M. Shen, "A recursive least squares algorithm with reduced complexity for digital predistortion linearization," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 4736–4739.



LIZAVETA MIASAYEDAVA (Graduate Student Member, IEEE) received the B.Sc. degree in computer systems engineering and informatics from Saint-Petersburg State Electrotechnical University, Russia, and the M.Sc. (Engineering) degree in e-governance technologies and services from the Tallinn University of Technology, Estonia, where she is currently pursuing the Ph.D. degree. She has professional experience in software engineering and web development. Her research interests include data-driven modeling and computational mathematics.



JAANUS KAUGERAND (Member, IEEE) received the M.Sc. and Ph.D. degrees in computer system engineering from the Tallinn University of Technology, in 2014 and 2020, respectively. He is currently the Head of the Laboratory for Proactive Technologies, Department of Software Science, Tallinn University of Technology. His current research interests include wireless sensor networks and large-scale environmental sensing.



JEFFREY A. TUHTAN (Member, IEEE) received the B.Sc. degree in civil engineering from California Polytechnic University, San Luis Obispo, CA, USA, in 2004, and the M.Sc. degree in water resources engineering and management and the Dr.-Eng. degree from the Institute for Modelling Hydraulic and Environmental Systems, University of Stuttgart, Germany, in 2007 and 2011, respectively. Since 2016, he has been with the Centre for Environmental Sensing and Intelligence, Department of Computer Systems, Tallinn University of Technology. His research interests include data-driven modeling and bio-inspired underwater sensing in extreme environments.

...

Curriculum Vitae

Personal data

Name Lizaveta Miasayedava
Date and place of birth 30 October 1996, Novopolotsk, Belarus
Nationality Belarusian

Contact information

Address Tallinn University of Technology, School of Information Technologies,
Department of Computer Systems,
Ehitajate tee 5, 12616 Tallinn, Estonia
E-mail lizaveta.miasayedava@taltech.ee

Education

2020–... Tallinn University of Technology, School of Information Technologies,
Information and Communication Technology, PhD studies
2018–2020 Tallinn University of Technology, School of Information Technologies,
E-Governance Technologies and Services, MSc *cum laude*
2014–2018 Saint Petersburg State Electrotechnical University "LETI", Faculty of
Computer Science and Technology,
Computer Systems Engineering and Informatics, BSc *cum laude*

Language competence

Russian native
Belarusian native
English fluent
Portuguese intermediate
Estonian basic

Professional employment

2023– ... Finrate AG, Software engineer
2022–2022 Payward Inc., Software engineer
2020–2022 Tallinn University of Technology, Early stage researcher
2019–2020 Thorgate, Software engineer
2017–2018 Teambrella, Software engineer

Defended theses

- 2020, Automated Assessment of Environmental Flows Using Estonian Hydrological Open Government Data, MSc, supervisor Prof. Jeffrey A. Tuhtan, co-supervisor PhD Keegan McBride, Tallinn University of Technology, School of Information Technologies.
- 2018, Research and Implementation of Frequent Subgraph Mining Algorithms for Project Documentation Analysis, BSc, supervisor Prof. Natalie E. Novakova, Saint Petersburg State Electrotechnical University "LETI", Faculty of Computer Science and Technology.

Scientific work

Papers

1. Lizaveta Miasayedava, Keegan McBride, and Jeffrey A. Tuhtan. Automated Environmental Compliance Monitoring of Rivers with IoT and Open Government Data. *Journal of Environmental Management*, 303:114283, February 2022
2. Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty. *Environmental Modeling & Assessment*, June 2023
3. Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Open Data Assimilation of Pan-European Urban Air Quality. *IEEE Access*, 11:84670–84688, August 2023

Conference presentations

1. Jeffrey A. Tuhtan, Elizaveta Dubrovinskaya, Lizaveta Miasayedava, Vishwajeet Pattanaiik, Jürgen Soom, Bernd Mockenhaupt, Cornelia Schütz, Christian Haas, and Philipp Thumser. Smart Fish Counter for Monitoring Species, Size, Migration Behaviour and Environmental Conditions. In *The 2022 International Symposium on Ecohydraulics*, pages 1–4, 2022
2. Jeffrey A Tuhtan, Lizaveta Miasayedava, and Gert Toming. Data Assimilation of Acoustic Doppler Velocimeter and Total Pressure Sensors. Presentation at the 40th IAHR World Congress, 2023

Elulookirjeldus

Isikuandmed

Nimi Lizaveta Miasayedava
Sünniaeg ja -koht 30. oktoober 1996, Novopolotsk, Valgevene
Kodakondsus Valgevenelane

Kontaktandmed

Adress Tallinna Tehnikaülikool, Infotehnoloogia Teaduskond,
Arvutisüsteemide Instituut,
Ehitajate tee 5, 12616 Tallinn, Eesti
E-post lizaveta.miasayedava@taltech.ee

Haridus

2020-... Tallinna Tehnikaülikool, Infotehnoloogia Teaduskond,
Info- ja kommunikatsioonitehnoloogia, doktoriõpe
2018-2020 Tallinna Tehnikaülikool, Infotehnoloogia Teaduskond,
E-Riigi Tehnoloogiad ja Teenused, MSc *cum laude*
2014-2018 Tallinna Tehnikaülikool, Infotehnoloogia Teaduskond,
Arvutisüsteemide Tehnika ja Informaatika, BSc *cum laude*

Keelteoskus

vene keel	emakeel
valgevene keel	emakeel
inglise keel	kõrgtase
portugali keel	kesktasemel
eesti keel	algajatasemel

Teenistuskäik

2023- ...	Finrate AG, Tarkvara insener
2022-2022	Payward Inc., Tarkvara insener
2020-2022	Tallinna Tehnikaülikool, Algtaseme uurija
2019-2020	Thorgate, Tarkvara insener
2017-2018	Teambrella, Tarkvara insener

Kaitstud lõputööd

- 2020, Keskkonnavoolumulga Automaatne Hindamine Kasutades Eesti Avaliku Sektori Hüdroloogilisi Avaandmeid, MSc, juhendaja Prof. Jeffrey A. Tuhtan, kaasjuhendaja PhD Keegan McBride, Tallinna Tehnikaülikool, Infotehnoloogia Kool.
- 2018, Sagedaste Alamgraafide Kaevandamise Algoritmide Uurimine ja Rakendamine Projektdokumentatsiooni Analüüsiks, BSc, juhendaja Prof. Natalie E. Novakova, Peterburi Riiklik Elektrotehniline Ülikool "LETI", Arvutiteaduse ja Tehnoloogia Teaduskond.

Teadustegevus

Artiklid

1. Lizaveta Miasayedava, Keegan McBride, and Jeffrey A. Tuhtan. Automated Environmental Compliance Monitoring of Rivers with IoT and Open Government Data. *Journal of Environmental Management*, 303:114283, February 2022
2. Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Assimilation of Open Urban Ambient Air Quality Monitoring Data and Numerical Simulations with Unknown Uncertainty. *Environmental Modeling & Assessment*, June 2023
3. Lizaveta Miasayedava, Jaanus Kaugerand, and Jeffrey A. Tuhtan. Lightweight Open Data Assimilation of Pan-European Urban Air Quality. *IEEE Access*, 11:84670–84688, August 2023

Konverentsi ettekanded

1. Jeffrey A. Tuhtan, Elizaveta Dubrovinskaya, Lizaveta Miasayedava, Vishwajeet Pattanaiik, Jürgen Soom, Bernd Mockenhaupt, Cornelia Schütz, Christian Haas, and Philipp Thumser. Smart Fish Counter for Monitoring Species, Size, Migration Behaviour and Environmental Conditions. In *The 2022 International Symposium on Ecohydraulics*, pages 1–4, 2022
2. Jeffrey A Tuhtan, Lizaveta Miasayedava, and Gert Toming. Data Assimilation of Acoustic Doppler Velocimeter and Total Pressure Sensors. Presentation at the 40th IAHR World Congress, 2023

ISSN 2585-6901 (PDF)
ISBN 978-9916-80-125-3 (PDF)