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 KAUGERAND[1], (Member, IEEE), AND JEFFREY A. TUHTAN[2], (Member, IEEE)

[1]Research Laboratory for Proactive Technologies, Tallinn University of Technology, 12616 Tallinn, Estonia
[2]Department 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 NO_2, 462 stations for O_3, 137 stations for SO_2, 254 stations for $PM_{2.5}$, and 445 stations for PM_{10} 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 Jiang.

I. INTRODUCTION

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

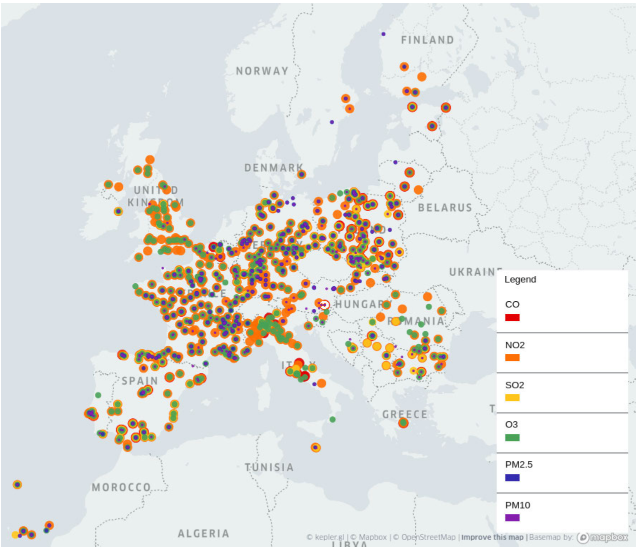


FIGURE 1. Map showing the ground station locations of the open access air quality monitoring network of the European Environment Agency (EEA) [11] used in this work for testing and validation. The map is generated using [12], [13] and [14].

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 NO_2, 462 stations for O3, 137 stations for SO2, 254 stations for PM2.5, and 445 stations for PM10. The locations are shown in Fig 1), and the observations from the location were assimilated with hourly and daily 0.2Â° numerical model simulation results obtained from the openly available System for Integrated modeLling 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 valuesx_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 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).

Algorithm 1 Recursive Daily Average Estimator

```
$\overline{x^h}$ - current average data value, $\overline{\epsilon^h}$ - current average error (uncertainty), $x^d$ - last full-day daily average data value, $\epsilon^d$ - last full-day daily average error (uncertainty), $N$ - counter of previous hours (reset after each 24 hour interval).
procedure INIT( )
    $\overline{x^h}, \overline{\epsilon^h}, x^d, \epsilon^d = 0, 0, 0, 0$
    $N = 0$
end procedure
procedure RESET( )
    $\overline{x^h}, \overline{\epsilon^h} = 0, 0$
    $N = 0$
end procedure
procedure UPDATE($x^h_{new}, \epsilon^h_{new}$)
    if N == 24 then
        $x^d = \overline{x^h}$
        $\epsilon^d = \overline{\epsilon^h}$
        $RESET()$
    end if
    $N = N + 1$
    $\overline{x^h} = \frac{1}{N} \cdot (\overline{x^h} \cdot (N-1) + x^h_{new})$
    $\overline{\epsilon^h} = \frac{1}{N} \cdot (\overline{\epsilon^h} \cdot (N-1) + \epsilon^h_{new})$
end procedure
procedure RD( )
    INIT()
end procedure
```

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,

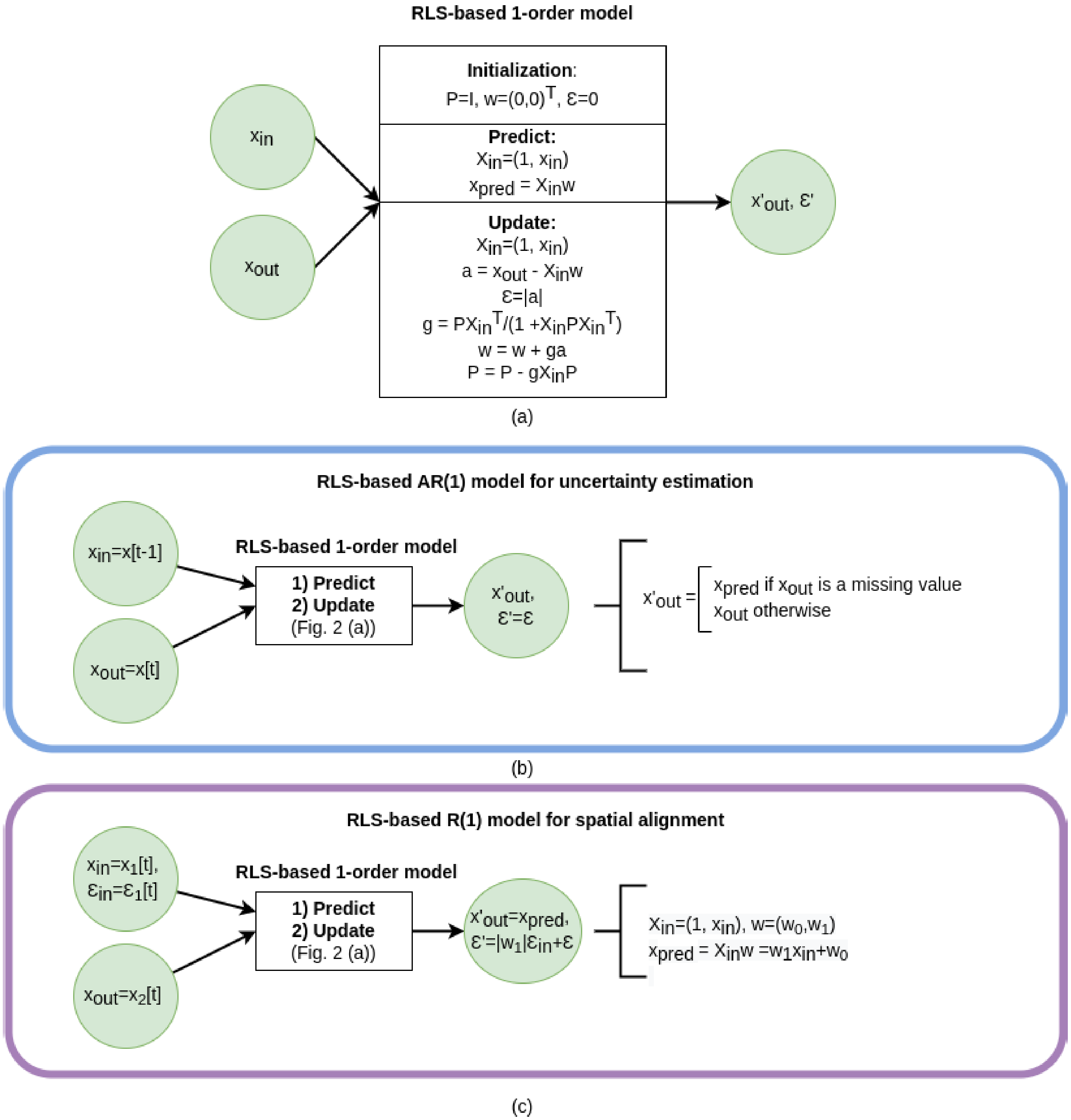


FIGURE 2. Recursive algorithms for least-squares data assimilation (LSDA) from [22]. (a) recursive least squares (RLS)-based first-order model used as a core for data-driven uncertainty estimation and spatio-temporal alignment (calibration). (b) RLS-based first-order autoregression AR(1) model for sequential 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

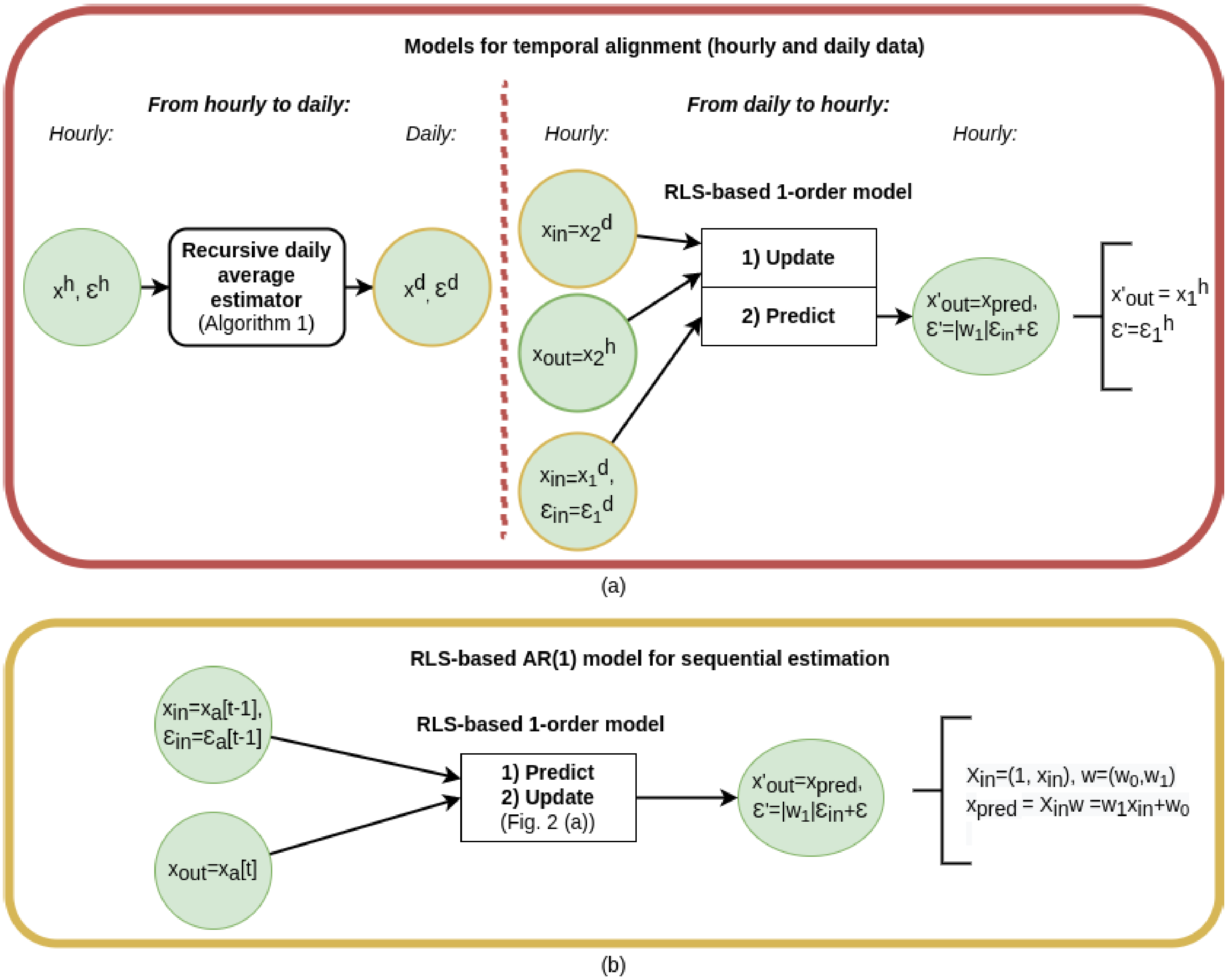
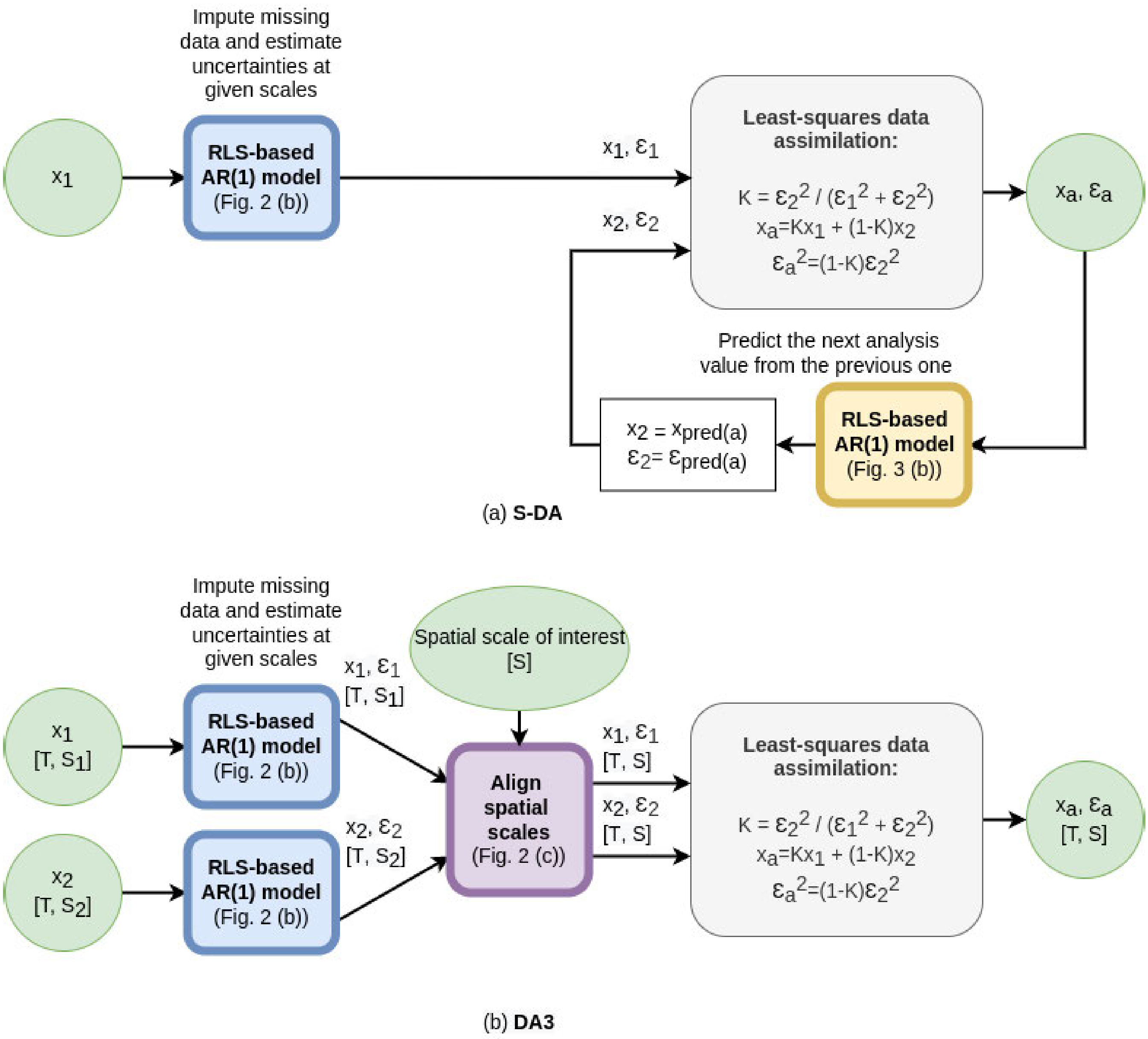

Models for temporal alignment (hourly and daily data)
From hourly to daily:
Hourly:
Daily:
x^h, ε^h
Recursive daily average estimator (Algorithm 1)
x^d, ε^d
From daily to hourly:
Hourly:
Hourly:
RLS-based 1-order model
x_in=x_2^d
x_out=x_2^h
x_in=x_1^d, ε_in=ε_1^d
1) Update
2) Predict
x'_out=x_pred, ε'=|w_1|ε_in+ε
x'_out = x_1^h
ε'=ε_1^h
(a)
RLS-based AR(1) model for sequential estimation
x_in=x_a[t-1], ε_in=ε_a[t-1]
x_out=x_a[t]
RLS-based 1-order model
1) Predict
2) Update
(Fig. 2 (a))
x'_out=x_pred, ε'=|w_1|ε_in+ε
X_in=(1, x_in), w=(w_0,w_1)
x_pred = X_in w =w_1x_in+w_0
(b)

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 (2×2) 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.


Impute missing data and estimate uncertainties at given scales
x1
RLS-based AR(1) model (Fig. 2 (b))
x1, ε1
x2, ε2
Least-squares data assimilation:
K = ε2^2 / (ε1^2 + ε2^2)
xa=Kx1 + (1-K)x2
εa^2=(1-K)ε2^2
xa, εa
Predict the next analysis value from the previous one
x2 = xpred(a)
ε2= εpred(a)
RLS-based AR(1) model (Fig. 3 (b))
Impute missing data and estimate uncertainties at given scales
Spatial scale of interest [S]
x1 [T, S1]
x2 [T, S2]
RLS-based AR(1) model (Fig. 2 (b))
RLS-based AR(1) model (Fig. 2 (b))
x1, ε1 [T, S1]
x2, ε2 [T, S2]
Align spatial scales (Fig. 2 (c))
x1, ε1 [T, S]
x2, ε2 [T, S]
Least-squares data assimilation:
K = ε2^2 / (ε1^2 + ε2^2)
xa=Kx1 + (1-K)x2
εa^2=(1-K)ε2^2
xa, εa [T, S]

(a) **S-DA**

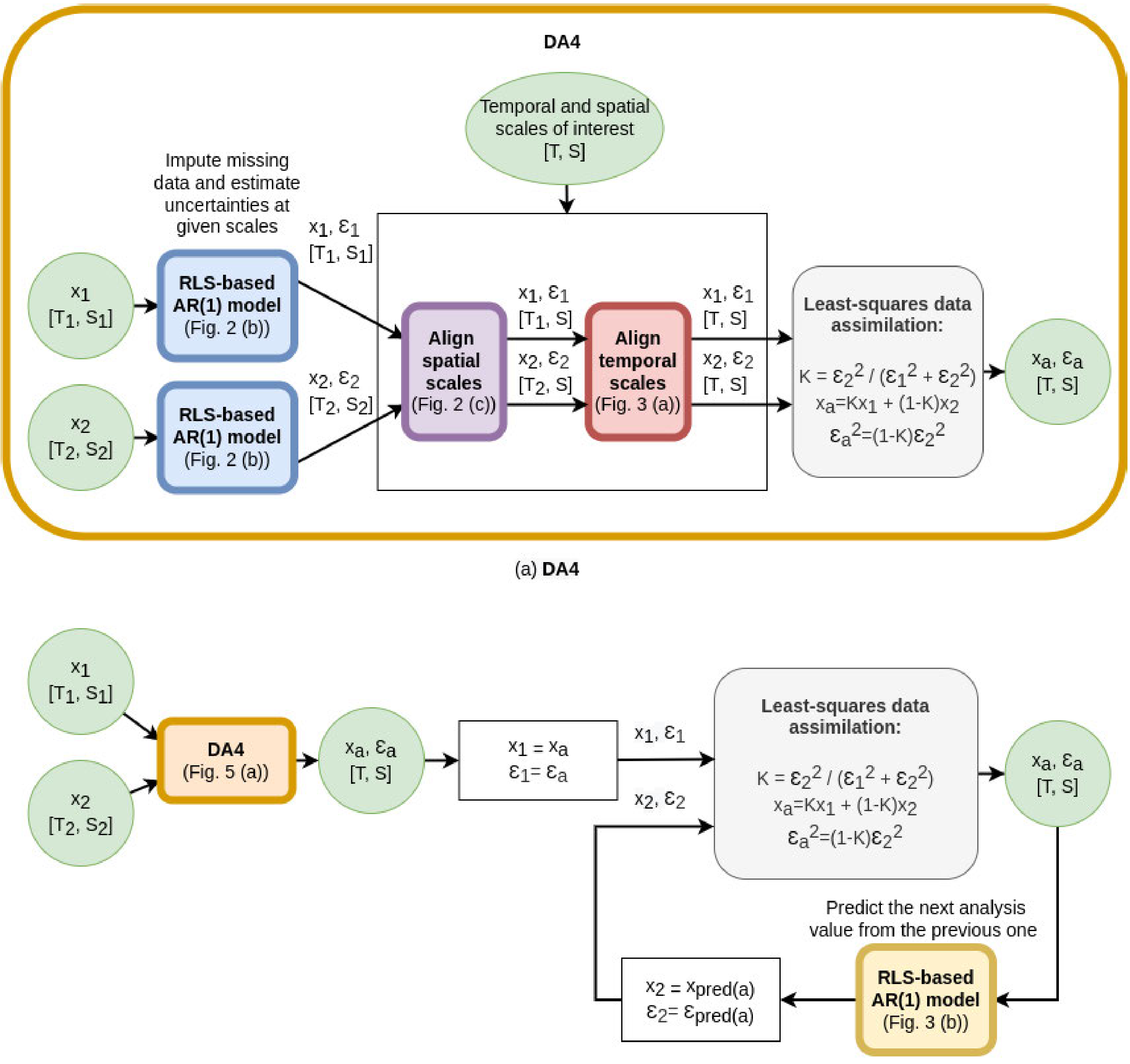
(b) **DA3**

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


DA4
Temporal and spatial scales of interest [T, S]
Impute missing data and estimate uncertainties at given scales
x1 [T1, S1]
x2 [T2, S2]
RLS-based AR(1) model (Fig. 2 (b))
RLS-based AR(1) model (Fig. 2 (b))
x1, ε1 [T1, S1]
x2, ε2 [T2, S2]
Align spatial scales (Fig. 2 (c))
x1, ε1 [T1, S]
x2, ε2 [T2, S]
Align temporal scales (Fig. 3 (a))
x1, ε1 [T, S]
x2, ε2 [T, S]
Least-squares data assimilation:
K = ε2^2 / (ε1^2 + ε2^2)
xa=Kx1 + (1-K)x2
εa^2=(1-K)ε2^2
xa, εa [T, S]
x1 [T1, S1]
x2 [T2, S2]
DA4 (Fig. 5 (a))
xa, εa [T, S]
x1 = xa
ε1= εa
x1, ε1
x2, ε2
Least-squares data assimilation:
K = ε2^2 / (ε1^2 + ε2^2)
xa=Kx1 + (1-K)x2
εa^2=(1-K)ε2^2
xa, εa [T, S]
Predict the next analysis value from the previous one
RLS-based AR(1) model (Fig. 3 (b))
x2 = xpred(a)
ε2= εpred(a)

(a) **DA4**

(b) **S-DA4**

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: $(1 - \frac{\text{RMSE(true, assimilated)}}{\text{RMSE(true, less accurate source)}}) \cdot 100\%$.

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

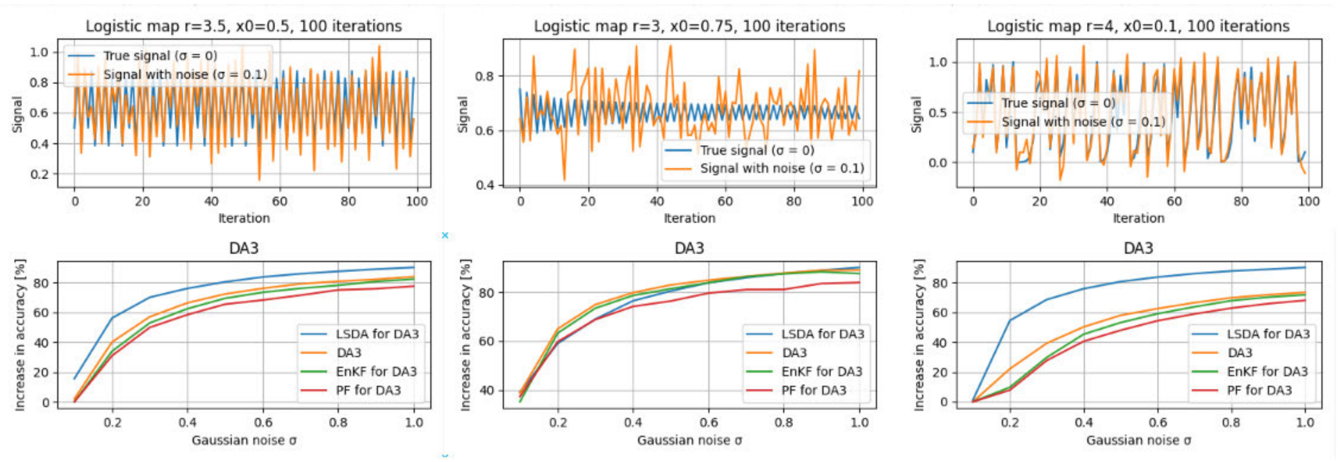


FIGURE 6. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labeled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members 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 0.2Â° 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 (SO_2), nitrogen dioxide (NO_2), carbon monoxide (CO) and ozone (O_3), and particulate matter ($PM_{2.5}$ and PM_{10}).

SILAM generates global 4-day forecasts of AQ data including SO_2, NO, NO_2, O_3, $PM_{2.5}$, and PM_{10}. 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 NO_2, 462 stations for O_3, 137 stations for SO_2, 254 stations for $PM_{2.5}$, and 445 stations for PM_{10}. 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 observations) and Fig. 8 (DA4 and S-DA4, calibration to model simulations).

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 × 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]

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.

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

where $\boldsymbol{x1}, \boldsymbol{x2}$ are vectors of data values of length n from 2 data sources.

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

where $\boldsymbol{\epsilon}$ 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.

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

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

where $\boldsymbol{\epsilon}$ 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 $DA3$,

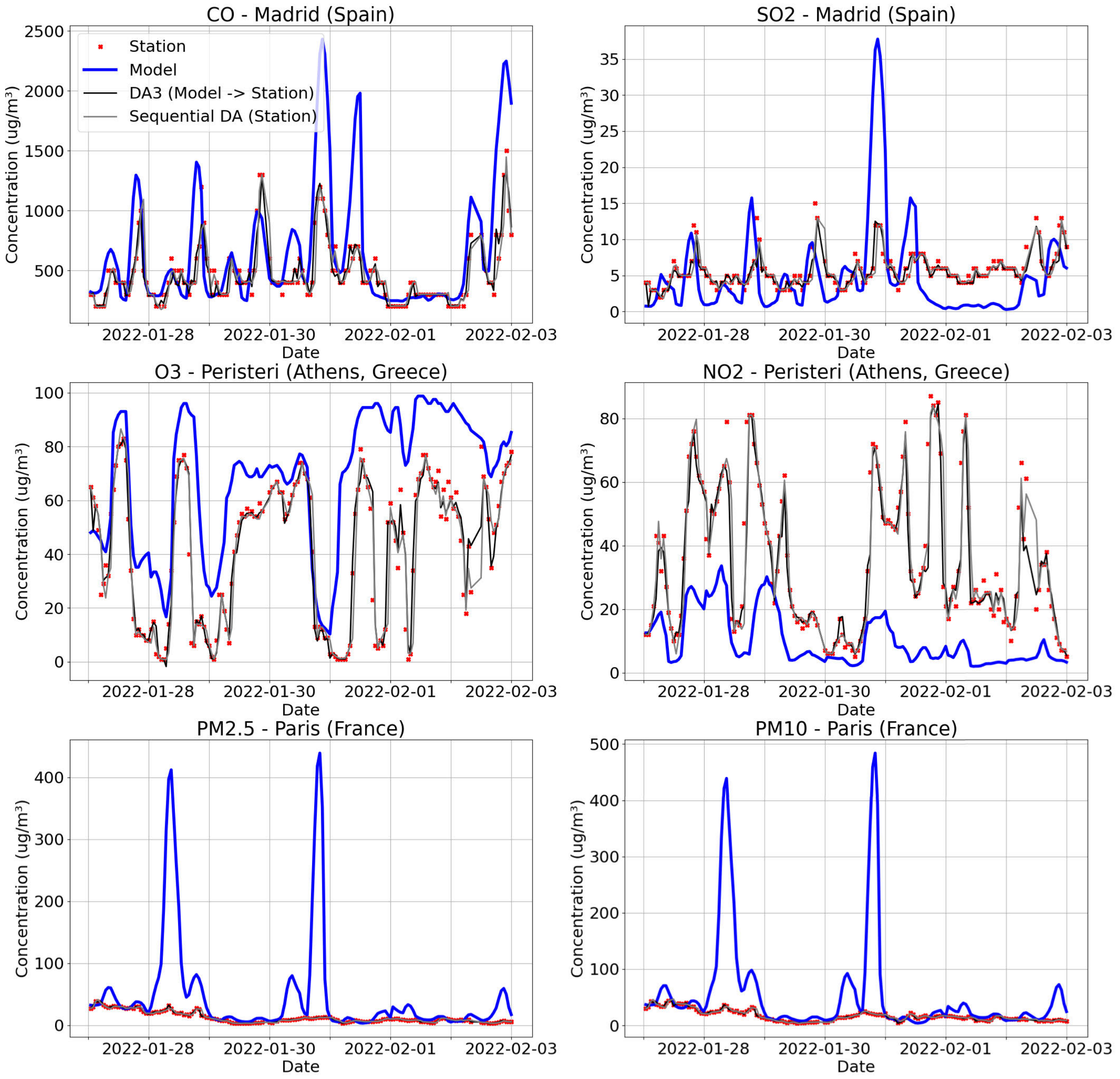


FIGURE 7. Time series plots of input data sources and assimilated values for CO, SO_2, $PM_{2.5}$, NO_2, O_3, PM_{10} AQ variables. "Station" corresponds to observations made by the AQ monitoring stations in Madrid (Spain, CO, SO_2), Peristeri (Athens, Greece, O_3, NO_2), Paris (France, $PM_{2.5}$, PM_{10}). "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 *DA*3 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, $\boldsymbol{x}_m^h$ with the last available daily averages from the previous day, $\boldsymbol{x}_m^d$. We define hourly as the temporal scale of interest, T and the spatial

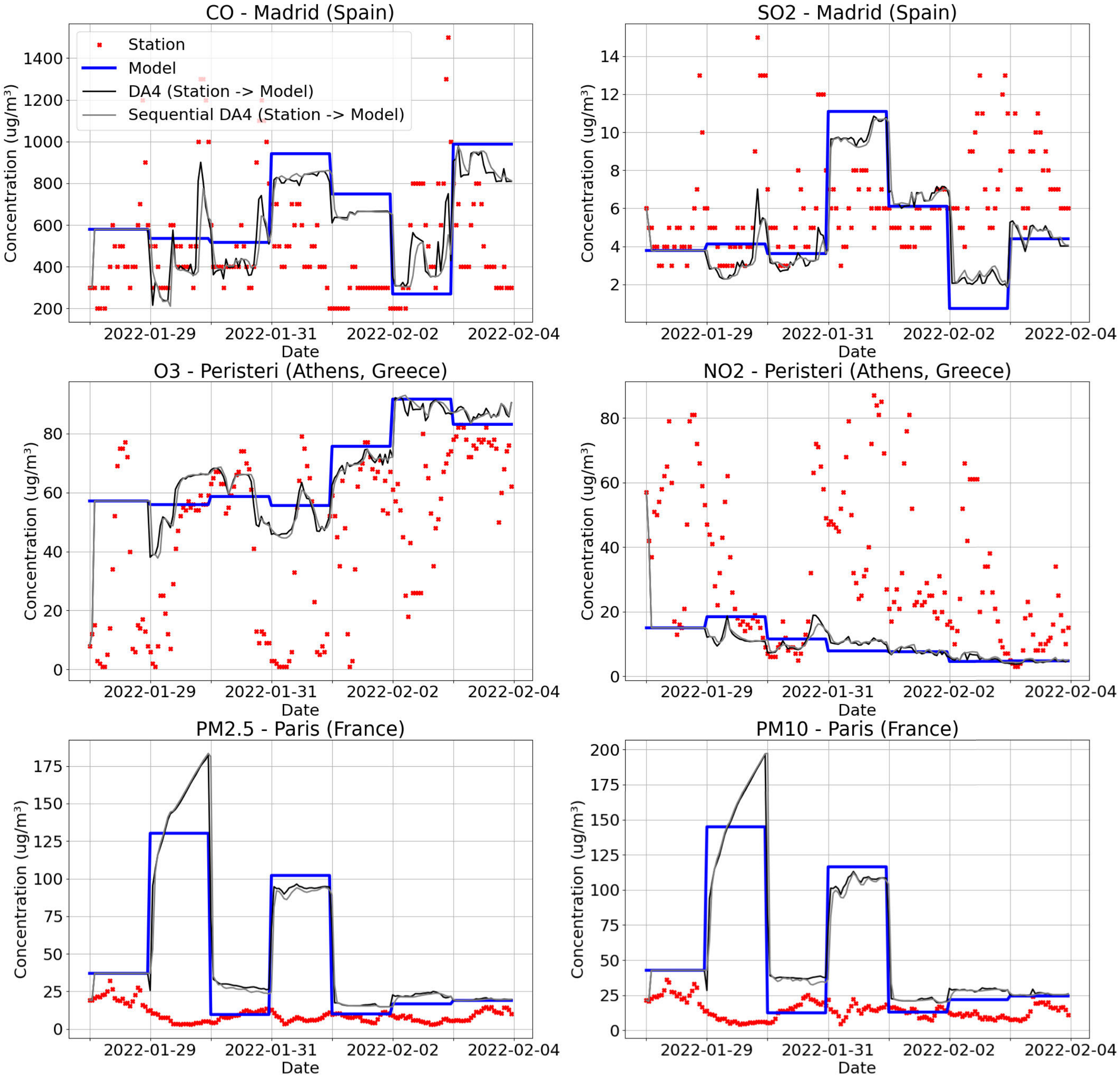


FIGURE 8. Time series plots of the input data and assimilated values for CO, SO_2, $PM_{2.5}$, NO_2, O_3, PM_{10} AQ variables. "Station" corresponds to observations made by the AQ monitoring stations in Madrid (Spain, CO, SO_2), Peristeri (Athens, Greece, O_3, NO_2), Paris (France, $PM_{2.5}$, PM_{10}). "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 ± sd [min; max]	r_{MAU} mean ± sd [min; max]
CO (N=86)	1.116 ± 0.149 [0.660; 1.621]	0.925 ± 0.056 [0.797; 1.034]
SO_2 (N=137)	1.043 ± 0.172 [0.286; 2.000]	0.914 ± 0.083 [0.689; 1.128]
$PM_{2.5}$ (N=254)	1.158 ± 0.181 [0.792; 1.865]	0.902 ± 0.064 [0.575; 1.372]
NO_2 (N=593)	1.257 ± 0.182 [0.892; 2.036]	0.919 ± 0.044 [0.771; 1.262]
O_3 (N=462)	1.367 ± 0.176 [0.651; 2.068]	0.902 ± 0.040 [0.801; 1.273]
PM_{10} (N=445)	1.136 ± 0.156 [0.697; 1.923]	0.899 ± 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)		
	$r^{d\to h}_{RMSE}$ for DA4 mean ± sd [min; max]	$r^{d\to h}_{RMSE}$ for S-DA4 mean ± sd [min; max]	$r^{d\to h}_{MAU}$ mean ± sd [min; max]
CO (N=86)	0.980 ± 0.125 [0.708; 1.242]	0.933 ± 0.102 [0.684; 1.215]	0.593 ± 0.024 [0.536; 0.639]
SO_2 (N=137)	0.972 ± 0.064 [0.863; 1.472]	0.967 ± 0.043 [0.849; 1.116]	0.638 ± 0.074 [0.480; 0.869]
$PM_{2.5}$ (N=254)	0.959 ± 0.074 [0.799; 1.410]	0.969 ± 0.072 [0.819; 1.403]	0.601 ± 0.046 [0.336; 0.818]
NO_2 (N=593)	0.911 ± 0.062 [0.770; 1.335]	0.926 ± 0.057 [0.795; 1.339]	0.592 ± 0.037 [0.415; 0.823]
O_3 (N=462)	0.868 ± 0.084 [0.675; 1.449]	0.885 ± 0.080 [0.697; 1.466]	0.614 ± 0.033 [0.479; 0.770]
PM_{10} (N=445)	0.940 ± 0.075 [0.787; 1.611]	0.949 ± 0.075 [0.793; 1.665]	0.594 ± 0.038 [0.509; 0.790]

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

$$r^{d\to h}_{RMSE} = \frac{RMSE(\boldsymbol{x_a}; \boldsymbol{x^h_m})}{RMSE(\boldsymbol{x^d_m}; \boldsymbol{x^h_m})}, \tag{5}$$

where $\boldsymbol{x_a}$ are analysis values for the DA4 and S-DA4 algorithms.

The MAU ratios $r^{d\to h}_{MAU}$ are calculated similarly to Equation (4), but by dividing $MAU(\boldsymbol{\epsilon_{a(S-DA4)}})$ by $MAU(\boldsymbol{\epsilon_{a(DA4)}})$.

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 for $PM_{2.5}$, NO_2, O_3 and PM_{10} and S-DA4 for CO and SO_2. 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.

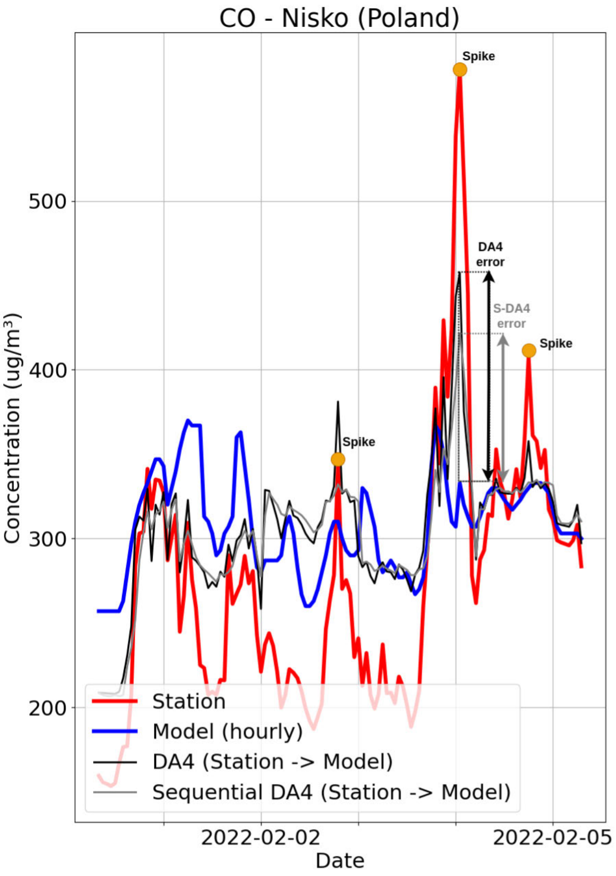


FIGURE 9. Demonstration of situations when S-DA4 can outperform DA4. "Station" corresponds to observations made by the Nisko AQ monitoring station (Nisko, Poland), "Model (hourly)", hourly SILAM simulations, "DA4 (Station → Model)", algorithm DA4 with calibration of hourly station observations to daily model simulations, "Sequential DA4 (Station → 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.

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

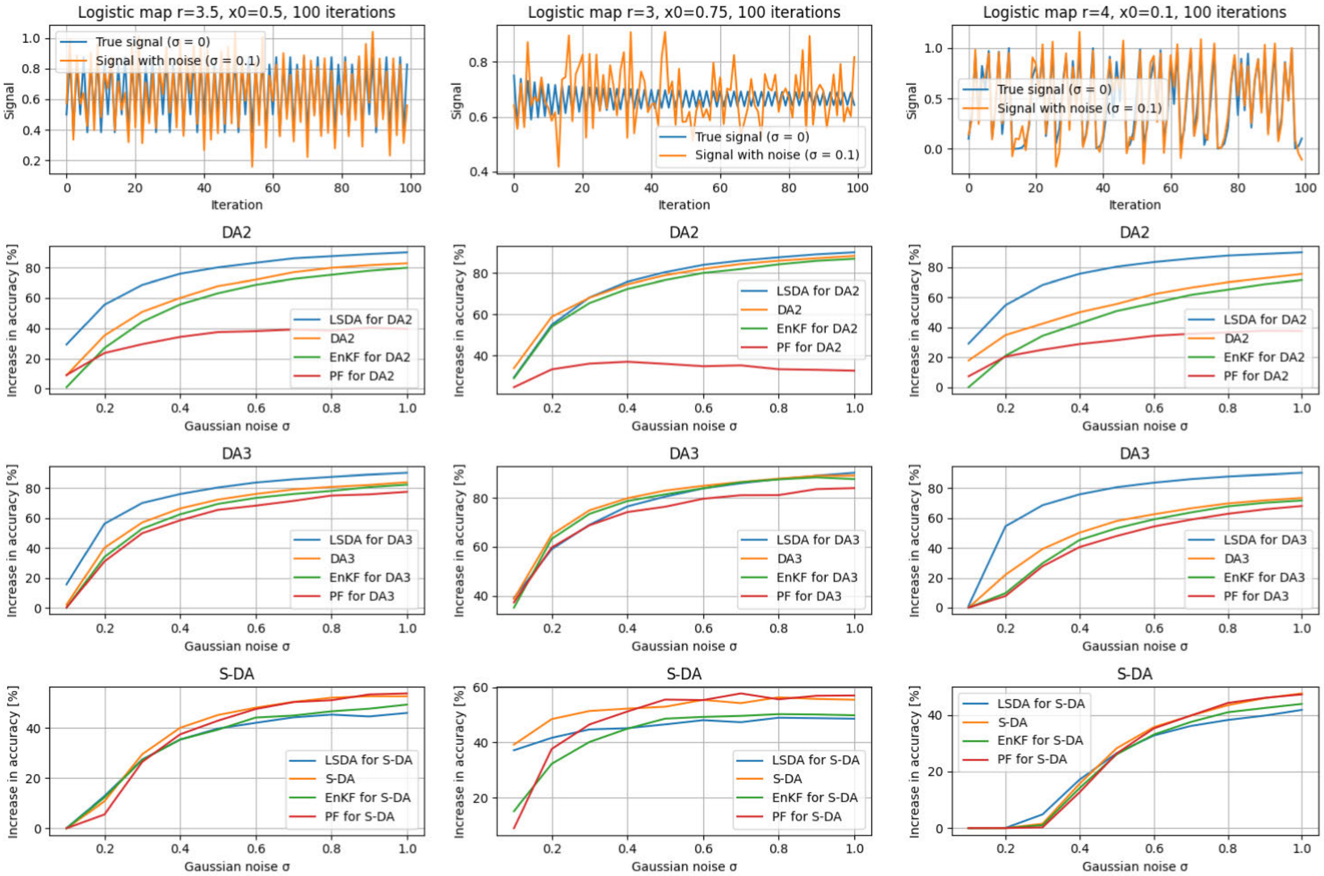


FIGURE 10. **Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members 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 0.2Â° and 0.4Â° spatial resolution, the algorithms allow only for retrieving analyses of 0.2Â° or 0.4Â° 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 of 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

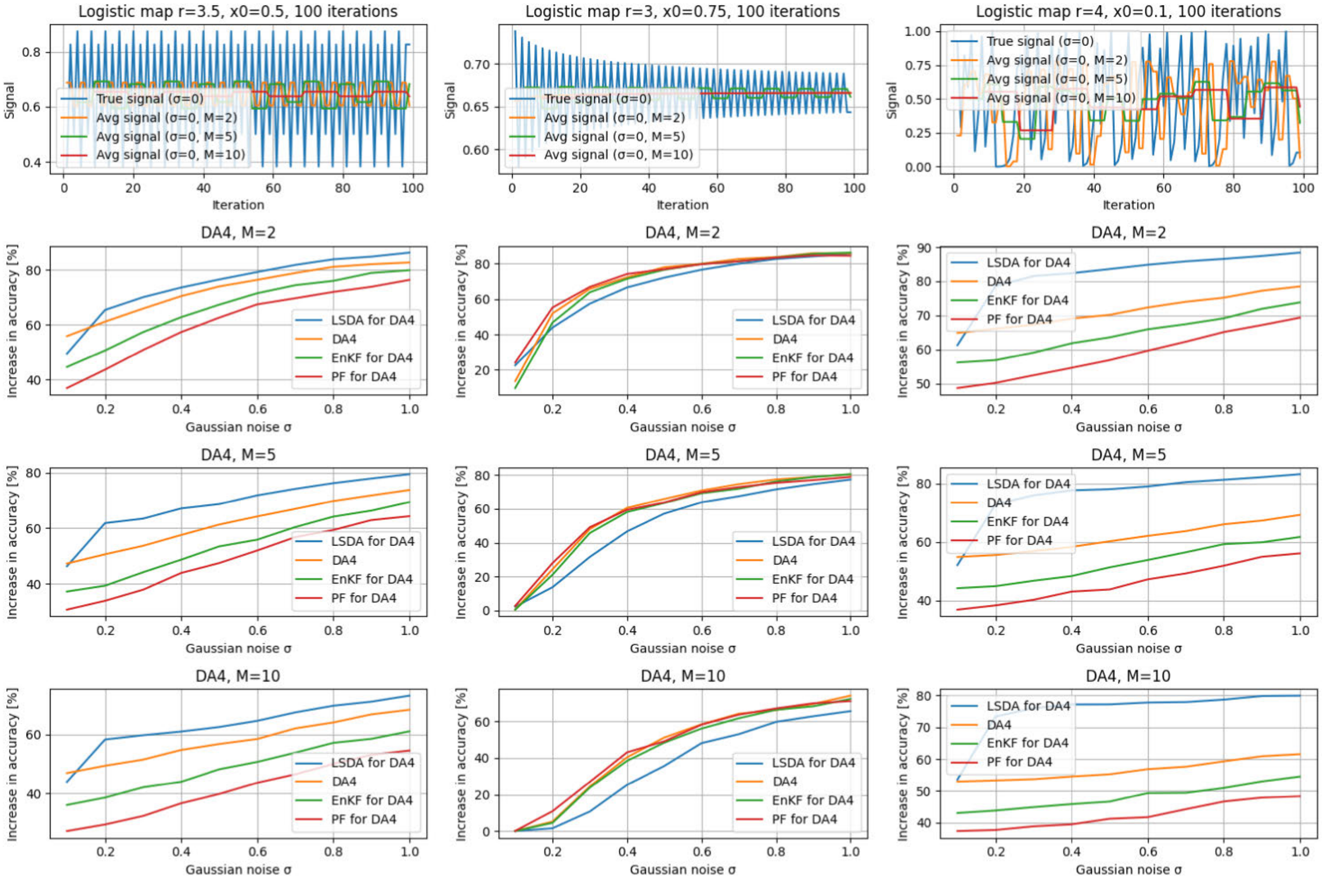


FIGURE 11. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members 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

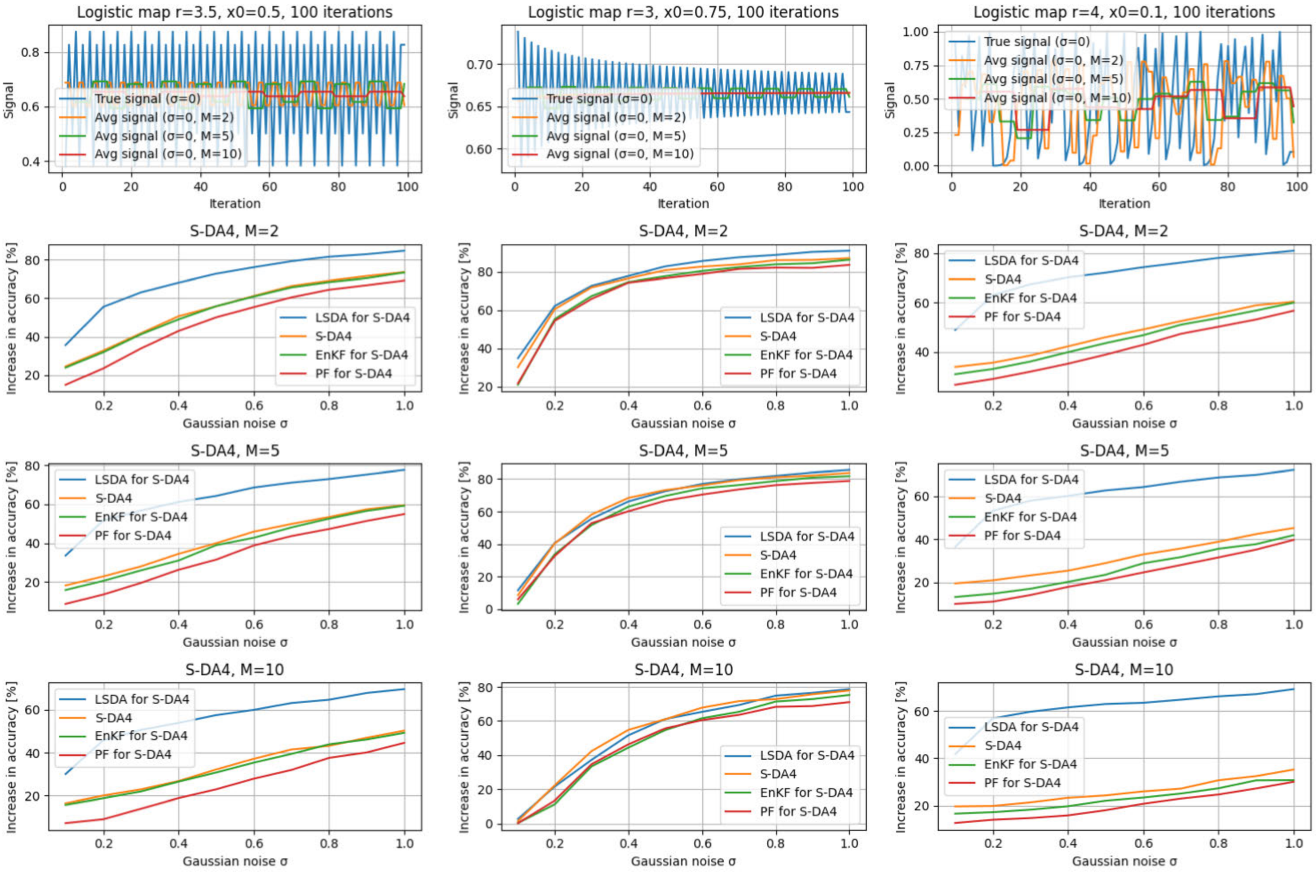


FIGURE 12. Sensitivity analysis for LSDA with known uncertainties, LSDA with unknown uncertainties (uncertainties are estimated using the authors' methods, labelled by the name of a scenario), ensemble Kalman filter (EnKF, uncertainties are estimated using the authors' methods, number of ensemble members 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. PM_{10}, $PM_{2.5}$, Ozone, NO_2 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 NO_2 and O_3 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 V.5.7: 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.

• • •