Information Fusion 131 (2026) 104124

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Data fusion for low-cost sensors: A systematic literature review

Gabriel Oduori [a,*], Chaira Cocco [a], Payam Sajadi [b,c], Francesco Pilla [a]

[a] School of Architecture, Planning and Environmental Policy, Spatial Dynamics Lab, University College Dublin, Belfield, Dublin, D04 C1P1, Leinster, Ireland
[b] Irish Climate Analysis and Research UnitS (ICARUS), Department of Geography, Maynooth University, Maynooth, Ireland
[c] Water Management Consultancy (WMC), AtkinsRealis, Dublin, Ireland

ARTICLE INFO

Keywords:
Data fusion
Low-cost sensors
Environmental monitoring
Systematic literature review
Satellite imagery
Geostatistics
Machine learning
Spatio-temporal data
Uncertainty quantification

ABSTRACT

Data fusion (DF) addresses the challenge of integrating heterogeneous data sources to improve decision-making and inference. Although DF has been widely explored, no prior systematic review has specifically focused on its application to low-cost sensor (LCS) data in environmental monitoring. To address this gap, we conduct a systematic literature review (SLR) following the PRISMA framework, synthesising findings from 82 peer-reviewed articles. The review addresses three key questions: (1) What fusion methodologies are employed in conjunction with LCS data? (2) In what environmental contexts are these methods applied? (3) What are the methodological challenges and research gaps? Our analysis reveals that geostatistical and machine learning approaches dominate current practice, with air quality monitoring emerging as the primary application domain. Additionally, artificial intelligence (AI)-based methods are increasingly used to integrate spatial, temporal, and multimodal data. However, limitations persist in uncertainty quantification, validation standards, and the generalisability of fusion frameworks. This review provides a comprehensive synthesis of current techniques and outlines key directions for future research, including the development of robust, uncertainty-aware fusion methods and broader application to less-studied environmental variables.

1. Introduction

1.1. Context

Rapid pace of urbanisation, coupled with population growth and climate change, continues to challenge the sustainability and liveability of urban environments [1]. To respond to these multi-faceted challenges, researchers are continuously adopting new and innovative approaches to measure, monitor and manage environmental impacts. At the core of these advancements lies the integration of diverse environmental datasets and sources, ranging from traditional topographic maps, satellite imagery, and more recently, social media, low-cost sensors (LCS) integrated with Internet of Things(IoT).

Traditionally, environmental monitoring has often been done by a network of regulatory monitoring sites, strategically located, to enable continuous reporting of concentration levels of parameters of importance [2]. These standard-graded stations, while provide high quality data, usually require high initial and thereafter running costs. Consequently, they are never adequate to cover wider spatial areas and support community-wide exposure assessment [2–4], leading to a coarse spatial resolution [5].

Information science research regarding the development of sensing systems focuses on how information can be extracted from sensory data [6]. Recent advancements in computing, sensor technology, and wireless communication have resulted in a new paradigm in environmental monitoring [3]. More importantly, the improvements in wireless data transmission and location-sensing devices contribute to real-time data collection [1]. These developments have heralded a new era of LCS. The LCSs are continuously being included as part of a network of individuals and sensors participating in activities of environmental monitoring [7].

LCS are affordable devices, designed to measure and collect data on environmental, physical, or chemical parameters. Compared to traditional sensors, LCS are significantly more economical, making them accessible for widespread use in various applications, including environmental monitoring, home automation, and Internet of Things (IoT) projects. Snyder et al. [8] describes these types of sensor platforms as relatively inexpensive, easier to use, and less bulky compared to traditional

* Corresponding author.
E-mail addresses: gabriel.oduori@ucdconnect.ie (G. Oduori), payam.sajadi@mu.ie (P. Sajadi).

https://doi.org/10.1016/j.inffus.2026.104124
Received 28 June 2025; Received in revised form 20 November 2025; Accepted 4 January 2026
Available online 18 January 2026

Table 1
Characteristics of existing literature reviews.

Paper	Year	Coverage	Objective and topics
Castanedo, [15]	2013	DF Techniques	The review focused on data fusion techniques
Alam et al.[16]	2017	DF in IoT	Data fusion application methods for IoT environment.
Lau et al.[17]	2019	DF General applications	Smart City Fusion applications.
Ding at al.[18]	2019	Local	Data Privacy and security
Krishnamurthi et al.[19]	2020	DF general techniques	IoT sensor data techniques including processing, analysis, and fusion.
Meng et al.[20]	2020	DF General applications	Application of machine learning in data fusion.
Ouhami et al.[21]	2021	DF crop disease	Explores Computer Vision, IoT and data fusion with a focus on crop pests..
Himeur et al.[22]	2022	DF General application	Artificial intelligence and data fusion for environmental impacts monitoring dams construction using RS images.
Karagiannopoulou et al.[23]	2022	DF general	Scoping review presenting data fusion algorithms and methodological procedures with a focus on possible and demonstrable applications.
Ounoughi and Ben H.[24]	2023	DF in intelligent transport systems	Fusion techniques, applications to extract issues, and challenges of using these techniques in intelligent transportation systems.
Fadhel et al.[25]	2024	IF methods in smart cities and urban environments	Examined smart city applications in detail, incorporating quality evaluation and IF techniques and identifying critical issues while outlining promising research directions.

equipment. Furthermore, they enable citizens and communities to monitor their local air quality, which may directly affect their health. A more recent work by [9] defines LCS units as electronic sensing devices that cost several orders of magnitude less than existing reference instruments. Additionally, [10] provides a more detailed definition of what constitutes a low-cost sensor, from a costing point of view.

Data Fusion (DF) addresses the challenge of how to combine or fuse data from these multiple sources. This combination is crucial for making decisions, and enables the merging of information, usually with the primary aim of forming a unified picture. DF systems are widely used in numerous fields, including sensor networks, robotics, image processing, cybersecurity, and intelligent transport systems, among many other applications. As a broad-ranging subject, many terminologies are used interchangeably across various research fields and applications. For instance, the JDL model [11,12], and the more recent review by [13] highlight the diverse definitions and applications of DF. These studies not only demonstrate a growing interest in the field but also provide definitions that fit most domains. Most researchers now agree with the following definition from Khaleghi et al (2013):

"Information fusion is the study of efficient methods for automatically or semi-automatically transforming information from different sources and different points in time into a representation that provides effective support for human or automated decision-making." [14]

1.2. Objective and scope

We find the interconnection between the domain of DF and that of LCS an exciting area of research and worthy of pursuit. To the best of our knowledge, previous studies reviewing the topic of DF have not focused on LCS. This review aims to fill that gap. By synthesising findings from peer-reviewed articles, we seek to explore key fusion methodologies, environmental parameters measured, levels of data fusion applied, and the thematic applications where data fusion has been implemented. Ultimately, this review provides insights into the current state of research on data fusion in LCS and offers a roadmap for future research in the field.

Several studies have systematically investigated the topic of data fusion with sensors and IoT. These studies have focused either on the technology, models, methodologies or applications. Table 1 summarises existing research. For instance, in their review of the fusion of IoT data, [16], focused on mathematical methods with specific attention to different IoT environments. They outlined three main algorithms namely probability probabilistic, artificial intelligence (AI), based fusion methods (including supervised, Artificial Neural Networks, ANN), and fuzzy logic. A key relevant mention by the authors is the application of Support Vector Machine (SVM)-based fusion in remote sensing for multispectral and pan-chromatic data. The authors did not include low-cost IoT environments. Regarding privacy and security of IoT devices for data collection, [18] focused on fusion for smart homes, grids, and transport systems. They proposed several privacy-preserving data fusion strategies as well as emerging research trends. They did not mention low-cost sensors in their research. Lau et al. [17] review focused on data application and fusion in smart city environments using standard-grade sensing technologies. They further proposed a multi-purpose classification for evaluating domain-specific fusion within an urban setting. A further review focusing on various IoT sensor data techniques, such as data processing, data analysis, and data fusion is offered by [19].

In a multi-disciplinary approach, [21] blended computer vision, Internet of Things data for crop disease detection, and machine learning. They concluded that the current technologies have limitations when focusing in earlier disease detection and hence the interest in multimodal data fusion amongst the research community. They also showed that the most widely used type of fusion in agriculture is the integration of multi-sensors data from aerial vehicles, fusion of multi-resolution satellites data and the fusion of satellite and UAV images. This fusion is employed to improve the detection process for tasks such as crop monitoring or plant classification." A trend towards the application of AI in data fusion continues to grow. For instance, [20] conducted a review that focused on data fusion using machine learning. A further study in the use of AI in data fusion is presented by [22]. A further analysis of fusion is presented by [23]. In this review, the authors took a special focus on the role of citizen science data. A more recent review focused on Intelligent Transport System (ITS) [24]. While there is a mention of low-cost, no further discussion is made regarding the use of these low-cost sensors especially within an environmental sensing context.

From the state of the art analysis, it is evident that a large amount of literature on data fusion already exists. However, these reviews have majorly focused on specific applications or industry grade sensor data. This leaves a gap with no publication dedicated to summarising the data fusion with data involving low-cost sensor network and environmental monitoring.

The aim of this paper therefore is three folds:

- Conduct a review of data fusion with a primary focus on low-cost sensors
- Outline key fusion algorithms and methods currently being used with low-cost sensors
- Identify application areas for fusion with low-cost sensor

Table 2
List of abbreviations.

Abbreviation	Description
ADMS	Atmospheric Dispersion Modelling System
AI	Artificial Intelligence
ANN	Artificial Neural Network
AOD	Aerosol Optical Depth
AQMS	Air Quality Monitoring Systems
BLUP	Best Linear Unbiased Predictor
CAMS	Copernicus Atmosphere Monitoring Service
CMAQ	Community Multiscale Air Quality
CNN	Convolutional Neural Network
CTM	Chemical Transport Model
CRPS	Continuous Ranked Probability Score
DBN	Deep Belief Network
DEM	Digital Elevation Models
DF	Data Fusion
DL	Deep Learning
DT	Decision Trees
ENFUSER	ENvironmental information FUsion SERvice
EPA	Environmental Protection Agency
ESCAPE	The European Study of Cohorts for Air Pollution Effects
EVT	Expected Value Theory
ExtraTrees	Extremely Randomized Trees
GP	Gaussian Process
IDW	Inverse Distance Weighting
IoT	Internet of Things
JDL	Joint Defence of Laboratory
KF	Kalman Filter
LCS	Low-Cost Sensor
LGBM	Light Gradient-Boosting Machine
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short Term Memory
LUR	Land Use Regression
PRISMA	Preferred Reporting Items for Systematic reviews and Meta-Analyses
ML	Machine Learning
MLP	Multilayer Perceptron
MLR	Machine Learning Regression
pXRF	portable X-ray Fluorescence
RF	Random Forest
RFR	Random Forest Regression
RMSE	Root Mean Squared Error
SGPR	Sparse Gaussian Process Regression
SHAP	SHapley Additive exPlanations
SILAM	System for Integrated modeLling of Atmospheric coMposition
SLR	Systematic Literature Review
SRTM	Shuttle Radar Topography Mission
SVM	Support Vector Machine
UK	Universal Kriging
vis-NIR	visible Near-InfraRed
XGBoost	Extreme gradient boosting

This paper therefore conducts a systematic review of recent literature on data fusion techniques and applications to extract methodologies, parameters and algorithms and the use of these techniques in low-cost sensors for environmental monitoring.

1.3. Contribution

Despite the rapid evolution of data fusion techniques in environmental monitoring, existing reviews have largely been limited to specific domains (such as air quality or hydrology), high-cost sensor networks, or the development of isolated fusion algorithms [7,26]. A substantial gap remains for a systematic, cross-domain review that critically addresses the particular demands and complexities associated with low-cost sensor (LCS) networks, systems now central to citizen science, urban analytics, and scalable environmental monitoring. This manuscript addresses that gap through several novel and impactful contributions. First, we provide the most comprehensive and up-to-date systematic review focused specifically on data fusion methods for LCS, synthesising not only traditional geostatistical and machine learning approaches, but also probabilistic, knowledge-driven, and emerging hybrid frameworks. Our work rigorously compares these methods across diverse real-world application domains, highlighting both their methodological strengths and their practical limitations under the noisy, heterogeneous, and resource-constrained realities of LCS deployments. Second, and uniquely, we introduce a structured and extensible taxonomy of data fusion techniques tailored to the low-cost sensor context. This taxonomy goes beyond previous classification efforts by organising fusion strategies according to their mathematical underpinnings, operational requirements, data quality dependencies, and target application domains. It serves as both a navigational aid for practitioners and a conceptual foundation for further research, supporting systematic model selection, benchmarking, and reproducibility. Third, we critically review and integrate cutting-edge advances in privacy-preserving data fusion, uncertainty quantification, and secure inference, areas of growing relevance for LCS networks but rarely treated in depth in previous reviews. Our analysis spans techniques such as geo-indistinguishability, federated learning, and cryptographically secure inference, directly addressing the new challenges that arise as LCS networks move from experimental deployments to large-scale, citizen-facing, or policy-relevant applications. Fourth, this review adopts a practical, deployment-oriented lens, systematically addressing issues of sensor calibration, data quality control, communication infrastructure, computational requirements,

and privacy/security risks. By synthesising recent field experience and technical innovation, we provide actionable recommendations and best practices for practitioners tasked with deploying robust data fusion systems in challenging real-world settings. Finally, we identify and articulate outstanding challenges and research gaps that must be addressed for data fusion with LCS to fulfil its scientific and societal promise. These include needs for standardized benchmarking datasets, more interpretable hybrid models, real-time uncertainty management, and robust privacy-preserving protocols. By providing both a clear synthesis of the state of the art and a strategic roadmap for future innovation, this paper aspires to set a new benchmark and reference point for the interdisciplinary community advancing data fusion for low-cost sensor networks. We have also explained all the abbreviated terms in Table 2.

1.4. Structure of the paper

The rest of the article is organized as follows: Our search and extraction of information from key databases is described in Methodology and Research protocol in Section 2. Section 3 provides a taxonomy for organising our review before presenting results of the review in Section 4. Section 5 is focused on the discussion of the results. The final Section 6 concludes the article with suggestions for future research.

2. Methods and protocol

Systematic Literature Review (SLR) is a standard method of extracting, identifying and synthesizing information from currently existing research studies. This is usually done through a systematic procedure [27]. A number of authors have proposed various methods for conducting systematic literature review [27,28]. The current study was accomplished by following recommended guidelines of "Preferred Reporting Items for Systematic reviews and Meta Analyses" (PRISMA) criterion [29,30]. The review is done by performing an exhaustive search of papers and reporting the main findings. For SLR, PRISMA cautions against using a single database search for literature. We therefore followed the protocol outlines and recommended in [28]. We conducted a thorough search against, multiple academic journal databases. Specifically, three prominent digital databases were used to identify relevant literature: Science Direct, Scopus and Web of Science. These databases were selected based on their academic reliability and wider availability of relevant articles to discover the research gap and provide critical implications. Utilising databases that collectively cover a wide range of technological and scientific fields ensures a comprehensive coverage of relevant research.

2.1. Research questions

The key research question in this systematic review was "What is the current status of data fusion and low-cost sensors in environmental monitoring implication? To answer this question, the main question was further split into the following sub-questions:

- What data fusion methods are used with low-cost sensors in environmental monitoring?
- What parameters are being monitored using data fusion?
- In what context are these fusion methods applied?
- What are the challenges and future directions of DF for LCS? applications

2.2. Search strategy

The current study followed PRISMA, an evidence-based minimum set of items for reporting in systematic reviews and meta-analyses. The objective was to identify relevant studies focusing on data fusion, environmental monitoring and low-cost sensors. To achieve this, explicit statements of research were developed as a set of criteria along the so-called PICOC, an acronym for Population of concerns, Interventions for addressing the conditions observed, Comparisons involved in the study, Outcomes on which there are improvements and lastly Context of the study. To encourage researchers to consider elements as their questions, PICO is widely used in social and medical sciences [27]. In this review, we adopt a list of components and terms developed for the field of computer science [28].

This strategy involved identifying words that associated with data fusion, environmental monitoring and low-cost sensors. This approach enabled the study to gather diverse range of studies exploring fusion of data from low-cost sensors, thereby extending research in this new area.

2.3. Selecting digital libraries

An extensive search was conducted on 21stOctober 2023 across key academic platforms, namely Web of Science, and two academic databases viz Scopus and ScienceDirect. These sources were selected primarily due to their availability through the university's digital library database, and secondarily for their size, quality assurance, and broad subject coverage. The inclusion of these three databases also meets the minimum standards for the number of platforms required for a systematic literature search, as outlined by Siddaway et al. (2019) [31].

To identify as many relevant studies as possible, a structured search was conducted using a predefined search string based on the keywords listed in Table 3. The initial search was carried out across the three major databases: Scopus, ScienceDirect, and Web of Science. This yielded 213, 343, and 107 records, respectively. An additional 38 records were identified through manual searches, resulting in a total of 701 records. All records were exported and saved in BibTeX format for further processing.

Finally, to update the dataset and ensure inclusion of more recent studies, a second search was conducted on 14th, April 2025, using the same search string, but with a publication date filter set between 2023 and April 2025. This search returned 178 new records: 94 from Scopus, 60 from ScienceDirect, and 60 from Web of Science. PRISMA workflow in Fig. 1 shows this workflow.

2.4. Inclusion and exclusion criteria

The scope of literature included in this study was restricted based on the following criteria:

i *Publication Timeline:* No a priori date restriction was applied to the primary search to ensure comprehensive coverage. The update search used the same windowing rules solely to avoid double-counting.
ii *Types of sources:* Published, peer-reviewed journal articles were the primary target.
iii *Accessibility:* Abstract and full text available.
iv *Relevance:* Studies explicitly involve low-cost sensors (LCS) and a data-fusion component in environmental monitoring.
v *Study type:*Original primary research presenting implemented methods and/or empirical evaluations. Systematic/structured reviews of data-fusion techniques were admissible for contextualisation only.

All search results (BibTeX) were imported into Rayyan[1], a web-based tool for systematic reviews, for de-duplication and management. Title-abstract screening against the predefined eligibility criteria was conducted by a single reviewer to maintain procedural consistency. To mitigate selection risk, the complete list of included studies and the key extracted methodological attributes (e.g., fusion level, primary method class, sensor type) were independently verified by two co-authors. Any discrepancies or uncertainties were resolved by consensus. This two-stage workflow enhances transparency and supports the completeness

[1] https://www.rayyan.ai/

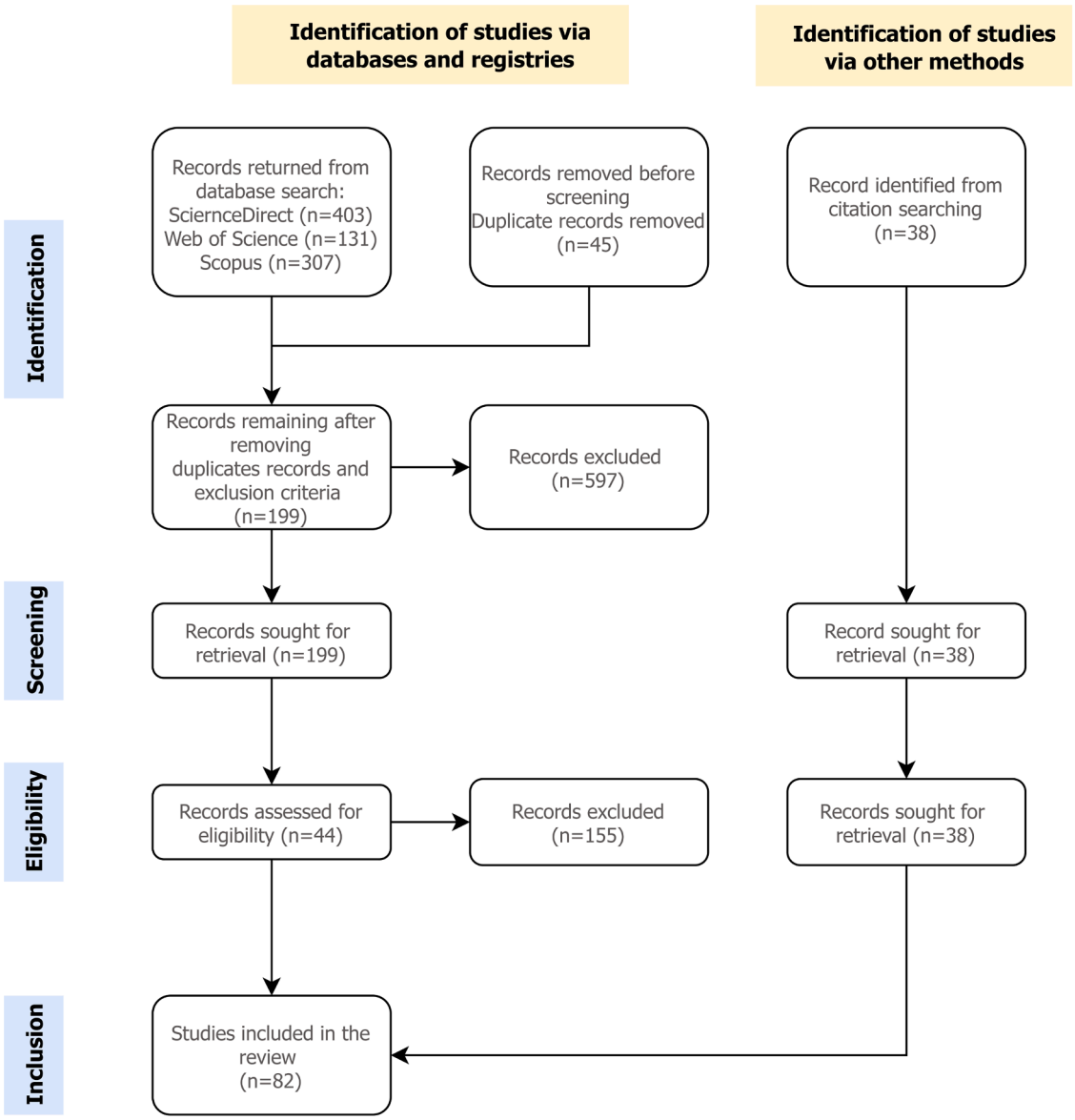

Fig. 1. PRISMA flow diagram showing identification, screening, eligibility, and inclusion. Title and abstract screening was performed by a single reviewer; the final included set and extracted attributes were independently verified by two co-authors, with consensus resolution of any discrepancies.

and internal consistency of the final corpus. Although the inclusion criteria emphasised peer-reviewed journal articles reporting original primary research, we retained a small, explicitly documented subset of sources to enrich context. In total, the 82 studies comprised 75 journal articles, 2 conference papers, 2 review/survey articles, and 1 stakeholder analysis. These exceptions were deliberately included where they provided unique methodological insights, datasets, or deployment/implementation perspectives not captured in the journal literature. They do not contribute to quantitative tabulations (e.g., distributions or performance metrics) unless an implemented method with evaluable results is reported; otherwise, they inform the qualitative synthesis. Any other article that did not meet the eligibility criteria after this process were excluded. The workflow is illustrated in Fig. 1.

2.5. Supplementary search and terminology sensitivity

The primary search strategy was designed to be systematic and reproducible across databases using established formulations of *"low-cost sensor"* terminology. Because related fields sometimes employ alternative descriptors, we conducted an exploratory supplementary search to assess the robustness of corpus coverage to terminology choices, providing additional context for the application distribution reported in Section 4.

Configuration: We queried ScienceDirect, Scopus, and Web of Science using three additional terms, *"affordable sensor," "community monitoring,"* and *"in-situ sensor"*. Database-level de-duplication was applied, followed by cross-database reconciliation against the screened primary set, using the same inclusion window and document types specified in our protocol.

Findings: This supplementary assessment identified 32 additional unique records (after de-duplication) that are thematically aligned with low-cost sensing and data-fusion approaches. Relative to the primary corpus, these records show greater representation in water quality, soil moisture, and noise monitoring. A concise summary is provided in Appendix A to support transparency and reuse.

Integration policy: The supplementary search was conducted after the formal screening cut-off and therefore serves as a terminology-sensitivity check rather than an extension of the quantitative synthesis. Consistent with our protocol and PRISMA reporting, all quantitative

Table 3
Search criteria on different literature sources.

Database	Search String	Articles
ScienceDirect	data fusion AND (environment monitoring OR environmental monitoring) AND ("low-cost sensor" OR "low cost sensor")	403
Scopus	ALL(data AND fusion AND ("environment monitoring" OR "environmental monitoring") AND ("low-cost sensor" OR "low cost sensor")) AND (LIMIT-TO (DOCTYPE,"ar")) AND (LIMIT-TO (LANGUAGE,"English"))	307
Web of Science	data fusion AND (environment monitoring OR environmental monitoring) AND (low-cost sensor OR low cost sensor)	131
Total		841

results and proportions, including those in Section 4, are computed on the finalised cohort of 82 studies from the primary search. The supplementary records (Table A.6) are documented to inform future updates and to facilitate replication.

Implications and reproducibility: Terminology can influence retrieval patterns across adjacent disciplines. To enable straightforward replication and protocol extension, we report the additional terms, date limits, databases, and de-duplication steps here and in Appendix A.

2.6. Quality control and final selection

This section documents the control steps used to ensure basic eligibility and internal consistency of the final corpus. After de-duplication and full-text checks, title-abstract screening against the predefined inclusion criteria was performed by a single reviewer to maintain procedural consistency; the complete list of included studies and key extracted methodological attributes (fusion level, primary method class, sensor type) were then independently verified by two co-authors, with consensus resolution of discrepancies.

To ensure that only high-quality evidence contributed to the review, each identified article was assessed based on the following quality criteria:

- Papers that featured low-cost sensors/IoT and data fusion.
- Exclusion of duplicate articles extracted from more than one database or catalogue.
- Articles published in more than one journal.
- Availability of the paper as a PDF file.

All the articles underwent a multi-stage screening process against these inclusion and exclusion process, as illustrated in Fig. 1. The final set of eligible papers were downloaded, and a comprehensive full-text review was conducted to confirm their relevance. The documents were systematically organised using Zotero[2], a reference management software designed for managing bibliographic data and research materials.

Scoping orientation: Consistent with the objectives of a scoping review, we did not conduct a formal risk-of-bias appraisal; all records meeting the eligibility criteria were included to provide a comprehensive map of the field. A consolidated qualitative assessment of common strengths and limitations observed across studies is presented in Section 4.7

2.7. Composition of included sources and role in the synthesis

The finalised corpus consists of 82 studies: 75 journal articles, 2 conference papers, 2 review/survey papers, and 1 stakeholder analysis. Conference items were retained where they contributed novel methods or datasets. Review/survey papers and the stakeholder analysis were used to contextualise trends, terminology, and deployment constraints. Unless a given item reported an implemented method or empirical results, it was not counted in quantitative tabulations (e.g., method/fusion distributions, evaluation metrics). This policy preserves methodological consistency while acknowledging valuable contributions beyond the journal literature.

2.8. Data extraction

A detailed data extraction matrix was developed, incorporating elements such as a description of each article, the geographical location of the study, thematic applications, data utilised, fusion methods, algorithms applied, and potential areas for future research. For the data utilised, an additional classification of satellite data usage was included. All relevant information from the journal articles was systematically compiled into a Microsoft Excel spreadsheet, which formed the foundation for further analysis aimed at addressing the research questions identified during the planning stage of the review. To extract keywords, all the downloaded articles PDF files were uploaded to NVivo[3] software to facilitate qualitative text and keyword analysis.

2.9. Annual scientific production

A detailed data extraction matrix was developed, encompassing various elements such as a description of each article, the theoretical framework, the geographical location of the study, thematic applications, data utilised, fusion methods, algorithms applied, and potential areas for future research. Additionally, a classification of satellite data usage was included. Each study was coded by a single reviewer based on these criteria, and the extracted data was systematically incorporated into the collection matrix.

3. Taxonomy and comparative review of data fusion for low-cost sensors

The rapidly expanding literature on data fusion encompasses a wide range of methodologies, architectures, and application domains. However, most prior frameworks and reviews have focused on high-cost, multimodal sensor networks or domain-specific scenarios, often overlooking the particular constraints and challenges associated with LCS deployments. In this section, we present a structured taxonomy of data fusion for LCS, rigorously benchmarking each branch against leading surveys, and critically analysing their field relevance, methodological gaps, and practical strengths. This approach aims to both clarify the landscape and highlight the novel contributions of this review, as shown in Fig. 3.

[2] https://www.zotero.org/

[3] https://lumivero.com/products/nvivo/

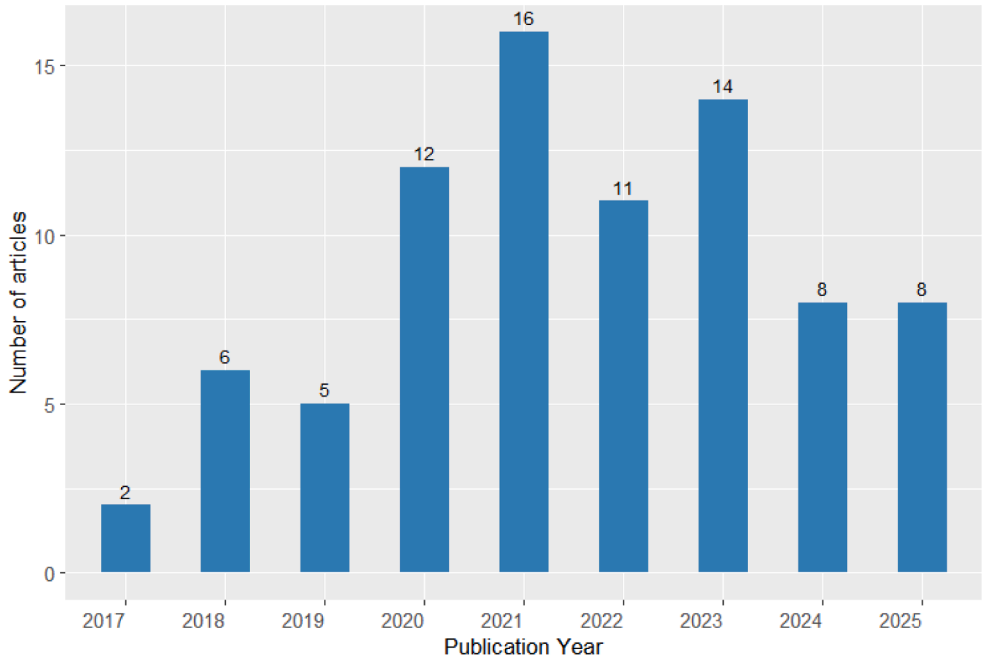

Fig. 2. Annual scientific production.

3.1. Taxonomies in context: benchmarking against the literature

Several influential reviews have shaped the foundational understanding of data fusion. For instance, [14] provide a seminal taxonomy based on fusion architecture (data, feature, decision level), with an emphasis on multi-sensor theoretical developments. Their framework is comprehensive but is largely technology-agnostic, and does not explicitly address issues such as data quality, privacy, or scalability critical to LCS systems. Similarly, [24] focus on knowledge-based data fusion for intelligent transport systems providing important insights into semantic and expert-driven approaches; yet, their survey is anchored in well-resourced, high-quality sensor networks, and less applicable to the decentralised, noisy, and resource-constrained settings common in LCS deployments. A deep dive into spatio-temporal fusion, mainly in the context of remote sensing imagery, providing a sophisticated algorithmic taxonomy but one focused on satellite-based, high-fidelity inputs rather than low-cost in situ sensors is provided in [32]. Baltrušaitis et al. [33] expand the view to multimodal machine learning, summarizing advances in fusing vision, language, and audio, but their survey primarily addresses high-quality, high-bandwidth modalities rarely encountered in low-cost environmental sensor applications. Regarding information quality, [34] reviewed the use of information fusion techniques to improve information quality and identify challenges and research towards improving quality. A more recent taxonomy is provided in [25] where apart from technical exploration, the study also delves into the ethical and privacy implications arising in smart cities, and examines challenges that must be addressed to realise the full potential of fusion within urban settings.

In contrast, our taxonomy explicitly builds on these works while directly addressing the operational realities of LCS: noisy and incomplete data streams, severe hardware constraints, privacy/security risks, calibration needs, and the necessity for scalable, interpretable fusion frameworks. Throughout this section, we highlight where our categories overlap with or diverge from these foundational reviews, and we provide specific examples from recent LCS-focused literature.

3.2. By levels of data fusion

The distinction between data/sensor-level, feature-level, and decision-level fusion remains a useful organising principle, as set out in [14]. However, their generic definitions do not address the trade-offs that arise in LCS settings. For instance, data-level fusion the direct integration of raw sensor measurements, is often attractive for its simplicity and potential to exploit redundancy, but can amplify noise or bias when low-cost sensors lack calibration or quality control. Feature-level fusion can provide robustness by extracting and merging engineered or learned representations but may require sophisticated preprocessing and is sensitive to missing or misaligned data, a frequent issue in distributed LCS networks. Decision-level fusion offers modularity and flexibility, allowing the combination of heterogeneous model outputs (e.g., from different sensor types or ML algorithms), but may sacrifice fine-grained uncertainty quantification or spatio-temporal coherence. While these levels of fusion are recognised in classical frameworks, our taxonomy connects them to the practicalities of LCS deployment, emphasising their respective strengths and limitations for real-world applications.

3.3. By methodological approach

3.3.1. Definitions and scope

To ensure terminological precision throughout this review, we first define the key concepts of Artificial Intelligence (AI), Machine Learning (ML), and Deep Learning (DL) as they are used in our analysis and classification of studies. The field is often characterised by a hierarchical relationship where DL is a subset of ML, which in turn is a subset of the broader AI field [35,36]. Our scope and definitions are inspired by [37], as follows:

- Machine Learning (ML): Non-neural learners (Random Forests, Gradient Boosting, SVMs, k-NN, regularised linear models) and shallow neural regressors/classifiers that do not perform hierarchical feature learning.
- Deep Learning (DL): Multi-layer neural networks with hierarchical feature learning (CNNs, RNNs/LSTM/GRUs, transformers,

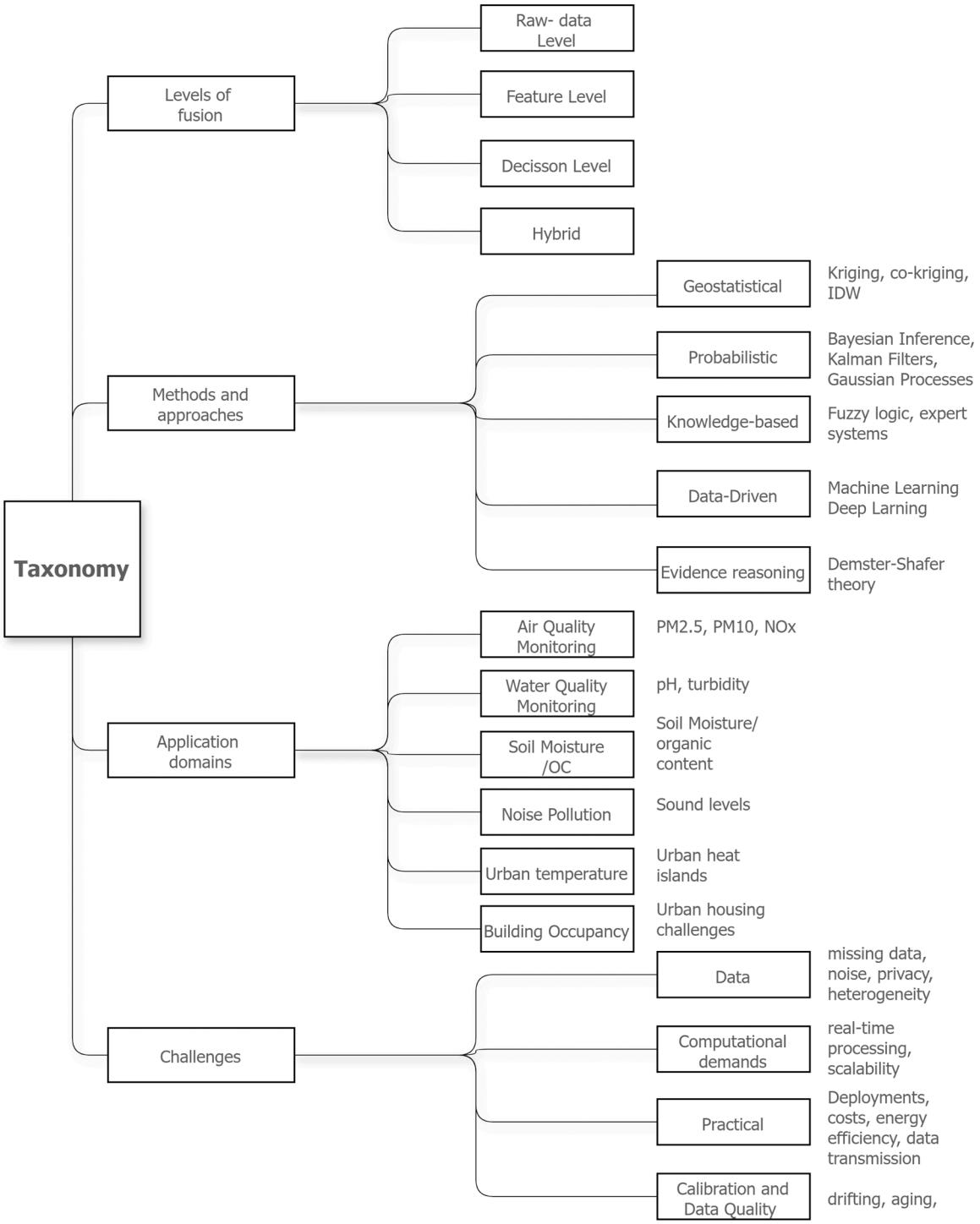

Fig. 3. Taxonomy of data fusion for low-cost sensors applications.

autoencoders), used for fusion, representation learning, or end-to-end mapping

Coding policy: A study is coded ML and/or DL if the corresponding learner appears in the fusion pipeline (feature-, model-, or decision-level). Multi-label coding is permitted (e.g., a study can be both Geostatistical and ML). Hybrid is assigned when ML/DL is explicitly combined with another class (e.g., kriging residuals and RF)

Recent reviews, such as [24] and [33] categorise fusion methods according to knowledge-based, statistical, or machine learning paradigms. Our taxonomy refines this approach for LCS by explicitly considering these four methodological branches, plus an additional hybrid branch, as follows:

Geostatistical methods:

Geostatistical methods (kriging, co-kriging, inverse distance weighting (IDW) explicitly model and exploit spatial autocorrelation, typically represented by a variogram or covariance function, to perform spatial interpolation and uncertainty estimation for point observations. They remain foundational in environmental monitoring because they provide spatially continuous fields and analytically tractable uncertainty

measures. However, their reliance on the stationarity assumption and dense spatial sampling limits scalability to large or highly dynamic sensor networks. For LCS applications, adaptations such as local kriging, moving-window estimators, and sparse variogram models [38] have been proposed to maintain computational feasibility while preserving spatial fidelity. These refinements reflect an evolution from traditional, resource-intensive frameworks toward operationally deployable geostatistical pipelines optimised for real-time or near-real-time LCS fusion.

Probabilistic methods:

Probabilistic approaches (e.g., Bayesian inference, Hidden Markov Models) focus on modelling uncertainty and system dynamics through explicit probability distributions rather than deterministic interpolation. They treat both the state variables and measurement processes as random variables with evolving posterior estimates, enabling dynamic data assimilation and real-time updating. While comprehensive overviews exist in the broader data-assimilation literature [14] and [32], their direct application to low-cost sensor (LCS) networks remains challenging because of computational constraints, communication bandwidth limits, and the difficulty of achieving scalable uncertainty propagation. Recent studies demonstrate a shift toward lightweight and computationally efficient algorithms, such as ensemble Kalman filters, unscented Kalman filters, and sparse Gaussian process (GP) frameworks [39,40], which balance probabilistic rigour with the operational practicality required for field deployments. These advances mark a move away from the more resource-intensive Bayesian implementations typically designed for high-performance computing environments.

Data-driven methods:

[33] comprehensively review ML and deep learning in the context of multimodal fusion; however, their survey emphasizes high-bandwidth data and cloud-based processing. In LCS settings, ML approaches must contend with limited sample sizes, non-stationarity, and sensor drift. We synthesize examples where decision trees, random forests, and compact neural networks are tailored for resource-constrained, distributed processing and for robust feature extraction from noisy inputs.

Knowledge-based methods:

Knowledge-based fusion is provided in [24], but their review is anchored in high quality and expert annotated data. For LCS networks, lightweight rule-based systems and fuzzy logic can supplement or guide data-driven methods, particularly where ground truth is sparse or calibration is problematic.

Hybrid approaches:

Fusion of methodologies is a central theme in modern LCS data analysis. This is reflected in our review's classification system (see Table 4), where the sum of percentages exceeds 100 % because studies were permitted to belong to multiple, non-mutually exclusive categories. This design intentionally captures the prevalence and importance of hybrid approaches, which combine geostatistical, probabilistic, and machine learning techniques to leverage their complementary strengths.

While the potential of such hybridisation is increasingly recognised in the literature[7], few prior reviews offer practical guidance on their integration in LCS systems. This section fills that gap by critically analysing the design, benefits, and operational challenges of specific hybrid fusion frameworks. These include kriging with ML ensembles, fuzzy logic-LUR combinations, and Bayesian-deep learning architectures. We highlight their capacity to improve robustness and accuracy in noisy, dynamic, and data-sparse environments. Particular attention is given to the computational trade-offs and interpretability barriers that arise when deploying these integrated systems at scale.

3.4. By application domain

Most previous surveys (Table 1) focus on either remote sensing or specific verticals such as air quality, smart urban areas etc. Our taxonomy, while drawing from these examples, generalises across the diverse domains in which LCS are increasingly deployed, including air quality, urban microclimate, soil moisture, noise, and traffic monitoring. For each domain, we critically review trends in fusion methodology adoption, discuss field-relevant challenges (such as spatio-temporal gaps, mobility, or privacy risks), and highlight exemplar studies demonstrating the adaptability or limitations of current techniques.

Table 4
Distribution of method classes across the 82 studies (multi-label coding).

Fusion Method	Percentage
Data-Driven (ML/DL)	34 %
Machine Learning (ML)	(24 %)
Deep Learning (DL)	(10 %)
Others	20 %
Geostatistical	15 %
Probabilistic	15 %
Hybrid Approaches	11 %
Knowledge-based	6 %
Total	**100 %**

Note: Multi-label coding is applied; a study may appear in more than one methodological class (e.g., Geostatistical + ML). "Geostatistical" refers to variogram-based spatial interpolation and prediction frameworks (e.g., kriging, co-kriging, IDW) that explicitly model spatial autocorrelation. "Probabilistic" denotes non-variogram probabilistic inference and data-assimilation techniques (e.g., Bayesian inference, Kalman/Ensemble Kalman filters, Hidden Markov Models) that model uncertainty and system dynamics. "ML" includes traditional machine-learning algorithms such as Random Forest (RF), Gradient Boosting Machine (GBM), Support Vector Machine (SVM), and k-Nearest Neighbour (k-NN); "DL" refers to multi-layer neural architectures (CNN, RNN/LSTM/GRU, transformer, autoencoder). "Hybrid" indicates pipelines explicitly combining classes (e.g., kriging + RF, LUR + fuzzy logic). Percentages are computed over N = 82.

3.5. Summary and synthesis

In summary, our taxonomy builds on and extends the foundational work of previous surveys by directly addressing the methodological, operational, and ethical complexities of data fusion in low-cost sensor deployments. Through comparative analysis, critical synthesis, and LCS-centric classification, we provide a practical roadmap for researchers and practitioners aiming to design, benchmark, and advance robust data fusion systems in real-world, resource-constrained settings.

4. Result

4.1. Overview of the studies

Our search and screening produced 82 studies (January 2017-April 2025; Table 5; Fig. 2): 75 journal articles, 2 conference papers, 2 review/survey papers, and 1 stakeholder analysis. Quantitative summaries (e.g., distributions of method classes and fusion levels, evaluation metrics) are computed from studies reporting implemented methods or

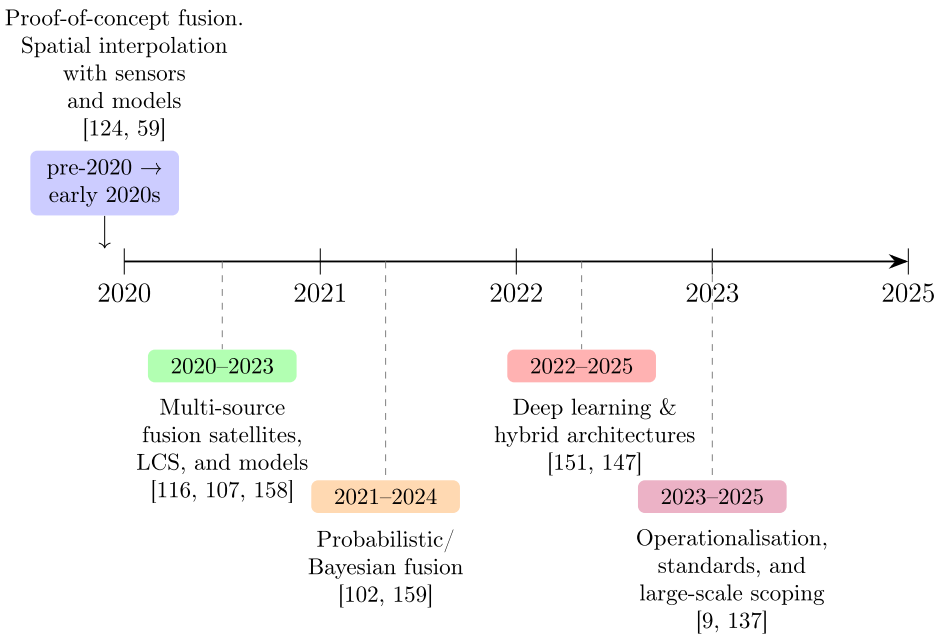

Fig. 4. Selected key phases in evolution of data fusion with low-cost sensors.

empirical results, while non-implementation items inform the qualitative analysis of methodological context and deployment practice.

The temporal distribution of the studies reveals a clear trend towards more recent research, with 79 % of the studies published within the past five years. This surge in publications underscores the growing interest and advancements in LCS data fusion, likely driven by the rapid development of sensor technologies, data analytics, and the increasing emphasis on environmental sustainability. A notable observation is the regional disparity in research output. A substantial proportion of these publications originated from the Global North, which may reflect greater resources and infrastructure dedicated to environmental monitoring and sensor technology development in these regions. This discrepancy highlights potential gaps in the application and accessibility of LCS research in the Global South, where sensor networks and related technologies may not be as widespread. Below, we respond to the key research questions at hand.

4.2. What data fusion methods are being used with LCS?

In order to address this question, we classified the reviewed studies into five broad categories of data fusion methods: geostatistical, data-driven methods, probabilistic, knowledge-based approaches, and lastly hybrid-based methods. Table 4 summarises fusion methodologies while Table 5 provides an overview of representative studies within each category, including details on data sources, fusion techniques, fusion levels and lastly, fusion methods. Based on our results, we expand on these categories in the subsequent subsections, highlighting key methodological principles and illustrating them with examples drawn from the studies summarised in Table 5. We also highlight selected key evolution phases of fusion with low-cost sensors in Fig. 4.

4.2.1. Geostatistical methods

Geostatistical methods form the backbone of spatial data fusion in environmental data, particularly in applications involving LCS and constitutes 15 % of our results, as shown in Table 4. They are also the methodologies used as proof of concept in seminal studies, as indicated in Fig. 4. Their primary strength is in modelling and interpolating spatially autocorrelated phenomena, such as air pollution, soil moisture, and temperature fields, across complex sensor networks, a capability that has led to kriging and its variants being regarded as gold standards for spatial prediction and uncertainty quantification in the geosciences [26,41,42]. The choice of geostatistical model is fundamentally determined by the spatial properties of the data, the number and distribution of observations, and the specific application, whether it is for interpolation, mapping, or exposure assessment. Model selection typically begins with empirical semivariogram analysis to assess how measurement similarity decays with distance, which informs whether ordinary kriging is sufficient (for stationary fields) or if universal kriging is required to account for deterministic spatial trends. In situations where LCS networks are sparse or display uneven coverage, kriging predictions may become unstable or overly smoothed, prompting the adoption of hybrid approaches such as co-kriging with auxiliary covariates from other sensors or the integration of environmental predictors through models e.g. LUR and regression-kriging) to refine estimates in complex urban settings [43].

Computational considerations for geostatistics methods are non-trivial, with classic kriging computational costs scaling cubically with the number of observations. This quickly becomes prohibitive for large scale sensor networks for high resolution data fusion [7]. As such, researchers are increasingly turning to more efficient sparse Gaussian Processes (GP) methods, stochastic approximations or spatial partitioning to maintain tractability without sacrificing predictive performance [7,44].

Within the literature, ordinary kriging [45,46] and universal kriging (UK) [47,48] are most widely used for interpolating pollutant concentrations from LCS networks such as $PM_{2.5}$ and NO_2. UK is especially relevant when spatial trend modelling is needed, for instance by incorporating coordinates or environmental covariates directly into the main function. Co-kriging extends these concepts by allowing secondary data, such as satellite derived information in areas where LCS coverage is limited [49]. Inverse Distance Weighting (IDW), while deterministic, is computationally efficient and is often used as a baseline method for majorly mapping air quality [43,50].

The studies summarised in Table 5 illustrate how geostatistical approaches have been deployed across various environmental applications. Wallek et al. (2022) [45] employed ordinary kriging to fuse dense urban PM_{10} sensor networks with LUR covariates, reporting cross-validated MAE and RMSE to characterise predictive uncertainty. Schneider et al. (2017) [26] implemented universal kriging to integrate NO_2 observations with a dispersion model, evaluating performance through R^2 but without explicit uncertainty quantification. Kibirige et al. (2020)

Table 5
Studies used in this review.

Year	Article	Data Used	Fusion Model	Fusion Level	Methods
2017	Schneider et al. [26]	LCS and EPISODE Model	Geostatistical	Decision level	Kriging
2017	Fiebig et al.[97]	LCS	Data-Driven	Sensor level	MLP, kNN, DT and RF
2018	Schneider et al.[47]	LCS and Model	Geostatistical	Sensor level	Universal Kriging
2018	Jiang et al.[86]	Survey	Survey	Survey	Survey
2018	Castell et al.[59]	LCS, AQMS and EPISODE	Geostatistical	Feature level	Kriging
2018	Bebelaar et al.[138]		Applied Research		
2018	Wang et al.[66]	Environmental: CO_2 Temperature, Relative Humidity - Low cost Wi-Fi probes, Low-cost Camera	Data-Driven	Decision level	ML- ANN, SVM and kNN
2019	Shen et al.[71]	Satellite Imagery and Social Sensing data	Data-Driven	Feature level	DL Deep Belief Network (DBN)
2019	Weissert et al.[62]	LCS and LUR	Data-Driven	Feature level	Linear Regression
2019	Chen and Yang, [85]	N/A	Algorithmic development	N/A	N/A
2019	Zappa et al.[95]	LCS and RS derived data	Data-Driven	Feature level	Random Forest
2019	Huang et al.[139]	LCS, AQMS and ADO	Data-Driven	Decision level	Random Forest
2019	Lai et al.[72]	LCS	Probabilistic	Feature Level	Kalman filter
2020	Gressent et al.[7]	LCS and Dispersal Model	Geostatistical	Sensor level	Kriging
2020	Okafor et al.[64]	LCS	Data-Driven	Feature level	MLR, SLR and ANN
2020	Li et al.[46]	Low-cost, AQMS and satellite imagery	Geostatistical	Feature level	Ordinary kriging
2020	Ferrer-Cid et al.[89]	LCS	Hybrid	Sensor level	Weighted Averages, MLR, Support vector regression, kNN
2020	Feenstra, et al.[84]		Software Package	N/A	
2020	Lin et al.[5]	LCS and AQMS	Hybrid	Sensor level	Optimal Linear Data Fusion and Ordinary Kriging
2020	Kibirige and Dobos [49]	LCS, Terrain, Remote Sensing and Landsat derived NDVI data	Hybrid	Feature level	MLR, Regression Kriging and Ordinary Co-kriging
2020	Xaver et al.[140]	LCS	Applied Research	Feature level	N/A
2020	Zappa et al.[53]	LCS, Satellite derived	Other	Feature level	
2020	Vidaña-Vila et al.[141]	LCS	Data-Driven	Decision level	Deep Learning
2020	Carbajales et al.[142]	LCS, Crowdsourced	Probabilistic	Sensor level	Time series
2020	Fehri et al.[96]	Mobile based crowdsourcing	Geostatistical	Sensor level	Best Linear Unbiased Predictor (BLUP)
2021	Mani, Volety [73]	LCS	Probabilistic	Sensor level	Kalman filter
2021	Kelly et al.[75]	LCA	Probabilistic	Feature level	Gaussian process (GP) model
2021	Shafran-Nathan et al.[61]	LCS and LUR derived data	Knowledge-Based	Decision level	Fuzzy Logic
2021	Novak et al.[143]	LCS	Other	Sensor level	Data harmonisation
2021	İçöz, et al.[82]	Platform Development	Platform Development	N/A	N/A
2021	Becerra et al.[144]	AQMS	Knowledge-Based	Ensemble	Information Quality, JDL
2021	Chao et al.[50]	LCS, Fixed stations, ADO	Geostatistical	Feature level	Inverse Distance Weighting (IDW)
2021	Idir et al.[145]	Mobile and crowd-sourced	Geostatistical	Sensor level	Simple Kriging (SK), Ordinary Kriging (OK), and Kriging with External Drift, IDW
2021	Zumwald et al.[98]	LCS, Reference and personal weather stations	Data-driven	Decision level	ML - Quantile Regression Forest
2021	Okafor et al.[65]	LCS	Data-Driven	Decision level	Multiple Linear Regression, Decision Tree, Random Forest and XGBoost
2021	Kaginalkar et al.[87]	Review	Review	Review	N/A
2021	Huang et al.[92]	LCS and CMAQ Model	Geostatistical	Ensemble	Ordinary Kriging
2021	Babaeian et al.[91]	Drone Images, Soil Samples	Data-Driven	Decision level	AutoML
2021	Kibirige et al.[146]	Flower Parrot Sensor at a depth of 10cm , Landsat and SAR	Hybrid	Decision level	Multiple Linear Regression (MLR), Cokriging (CK) and Regression Kriging (RK)
2021	Veiga et al.[147]	LCS	Data-Driven	Decision level	Random Forest
2021	Anachkova et al.[81]	LCS	Platform	N/A	
2022	Pu and Yoo [51]	AQMS, ADO, Auxiliary Data	Hybrid	Ensemble	XGBoost algorithm
2022	Nguyen et al.[52]	Satellite Imagery/Surface Samples	Knowledge-Based	Decision level	ML XGBoost
2022	Johansson et al.[60]	ENFUSER and Dispersal Models, Emissions Data	Knowledge-Based	Decision level	Data Assimilation
2022	Chen et al.[148]	LCS and AQMS	Geostatistics	Sensor level	Clustering-based Inverse Distanced Weighting (CIDW)
2022	Han et al.[43]	TROPOMI, AQMS and SRTM DEM	Hybrid	Feature level	IDW, OK, RF, and Random Forest combined with OK (RFK)
2022	Wallek et al.[45]	AQMS and LUR	Geostatistical	Decision level	Ordinary kriging
2022	Bobbia et al.[42]	LCS (micro-sensor)	Geostatistical	Sensor level	Kriging
2022	Briciu-Burghina et al.[149]	LCS	Knowledge-Based	Sensor level	N/A
2022	Corbari et al.[150]	LCS and Satellite	Probabilistic	Feature level	N/A
2022	Bush et al.[63]	LCS, AQMS	Data-Driven	Feature level	Random Forest

Table 5
Continued.

Year	Article	Data Used	Fusion Model	Fusion Level	Methods
2022	Tan et al [93]	LCS and Commodity Grade	Hybrid	Ensemble	CNN/LSTM feature extraction + ensemble fusion
2023	Fu et al.[67]	AQM, LCS, and TROPOMI	Data-Driven	Feature level	Fusion-Imputation-Gradient Boosting-Machine
2023	Criado et al.[48]	CALIOPE (model) + LUR (statistical) + monitors (observation)	Hybrid	Decision level	Universal Kriging
2023	Al Yammahi et al [68]	AQMS	Data-Driven	Feature level	ML LSTM
2023	Liang et al.[102]	LCS, Satellite Imagery	Data-Driven	Decision level	Deep Learning CNN
2023	Miasayedava et al [54]	AQMS and SILAM model	Probabilistic	Decision level	Sequential Least-Squares Data Assimilation
2023	Miasayedava et al.[151]	Platform	Other	Sensor level	Least square data assimilation
2023	Kaginalkar et al.[88]	Stakeholder Analysis	N/A	N/A	N/A
2023	Wu et al.[152]	Custom IoT sensor network	Data-Driven	Feature level	ML Deep Q Network
2023	Pradeep et al.[83]		Multi-Sensor Platform	Hardware/Platform-Level Integration (physical co-location only)	
2023	Tsanousa et al.[69]	LCS (Multi-sensor environmental data	Data-Driven	Ensemble	ML LGBM and ExtraTrees
2023	Zhu et al.[153]		Review Article	N/A	N/A
2023	Aix et al.[111]	LCS	Data-Driven	Sensor level	ML RFR, MLR
2023	Fritz et al.[154]	BEVO Beacon (IAQ), Smartphone GPS, Fitbit sleep data	Data-Driven	Ensemble	Multi-Layer Perceptron (MLP)
2024	Zhang et al. [55]	LCS and AQMS	Probabilistic	Ensemble	Ensemble Kalman Filter (EnKF)
2024	Okafor et al.[155]	IoT sensor data in peatlands	Probabilistic	Ensemble	Multiple Hypothesis Tracking (MHT) method
2024	Rodriguez et al.[156]	LCS	Hybrid	Sensor level	ANN and IDW interpolation
2024	Wei et al.[90]	Google traffic, mobile sensor data	Data-Driven	Feature level	Deep neural network (DNN) model and a modified CNN model, ResNet
2024	De vito [157]	LCS	Case Study		
2024	Song et al.[158]	LCS, Vis-NIR and pXRF	Data-Driven	Ensemble	ML Deterministic PLS-based multi-block regression (SO-PLS) t
2024	Tang et al. [159]	Observations, CTMs, AOD	Data-Driven	Sensor level	Obs-model ML Random Forest
2024	Das et al.[160]	IoT sensor data	Probabilistic	Infrastructure	Kalman filter
2025	Gamazo-Real et al.[161]	IoT sensor data	Data-Driven	Sensor level	Centralised vs distributed edge ML
2025	Tasnim et al.[77]	Mobile IoT air quality data	Probabilistic	Sensor level	Reputation mechanisms and context-aware fusion - Contextual Hidden Markov Model
2025	Hoogerbrugge et al.[162]	AQMS + low-cost sensor data	Others	Sensor level	Statistical framework integration of different measurement sources
2025	Ho wo Chen et al.[163]	Monitoring station data	Probabilistic	Sensor level	Source apportionment using Expected value theory
2025	Choi et al.[164]	LCS, Multimodal AQMS	Probabilistic	Sensor level	Stochastic advection-diffusion model
2025	Ganji et al [165]	Traffic and AQMS data	Data-Driven	Sensor level	Deep learning (CNN + LSTM)
2025	Shetty et al. [166]	AQMS, CAMS	Data-Driven	Sensor level	ML-based spatial downscaling
2025	Tang et al.[58]	LCS, AQMS	Data-Driven	Sensor level	Data augmentation via chained imputation (CI-DA) and harmonisation via interpretable semi-supervised ML
2025	Wang et al.[57]	Observational and simulation data	Data-Driven	Ensemble	DL CNN with dual attention mechanism

[49] combined sparse in-situ soil measurements with satellite data using co-kriging, highlighting trade-off between heterogeneity and computational cost. Chao et al. (2021) [50] and Han et al. (2022) [43] adopted inverse-distance weighting (IDW) as a computationally efficient baseline for air-quality mapping, while.Zhang et al 2021 [43] introduced a clustering-based IDW variant evaluated via leave-one-out cross-validation (LOOCV). Together, these examples demonstrate that Table 5 captures not only the diversity of geostatistical fusion methods but also their characteristic data structures, validation metrics, and varying degrees of uncertainty treatment.

Increasingly, model-based geostatistical techniques are combining LCS data with chemical transport or dispersion models including EPISODE, [26,59], ENFUSER, [60], and ESCAPE [61]. Other include Land Use Regression (LUR) model [62] and ADMS-Urban model [7].

Performance evaluation of geostatistical fusion models are grounded on robust statistical metrics including Root Mean Square Error(RMSE) and Mean Absolute Error (MAE) to quantify predictions, R^2 and explained variance to assess spatial variability captured, and prediction interval coverage, especially crucial in kriging due to its probabilistic uncertainty estimation. Cross-validation strategies, such as leave-one-out or block-based spatial validation, are widely used to ensure generalizability. Notably, comparative studies such as [43] and [50] have demonstrated that hybrid or co-kriging models consistently outperform

basic spatial interpolation, particularly when auxiliary data from satellite sensors or meteorological fields are leveraged.

Nevertheless, several limitations persist. The scalability challenge of kriging becomes acute as data volumes grow, a common scenario with contemporary high-resolution sensor deployments. This necessitates approximation strategies like sparse Gaussian processes, spatial partitioning, or high-performance computing [7,44,63]. Furthermore, geostatistical models are sensitive to assumptions regarding stationarity, isotropy, and the functional form of the variogram. Therefore, incorrect assumptions can result in bias and unreliable uncertainty quantification. The issue of sensor bias is also significant, as uncorrected systematic errors in LCS measurements can propagate through the fusion process, highlighting the need for rigorous pre-processing and calibration. Integrating heterogeneous data sources, such as harmonizing LCS with satellite-derived or regulatory-grade data, requires careful attention to temporal and spatial alignment, unit consistency, and differing noise profiles. Edge effects, where interpolation accuracy degrades near the boundaries of the domain with few neighbouring observations, are also a practical consideration. Despite these challenges, geostatistics remains an indispensable component of environmental data fusion, especially when integrated with satellite derived and physical modelling to provide spatially explicit and robust environmental intelligence.

4.2.2. Data-driven methods

Data-driven methods, particularly general machine learning (ML) and deep learning (DL), are rapidly transforming the paradigm of data fusion in environmental monitoring and remote sensing by enabling the extraction of meaningful insights from location-aware, geo-referenced, and heterogeneous datasets. This is reflected in our results in Table 4 with these methods taking the largest share at 34 %. Unlike geostatistical approaches, which fundamentally rely on spatial autocorrelation, data-driven methods offer the flexibility to model highly nonlinear, multi-source, and multi-scale relationships that are common in urban, atmospheric, or land surface environments. The selection of an appropriate ML model is inherently linked to the problem structure, whether it involves regression tasks such as pollutant concentration estimation, classification tasks such as occupancy detection or event detection/recognition, or time series forecasting where recurrent architectures like Long short-term memory (LSTM) are particularly effective. AI models are well-suited to address the complexities of missing data, high-dimensional feature spaces, and the integration of diverse sources such as satellite spectral bands, meteorological data, and LCS network outputs.

A key factor with AI models is interpretability, which remains an important criterion. Models like decision trees and random forests provide transparent feature importance scores, facilitating domain understanding and regulatory acceptance, whereas neural networks, particularly deep architectures, function as black boxes, leading to increased reliance on post hoc interpretability techniques such as SHAP or LIME. Scalability considerations also play a key role: deep learning architectures, including convolutional neural networks (CNNs) and LSTMs, are highly scalable and can exploit large training sets common in environmental sensing, but they also demand substantial computational resources and careful regularization to avoid overfitting.

Within the reviewed literature, traditional machine learning models appear to be the most dominant models in data fusion (Table 4) with random forest algorithms have emerged as a robust choice for both regression and classification tasks in air quality mapping, noise estimation, and soil moisture prediction, offering resilience to noisy data and the additional advantage of feature ranking [64,65]. Support Vector Machines (SVMs) are particularly effective for high-dimensional and smaller datasets, with their performance heavily influenced by the selection of kernel functions [66]. Artificial Neural Networks (ANNs) are widely deployed in sensor calibration, air quality forecasting, and soil moisture estimation. ANNs however require considerable expertise in hyper-parameter tuning, data normalization, and sometimes data augmentation to achieve optimal results [67]. In the realm of deep learning, CNNs have demonstrated efficacy in extracting spatial patterns from gridded satellite and sensor data, while LSTMs excel in modelling sequential dependencies and temporal evolution of environmental variables, such as air pollution time series or event detection [68]. Ensemble and hybrid models, such as stacked regressors or late-fusion frameworks, have also shown strong performance by integrating multiple ML algorithms or combining ML with geostatistical predictions, thereby achieving a superior balance between flexibility and spatial coherence [69].

Performance evaluation is rigorous, employing RMSE and MAE for continuous predictions, accuracy, precision, recall, and F1-score for classification or detection tasks, and cross-validation schemes such as k-fold or spatial-temporal splits, to ensure model robustness and generalizability. Many studies directly benchmark ML and DL methods against geostatistical baselines such as kriging, with results often showing that ML approaches provide substantial accuracy improvements when data richness permits.

Application of data-driven models in environmental data fusion is however not without challenges. DL methods, for instance, require extensive labelled datasets, which is often a limiting factor in environmental science, and are vulnerable to concept drift, where changing real-world conditions degrade model performance over time. The opacity of deep architectures can limit stakeholder trust, particularly in regulatory or mission-critical contexts, underscoring the importance of explainability. Overfitting is a persistent risk, especially in small or biased LCS networks, and transferability remains a major hurdle. Models trained in one geographic or temporal context may not generalize elsewhere without substantial re-training or domain adaptation. Finally, the computational burden of training and deploying deep models can be a significant drawback for real-time or edge applications, necessitating careful workflow optimization and sometimes the use of lightweight or approximate models.

Emerging best practices in the literature involve hybridizing AI with geostatistics or domain knowledge, implementing rigorous calibration and uncertainty quantification, such as via Bayesian neural networks or ensemble methods, and employing explainable AI techniques to support feature attribution and build stakeholder confidence. In sum, AI and ML are driving a revolution in LCS data fusion, enabling richer, more dynamic, and more accurate environmental intelligence, provided that methodological rigour and domain awareness are maintained throughout the modelling process.

As shown in Table 5, data-driven approaches span traditional ML algorithms and DL architectures, reflecting their growing influence in LCS fusion. Okafor et al. [64] applied Multiple Linear Regression (MLR) and Random Forest (RF) to fuse LCS and environmental covariates for air-quality mapping, using RMSE and MAE as performance metrics and feature-importance scores for interpretability. Wang et al. [70] integrated Artificial Neural Networks (ANN), SVMs, and k-NN for occupancy detection from environmental sensor networks, showing sensitivity to kernel selection but omitting uncertainty quantification . Fu et al. [67] used a Gradient Boosting Machine (GBM)-based fusion-imputation model to estimate hourly NO_2 levels, reporting MAE as the primary validation metric while noting hyperparameter-tuning challenges. For deep-learning applications Shen et al. [71] fused satellite imagery and social-media data to produce high-resolution surface-pollution maps, showcasing the ability of DL models to capture non-linear spatial dependencies. Collectively, these studies confirm that ML/DL-driven fusion provides flexibility for high-dimensional data while still facing limitations in interpretability and uncertainty handling.

4.2.3. Probabilistic methods

Probabilistic methods estimate parameter values at unsampled locations, a typical case when integrating LCS and official in -situ measurements. These methods, which formed 15 % of our results, (Table 4, take into account the spatial structure of the data and the uncertainty in predictions, relying on probability density functions [14]. Probabilistic data

fusion methods are particularly valued in remote sensing and environmental monitoring for their principled approach to uncertainty quantification and capacity to assimilate information from disparate sensor sources. In the context of LCS, these methods play a crucial role in both spatial and temporal integration, providing rigorous frameworks for inferring latent environmental fields, such as air quality or soil moisture, from incomplete, noisy, or heterogeneous data. Central to this family of approaches are Bayesian methods, including Bayesian filtering (e.g., Kalman Filter and its variants), Gaussian Processes (GP), and ensemble-based data assimilation strategies. The selection of a specific probabilistic approach is governed by factors such as the statistical properties of the observed data, the dimensionality of the state space, and computational feasibility. For instance, the Kalman Filter and its extensions, such as the Extended Kalman Filter (EKF) and Unscented Kalman Filter (UKF), are widely used for state estimation in linear or mildly non-linear systems where sequential data assimilation is required, such as updating pollution estimates with streams of sensor data [72,73]. These filters are especially advantageous when high-frequency, temporally resolved LCS data must be fused with sparser regulatory or remote sensing observations, as they systematically propagate uncertainty and enable real-time updating.

GPs represent a set of powerful class of probabilistic models, offering non-parametric, flexible approaches to spatial and spatio-temporal prediction [74]. GPs are particularly well-suited to environmental applications because they provide both predictive means and full posterior covariance estimates, capturing the spatial correlation structure inherent in phenomena like air pollution [75]. However, their practical deployment with LCS networks and large-scale remote sensing data is often limited by their cubic computational complexity with respect to the number of data points, a challenge that has spurred the adoption of sparse approximations, inducing-point methods, and variational inference frameworks to maintain scalability without sacrificing accuracy [76]. Least Squares Data Assimilation methods, while perhaps less commonly discussed, also find application in the fusion of LCS and satellite data, supporting robust estimation even in the presence of missing or biased measurements [5,54].

Performance assessment in probabilistic data fusion typically extends beyond classical error metrics such as RMSE, F1 and MAE [77] to include probabilistic scores, such as the Continuous Ranked Probability Score (CRPS), likelihood-based measures, and coverage probabilities of prediction intervals, reflecting the importance of both accuracy and calibrated uncertainty estimation in environmental risk assessment and decision support.

Despite their effectiveness, a key limitation of probabilistic approaches lie in their computational demands,especially as the dimensionality and volume of data increases, which grows cubically with the number of data points, [78]. This makes them impractical for large-scale applications without approximation techniques. Research to improve the use of GP include Sparse Gaussian Process Regression (SGPR) [44], which addresses the computational limitations by approximating the full GP with a smaller, more manageable set of locations, making it suitable for large datasets. Other areas of research include distributed computing, parallelization, and stochastic variational inference for large-scale applications [44,79]. In addition, successful application of probabilistic fusion requires careful attention to model specification, prior selection, and data harmonization, as unmodeled bias or misspecification can propagate through the inference process, compromising both predictions and uncertainty quantification. Nevertheless, probabilistic methods offer a compelling, theoretically rigorous foundation for environmental data fusion, enabling robust integration of LCS, satellite, and reference-grade observations with explicit management of uncertainty, an essential property for high-stakes environmental monitoring and regulatory applications.

Table 5 provides multiple examples of probabilistic frameworks emphasising uncertainty quantification and system dynamics Lai et al. [72] implemented a Kalman filter framework to fuse IoT-based LCS with regulatory monitors, demonstrating real-time state estimation and uncertainty propagation. Mani et al. [73] compared LSTM and ARIMA forecasting with Kalman filtering for air-pollution prediction, reporting higher accuracy than stand-alone models while acknowledging computational costs. Kelly et al. [75] applied a Gaussian process model for community-scale air-pollution mapping, using posterior covariance to derive uncertainty intervals validated via CRPS. Lin et al. [80] and Miasayedava et al. [54] employed least-squares data assimilation to combine numerical-model outputs with LCS observations, enhancing bias correction and robustness to missing data. Collectively, these representative studies show that probabilistic fusion supports explicit uncertainty quantification and dynamic state estimation, extending beyond purely spatial interpolation.

4.2.4. Knowledge-based methods

Knowledge-based data fusion methods offer a complementary paradigm to purely statistical or machine learning approaches by explicitly incorporating expert knowledge, rules, and contextual reasoning into the data integration process. They have therefore garnered increasing attention in the research community [24]. From our results, these models were the least applied at 6 %. See Table 4. In the context of LCS and remote sensing, these methods are particularly valuable in applications where physical models are incomplete, ground-truth data are sparse, or human expertise can provide essential constraints on possible system states. The core methodological frameworks within this class include fuzzy logic, rule-based systems, and hybrid approaches that combine expert-defined rules with data-driven inference.

Fuzzy-logic- and rule-based fusion frameworks enable the modelling of imprecise or linguistically defined relationships among environmental variables, providing transparency and explainability under data uncertainty [17]. As highlighted in Table 5, Lau et al. [17] used fuzzy inference to integrate traffic-based LUR models with LCS observations, generating high-resolution air quality maps in data-sparse urban environments [61]. Similarly, rule-based systems, often grounded in domain expertise, allow the definition of logical conditions and hierarchical relationships, which can govern sensor quality control, anomaly detection, or fusion of multi-source evidence in environmental monitoring networks. These systems are particularly valuable for embedding operational constraints, such as regulatory thresholds, physical plausibility checks, or decision heuristics that may not be easily learned from data alone. Johansson et al [60], utilised a combination of dispersion modelling approaches and data assimilation to extract and combine information from open-access sources. As highlighted in Table 5, knowledge-based methods are most commonly aligned with decision-level fusion, where outputs from multiple models or sensors are combined through logical rules or hierarchical reasoning. This reflects their strength in operational monitoring contexts where explainability, regulatory compliance, and stakeholder trust are central.

While knowledge-based methods can dramatically improve robustness and context-awareness, especially in real-world deployments where data are incomplete or subject to domain-specific artefacts, a key limitation is their dependence on the availability and quality of expert knowledge. If rules are too simplistic, static, or misaligned with evolving system dynamics, they may introduce bias or limit model adaptability. Furthermore, encoding and maintaining complex rule bases can become labour-intensive as monitoring networks scale or as system requirements change. Despite these challenges, the integration of knowledge-based methods with data-driven modelling is increasingly recognised as best practice in environmental informatics, supporting explainability, regulatory compliance, and stakeholder trust. As sensor networks become more ubiquitous and the need for context-aware, actionable intelligence grows, the role of knowledge-based data fusion, in conjunction with statistical and AI-driven methods is likely to expand, particularly in areas such as smart cities, environmental health, and climate adaptation planning.

4.2.5. Hybrid fusion approaches

Hybrid fusion approaches, defined here as the integration of two or more methodological paradigms such as geostatistical, probabilistic, knowledge-based, or machine learning frameworks, represent a promising and increasingly influential direction in low-cost sensor (LCS) data fusion. These methods are particularly effective in complex, data-scarce, or safety-critical environments, where neither data-driven nor knowledge-driven techniques alone provide sufficient robustness or generalisation. Across the corpus, 9 of the 82 studies (11 %) employed hybrid strategies (Table 5), confirming that hybridisation is emerging from conceptual exploration toward practical implementation. These studies typically exploit expert knowledge or physical models to constrain or regularise data-driven learning, thereby improving interpretability, uncertainty quantification, and transferability across spatial and temporal domains. Representative examples include Kibirige et al. [49], who integrated Multiple Linear Regression (MLR), Co-kriging, and Regression Kriging (RK) for soil-moisture estimation, demonstrating that statistical-ML integration enhances spatial fidelity in sparse networks; Lau et al. [17], who employed fuzzy inference to handle linguistic uncertainty and imprecise thresholds in noisy sensor data, using expert rule sets as a pre-fusion calibration layer; and Pu and Yoo [51], who developed an ensemble-level, gap-filling framework that fused multi-source AOD data with deep-learning models to produce high-resolution $PM_{2.5}$ estimates, achieving improved spatial resolution and robustness. Other relevant examples include Ounoughi et al. [24], who embedded expert rules alongside machine-learning algorithms to improve interpretability while maintaining predictive strength, and the CALIOPE-Urban v1.0 framework [48], which exemplifies a hybrid probabilistic-geostatistical approach that combines deterministic dispersion-model outputs, empirical LUR covariates, and in-situ observations through universal kriging to achieve uncertainty-aware NO_2 mapping at street-scale resolution. Collectively, these studies demonstrate that hybrid approaches are not merely theoretical constructs but constitute an evolving methodological paradigm that enables multi-scale inference, context-aware calibration, and explicit uncertainty propagation under heterogeneous data regimes. The growing prevalence of hybrid models thus reflects a broader shift toward model-informed learning, bridging physical interpretability and predictive capability, and aligning with the recent trajectory of physics-informed machine learning and hybrid probabilistic modelling in environmental informatics.

Accordingly, Table 4 has been updated to include “Hybrid (11 %)” as a distinct methodological class, consistent with the empirical evidence presented above, while the remaining studies that did not employ specific fusion models are grouped under “Other” (e.g., platform development [81–83], software applications [84], algorithmic design [85], review papers [86,87], and stakeholder analyses [88].

4.3. Data fusion levels

Complementing the discussion on fusion methods, we now consider the levels at which data fusion is performed. Following [15], three canonical levels can be distinguished namely signal-level, feature-level, and decision-level. In addition, our review revealed a fourth category - multi-level fusion, reflecting emerging practice where different stages are combined. Fig. 5 illustrates the conceptual framework, while Table 5 situates the reviewed studies across both methodological approaches and fusion levels, showing how particular techniques cluster at each level. These proportions provide a high-level answer to RQ2. Below, we unpack each of these levels with representative examples.

4.3.1. Level 1: sensor level fusion

Sensor-level fusion combines raw data from multiple sources to enhance accuracy and reliability. Our results in Table 5 shows that 37 % of the reviewed articles adopted this level. A representative example include foundation paper by [26], which applied geostatistical data fusion, between LCS and model data to improve spatial resolution. Similarly, [7] merged raw sensor data with results from a dispersal model, demonstrating the added value of fusing unprocessed signals. Ferrer-Cid et al. [89] also used this approach for sensor calibration, directly improving measurement quality.

4.3.2. Level 2: feature-Level fusion

Feature-level fusion extracts and combines descriptors derived from raw measurements. This level accounted for 26 % of studies (Table 5), with applications mainly in air quality and soil moisture monitoring. For instance, [49] fused features extracted from Sentinel-1B C-band Synthetic Aperture Radar (SAR), Landsat 8 data, and citizen observatory data to estimate surface soil moisture distribution. Okafor et al. [64] combined datasets from low-cost sensors with environmental covariates for calibration, while [75] highlighted that fused features revealed intra-urban $PM_{2.5}$ variations not captured by regulatory measurements alone. In a more recent study, [90] extracted features from Google traffic and mobile sensing data in a deep learning platform, out-performing existing ML models as validated by resulting R^2.

4.3.3. Level 3: decision-level fusion

Decision-level fusion represents the highest stage, where independent outputs are aggregated into a final classification or estimation [15]. This was the third most common approach,covering 24 %, as shown in Fig. 6 as well as Table 5. In many cases, this level was implemented via machine learning or rule-based systems. [52,60] demonstrated operational deployments of LCS-based air quality monitoring using decision-level aggregation. Babaeian et al. [91] investigated the feasibility of AutoML in estimating field-scale root zone soil moisture combining VIS and NIR reflectance measurements. Last but not least, [92] combined outputs from LCS, AQM and CMAQ to generate hourly spatio-temporal $PM_{2.5}$ exposures.

4.3.4. Level 4: multi-level fusion

A smaller but notable set of studies (15 %) employed multi-level/ensemble fusion levels, where sensors, feature and decision level approaches were integrated. This reflects increasing complexity of data fusion and the need for hybrid strategies. Pu et al(2022) [51] applied multi-modal learning across fusion levels, [93] used hierarchical fusion for environmental monitoring and [81] designed a framework that explicitly spanned multiple levels. Whilst less common, this category signals a trend towards more versatile and adaptive fusion architectures, frequently employing artificial intelligence techniques for fusion.

Taken together, the distribution across levels highlights clear methodological preferences. As Fig. 6 shows sensor-level fusion is the most dominant followed closely by feature and decision levels. Knowledge-based methods, though less frequent, appear to be applied at the decision level, while multi-level designs often integrate AI-driven pipelines spanning raw data to decision outputs is rarely used in geostatistical models. This pattern underscores both the methodological diversity of LCS data fusion and the emerging trend towards hybrid, multi-level frameworks.

The integration of Table 5 and Fig. 6 highlights how different methodologies align with specific fusion levels. Data-driven methods dominate sensor-level fusion, reflecting their strength in processing raw or pre-processed time-series data, whereas feature- and decision-level fusion increasingly rely on probabilistic and ML approaches for uncertainty quantification and decision support. Knowledge-based methods appear predominantly at the decision level, emphasising rule-based reasoning and interpretability, while multi-level designs often integrate hybrid or ML/DL pipelines spanning from raw data to decision outputs. Taken together, these patterns underscore both the methodological diversity of LCS data fusion and the emerging convergence toward hybrid, multi-level frameworks that couple physical interpretability with computational adaptability.

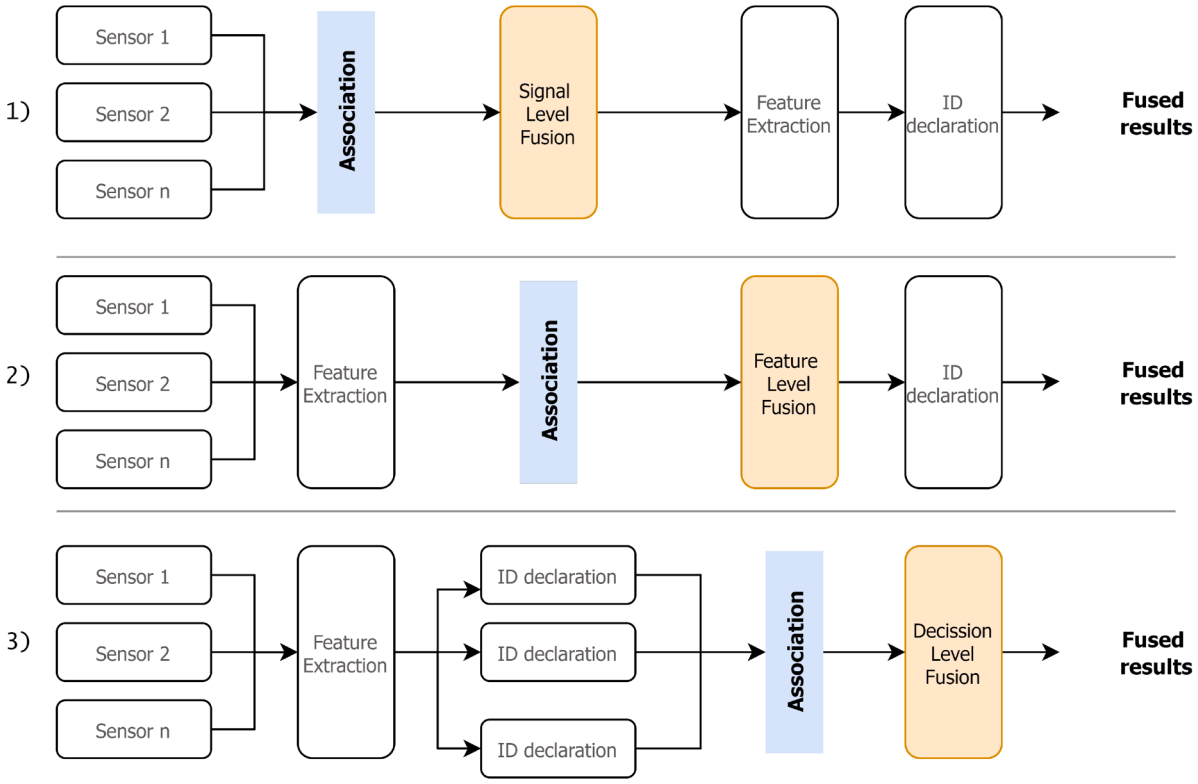

Fig. 5. Fusion levels flow.

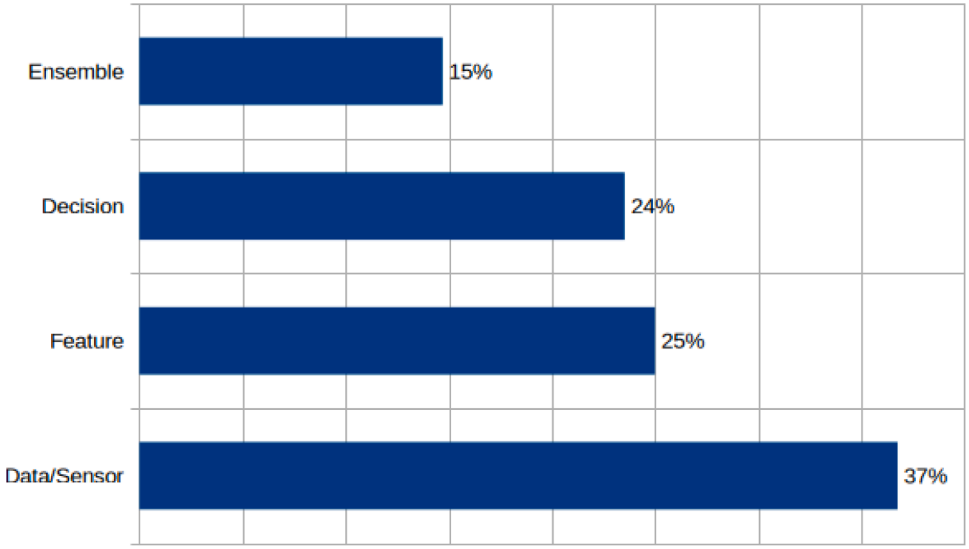

Fig. 6. Percentages of fusion levels by articles.

4.4. What are the application contexts?

Among the 82 studies included in this review, a clear majority, approximately 70 %, focused on air-quality monitoring, underscoring the dominant role of low-cost sensors (LCS) in urban environmental assessment and citizen-driven pollution tracking. This prominence reflects both the rapid maturity of low-cost air quality sensing technologies and the extensive availability of calibration reference networks in cities. Approximately 14 % of studies examined soil-moisture monitoring, highlighting the expanding use of LCS in agricultural and hydrological applications, particularly for irrigation management and drought assessment. Noise and building occupancy monitoring each accounted for about 5 %, demonstrating the growing integration of sensor networks within smart-city and indoor-environment frameworks. The remaining 8 % addressed a diverse range of emerging applications, including water quality assessment, urban-heat-island analysis, and noise pollution mapping, which together illustrate the versatility of data fusion techniques across environmental domains. While these proportions confirm the strong traction of LCS in air quality research, they also reveal significant opportunities for methodological transfer to under-represented fields such as hydrology, water quality, and climate adaptation.

The second question in this review was: What are the key application contexts of DF with LCS? To address this, we analysed the technologies underpinning data fusion applications across various domains.

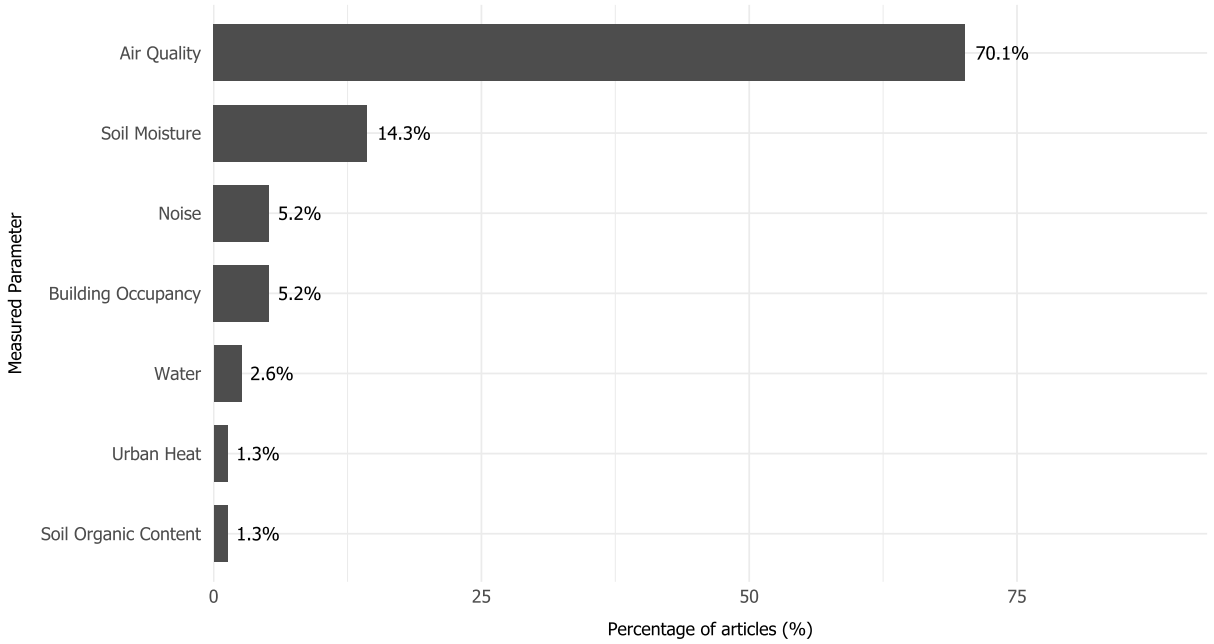

Fig. 7. Top parameters measured.

Our findings highlight that DF technologies are predominantly applied to air quality, soil moisture, water quality, building occupancy, urban heat, and noise monitoring, as shown in Fig. 7. Below, we discuss these applications from a technological perspective. This section details the parameters.

4.4.1. Air quality

Air quality emerges as the most researched application of DF with LCS, supported by advanced fusion technologies. Early studies, such as those by [26], implemented signal-level fusion by processing raw sensor data to enhance measurement accuracy and reliability. Signal-level fusion involves algorithms tailored for real-time data calibration and noise reduction, as illustrated in Fig. 5. This approach sets the foundation for integrating raw LCS data into actionable datasets.

Recent advancements emphasise multi-source fusion, combining LCS data with dispersion models and mobile or stationary platforms to improve spatial coverage and resolution. For instance, [7] demonstrated how technological integration can refine air quality measurements through dynamic calibration models and hybrid fusion algorithms. These approaches leverage LCS flexibility to augment traditional air quality monitoring infrastructure.

A growing trend in DF involves the integration of LCS data with satellite imagery, enabling high-resolution spatio-temporal environmental monitoring. Such applications often employ geostatistical techniques and machine learning models to address challenges in data heterogeneity and spatial sparsity. For instance, [94] used IoT-enabled LCS networks combined with satellite data to improve the resolution of NO_2 measurements. This study highlights the role of edge computing and IoT platforms in preprocessing and harmonizing heterogeneous datasets before fusion.

Similarly, [43] developed an innovative framework that integrates satellite imagery with regulatory-grade station data for $PM_{2.5}$ estimation. Their use of interpolation algorithms illustrates how data fusion technologies can bridge gaps in monitoring networks. Moreover, [67] combined data from official-grade sensors, LCS, and satellite retrievals using machine learning models to identify NO_2 hotspots in urban environments. These technologies underscore the potential of hybrid fusion methodologies, where machine learning algorithms process complex datasets to extract actionable insights.

4.4.2. Soil moisture

In the case of soil moisture, a notable technological approach combines LCS data with high-resolution satellite data, such as Sentinel-1 and Sentinel-2 imagery. For example, [95] and [53] evaluated the use of crowdsourced observations and satellite data to assess the temporal and spatial consistency of satellite-derived soil moisture products. Their work demonstrates the potential of citizen observatories (CO) in augmenting satellite data for more reliable and comprehensive soil moisture estimations.

The technological backbone for these fusion techniques involves geostatistical methods like regression-kriging, which integrates sensor measurements with satellite data for improved spatial interpolation. The fusion process uses microwave satellite imagery, such as from Sentinel-1B's Synthetic Aperture Radar (SAR), to capture the soil moisture at different depths. This approach leverages the technological synergy between LCS, satellite systems, and geospatial analytics to deliver higher-resolution and more consistent soil moisture maps.

Following these foundational studies, [49] explored advanced geostatistical techniques for estimating soil moisture, including multiple regression analysis, regression-kriging, and co-kriging with a digital elevation model. These methods integrate satellite imagery with citizen-generated data, such as CO data, to refine soil moisture estimates. The fusion of Sentinel-1B SAR, Landsat 8 data, and terrain information is enhanced by the incorporation of machine learning algorithms, such as Random Forest (RF) and Support Vector Machines (SVM), which can adaptively learn patterns from the data to improve the prediction accuracy.

The fusion methodologies implemented in these studies highlight the importance of integrating various data sources through advanced data processing pipelines. Machine learning models are increasingly used to fine-tune the fusion process by handling large volumes of heterogeneous data. These techniques facilitate better handling of complex spatial and temporal variations in soil moisture and allow for automated updates as new data becomes available.

As with air quality monitoring, fusion for soil moisture also often occurs at the signal level, where raw sensor data is pre-processed and calibrated before integration with satellite imagery or other data sources. This type of fusion involves the use of sophisticated algorithms for real-time data calibration, quality control, and noise reduction, making it

possible to process large datasets from multiple sources efficiently. As depicted in Fig. 5, signal-level fusion is essential for achieving high-accuracy estimates in soil moisture monitoring, enabling real-time updates to soil moisture maps across vast agricultural regions.

The integration of edge computing and IoT technologies has also played a key role in soil moisture monitoring. By processing data at the edge, where sensors and satellite data are combined in near real-time, it is possible to reduce latency and enhance the timeliness of moisture measurements. These technologies further enable the development of scalable systems that can be deployed in diverse geographic regions, improving the overall utility of soil moisture monitoring in agricultural and environmental management.

4.4.3. Water quality

Water quality monitoring is a critical component of environmental monitoring systems, and its integration with data fusion technologies offers significant potential. However, compared to other environmental parameters like air quality and soil moisture, water quality data fusion is less frequently explored in the literature. In the single study we reviewed, [96] highlights the use of low-cost sensor (LCS) data combined with crowdsourced data to monitor water quality. This is also an example of the potential of mobile app development for data collection as opposed to staging purposes-built sensors, as observed in most of the studies. This application further stressed the role of citizen science in environmental monitoring.

4.4.4. Building occupancy

Building occupancy detection is an emerging area of research with significant implications for urban planning, energy management, and smart building technologies. This is especially relevant given the increasing housing challenges in urban areas. In recent years, the integration of low-cost sensor (LCS) data with advanced data fusion techniques has opened new avenues for improving occupancy detection systems.

The work of [97] serves as a foundational example in the field. Their study focused on detecting occupancy in residential buildings using low-cost sensors. Their aim was to integrate occupancy detection into automated building energy management systems. By fusing data from low-cost sensors, such as motion detectors and light sensors, they enabled real-time monitoring of building occupancy to optimize energy usage. The resulting system successfully provided binary occupancy detection, which is crucial for controlling heating, cooling, and lighting systems in response to actual occupancy, thereby improving energy efficiency. The technological approach here emphasizes real-time data collection and the application of data fusion for automating building management systems.

In a different technological approach, [66] compared machine learning models with multiple data sources to enhance the performance of occupancy prediction systems. The study's focus was on applying data fusion techniques to integrate diverse sensor data (e.g., temperature, humidity, and motion data) and assess the impact of data fusion on machine learning model accuracy. Using error measurement metrics, they compared the performance of occupancy models trained on single-source data against those using fused data from multiple sensors. This approach demonstrated that fusion improves the reliability of occupancy predictions, particularly in dynamic environments where individual sensor readings may be insufficient.

Lastly, [69] explored occupancy detection in smart buildings using environmental sensors. Their work focused on comparing late fusion methods with early fusion followed by an ensemble classifier. Late fusion aggregates individual sensor outputs after each sensor has made its own occupancy prediction, whereas early fusion combines raw sensor data before the prediction step. By testing these two fusion strategies, their results demonstrated that late fusion could offer advantages in terms of computational efficiency, particularly when dealing with large, complex sensor networks. Their work provides valuable insights into how different fusion strategies can be leveraged for efficient and scalable occupancy detection systems in smart buildings.

4.4.5. Soil organic content

The study by [52] demonstrates the potential of integrating multispectral and Synthetic Aperture Radar (SAR) datasets for estimating agricultural soil organic carbon (SOC). By fusing these data sources, the research achieved fine-scale mapping at a 10-meter resolution, enhancing the precision of SOC monitoring for agricultural applications.

From a technological standpoint, the use of the XGBoost algorithm stood out as an advanced machine learning technique, delivering superior predictive accuracy compared to other methods. Its ability to handle complex data interactions and rank feature importance made it ideal for SOC mapping. The study also identified critical predictors, such as vegetation indices and radar backscatter, optimising the fusion process and reducing computational demands.

These advancements in data fusion and machine learning offer scalable solutions for SOC monitoring, supporting sustainable agriculture and carbon management. However, challenges like data availability and computational complexity remain areas for further development.

4.4.6. Urban heat

Accurately modelling urban temperature at high spatial and temporal resolutions is critical for informing urban policies related to climate change and extreme heat events. Zumwald et al. [98] proposed a novel data fusion approach that integrates low-cost citizen weather station (CWS) data with machine learning algorithms to produce high-resolution urban temperature maps. The study's innovation lies in the elimination of the need for additional measurement infrastructure, making the approach cost-effective and scalable.

A key aspect of this work is the emphasis on careful model evaluation and the quantification of uncertainties. By addressing these aspects, the study ensures the reliability of temperature predictions, highlighting the role of robust data fusion and machine learning methodologies in advancing urban heat monitoring and mitigation strategies.

4.4.7. Noise

Noise sensing research has centred around development of advanced sensing units to monitor environmental and health impacts on human's effectively. Pradeep and Nagendra [83] focused on the technological aspects of designing and deploying low-cost noise sensors, enabling large-scale environmental noise monitoring. Similarly, [81] addressed urban noise challenges by integrating low-cost sensor networks with urban noise mapping systems. These studies emphasize the role of data fusion in combining sensor readings with geospatial and urban data, enabling actionable insights for urban noise management and health impact assessments.

4.5. Data pre-processing, calibration, and quality control

Data quality is a defining factor in the success of data fusion systems for environmental monitoring, and LCS present unique challenges compared to regulatory-grade instruments. Common data quality issues include elevated sensor noise, calibration drift, non-linear response to environmental conditions (e.g., temperature, relative humidity), signal degradation over time, outliers due to hardware faults or extreme events, and high rates of missing data resulting from power loss or communication failure [99–102]. The prevalence and magnitude of these issues are typically greater in LCS networks, reflecting their reduced hardware robustness and variable deployment environments. These quality problems have a direct and sometimes compounding effect on the reliability, interpretability, and scientific utility of fused data products, particularly in multi-sensor, multi-temporal, or multi-modal fusion pipelines where errors may propagate or amplify.

A rigorous pre-processing and quality control workflow is therefore essential before undertaking data fusion. At a foundational level,

most studies employ procedures such as outlier removal, using statistical thresholds or robust estimators, alongside mean or regression-based imputation for missing values, as well as data standardization or normalization to harmonize units and value ranges across different sensor streams [103,104]. Temporal and spatial alignment is another critical pre-processing step, especially for studies integrating asynchronous LCS, satellite overpasses, and meteorological measurements. As reported in Wei et al., aligning datasets by both time and geographic coordinates is fundamental to minimize information loss and avoid spurious correlations in downstream analysis [56]. More advanced quality control strategies, though less commonly implemented, are gaining traction in the literature. Adaptive filtering (e.g., moving average, median, or Kalman filters), model-based imputation (using probabilistic or machine learning models to infer missing values), and dynamic recalibration techniques (periodic co-location with reference instruments or algorithmic drift detection) have all been demonstrated to enhance LCS data reliability. For example, [105] applied supervised machine learning models to diagnose and correct data quality issues, identifying key factors affecting calibration in diverse urban micro-environments. Similarly, [54] explored lightweight data assimilation and polynomial interpolation for gap-filling and noise suppression, while [135] discussed initial conversion of raw LCS outputs into calibrated gas concentrations, highlighting the confidentiality and variability of manufacturer-provided conversion settings.

Calibration, while an essential step in improving sensor observations, remains a persistent and often under-addressed challenge, especially under real-world field conditions where LCS performance may diverge from laboratory calibration curves [106–108]. Field calibration is highly contextual, typically involving co-location with reference stations or “transfer calibration” with higher-grade sensors, is vital to correct for both systematic bias and environmental response variability [109–111]. More advanced techniques include machine learning methods, such as Random Forests (RF) [109], Artificial Neural Network (ANN) [112], and hybrid frameworks like HypeAIR, which perform real-time calibration using dynamic environmental inputs [113]. Further work in [114] proposed support vector regression (SVR) based calibration to enhance calibration strategies for LCS. Nevertheless, our review indicates that many studies omit detailed calibration protocols or fail to account for temporal drift, especially in deployments spanning multiple seasons or environmental regimes. In addition, the degree to which datasets from disparate sources, such as LCS, satellites, and meteorological stations, are temporally and spatially harmonized is not always reported, creating uncertainty around the validity of fusion outcomes and complicating efforts to compare or reproduce results.

A key point to note is that while research on calibration for air quality sensors is expanding, there is a huge contrast compared to low-cost water sensors [108]. Nalakurthi et al. attribute this to the absence of regulatory pertaining to water quality sensing technologies. This is also reflected in the number of studies identified under water as a parameter. See Fig. 7.

Uncertainty quantification and propagation are further critical dimensions of quality control. While some advanced studies utilize probabilistic or Bayesian approaches to explicitly model and propagate measurement uncertainty through the fusion process, the majority of reviewed works either ignore or inadequately treat uncertainty, limiting the transparency and decision-readiness of their products [115]. Effective management of uncertainty thus depends not only on post-fusion analysis but, fundamentally, on the rigour and sophistication of pre-processing, calibration, and quality control at the earliest stages of the data pipeline.

Overall, our synthesis suggests that the operational sophistication of data quality control methods varies widely between studies, with only a minority implementing state-of-the-art workflows for robust uncertainty management. Best practices, emerging in the literature and recommended for future work, include systematic outlier detection, dynamic sensor drift correction, advanced data imputation, routine field calibration, and explicit reporting of uncertainty quantification methods. The effectiveness and credibility of data fusion for environmental monitoring are intimately linked to these foundational quality control steps, and advancing methodological rigour in this domain remains a priority for the field.

4.6. Practical deployment challenges: sensor calibration, data transmission, and computational resources

The practical deployment of data fusion systems in real-world scenarios presents a set of persistent challenges that are often underestimated in controlled experimental studies but become paramount in operational contexts. Sensor calibration is perhaps the most critical, as the performance and reliability of LCS are directly affected by calibration accuracy and stability. LCS are susceptible to numerous sources of error, including temperature and humidity dependencies, cross-sensitivity to other pollutants, ageing, and drift, all of which can degrade measurement fidelity over time [108,116]. Calibration is not a one-off activity but an ongoing process; it often requires both initial co-location with reference instruments and periodic re-calibration to maintain performance in the face of changing environmental conditions and sensor drift [113,117]. While traditional calibration relies on linear or multiple linear regression models developed during short-term co-location campaigns, recent advances incorporate machine learning techniques such as Random Forests, Artificial Neural Networks, and Support Vector Regression, which can model non-linear sensor behaviour and dynamically adjust to environmental context [109,110,112]. Nonetheless, these approaches are challenged by limited transferability across locations, as calibration models trained in one environment may underperform elsewhere, and by the logistical complexity of re-calibrating large distributed networks-especially for deployments in remote or inaccessible areas. Furthermore, the field of water quality sensing lags significantly behind air quality in calibration standardization and methodological rigour, in part due to the absence of regulatory frameworks and the scarcity of comprehensive, multi-parameter reference datasets [10]. Data transmission reliability constitutes a second major hurdle for real-world deployment. Many LCS platforms lack built-in connectivity, necessitating manual retrieval of data via SD cards or periodic uploads, which introduces significant delays and limits the utility of near real-time data fusion applications [10]. Even in cases where connectivity is available, networks may be constrained by bandwidth, power, or security requirements, particularly in remote or sensitive environments. Emerging wireless technologies, such as ZigBee, LoRaWAN, Wi-Fi, and cellular modules, are beginning to address these gaps by enabling remote, automated data collection; however, such advances also introduce cybersecurity vulnerabilities and operational constraints that must be carefully managed [118,119]. In certain high-security settings, cloud-based telemetry is prohibited, mandating offline operation and placing even greater demands on local data storage and processing. The rise of edge computing, where pre-processing, anomaly detection, or even partial data fusion occurs locally on the sensor node, offers a promising mitigation for bandwidth and reliability issues, but it requires careful system design and may be limited by the computational capabilities of low-power devices. Finally, the computational resource requirements of advanced data fusion techniques, particularly geostatistical and probabilistic (Bayesian) models, pose significant implementation challenges for real-time or resource-constrained environments. While geostatistical approaches like kriging remain popular for their interpretability and integration with standard geospatial toolboxes, they are computationally intensive and scale poorly to large sensor networks or high-resolution spatial domains unless supported by distributed or parallel computing infrastructure [7,120]. Probabilistic data fusion, especially Bayesian inference, traditionally incurred prohibitive computational costs, but recent advances in scalable algorithms such as variational inference and Hamiltonian Monte Carlo, alongside the development of probabilistic programming frameworks (Stan, PyMC, TensorFlow Probability), are narrowing the gap between methodological

sophistication and operational feasibility [121–123]. Cloud-based platforms have made it increasingly practical to deploy and manage these complex models in near-real-time settings, but resource constraints still limit their use in edge or embedded systems where memory and processor speed are at a premium. For such deployments, model simplification, quantization, or the use of lightweight approximations remains essential. Altogether, addressing sensor calibration, data transmission, and computational resource challenges is fundamental for transitioning data fusion from proof-of-concept studies to robust, scalable, and trustworthy real-world systems. Sustained progress in these areas will require ongoing collaboration between sensor developers, data scientists, network engineers, and end-users, as well as careful attention to emerging best practices in both research and operational domains.

4.7. Synthesis of methodological strengths and limitations

This subsection consolidates cross-cutting observations on methodological practice across the 82 included studies.

Geostatistical approaches

Geostatistics has historically underpinned LCS data fusion especially for spatial interpolation, bias correction, and uncertainty propagation, as illustrated in [7,26,92] and more recently [45]. Variogram modelling and kriging/co-kriging provide transparent assumptions, explicit prediction variance, and interpretable spatial structure, which are advantageous when the signal exhibits quasi-stationary behaviour and dense reference networks exist. Limitations remain where non-stationarity, anisotropy, or sparse/reference-poor settings prevail, and in computational scaling for large spatio-temporal domains unless approximations (e.g., sparse structures, local kriging) are used.

Data-driven (ML/DL) approaches

Data-driven methods, random forests, gradient boosting, SVMs, and deep architectures (CNN/LSTM/transformer variants), have gained traction due to their capacity to model non-linear, high-dimensional relationships that may elude classical geostatistics. Reported gains in point accuracy are common; however, interpretability is often reduced, and uncertainty quantification (UQ) is infrequently implemented beyond ad-hoc ensembling. Risks include spatio-temporal leakage in validation splits and overfitting to local calibration regimes. Good practice includes blocked (spatial/temporal) cross-validation, explicit drift compensation, and UQ via quantile regression, conformal methods, Bayesian approximations, or calibrated predictive intervals.

Knowledge-based and hybrid approaches

A smaller subset of studies employed knowledge-based models (6 %) and hybrid models (9 %), particularly in complex or data-scant contexts where expert judgement complements numerical inference. Knowledge-based schemes (fuzzy rules, rule-based systems) offer traceable decision logic and robustness under domain constraints but may depend on expert-defined rules that limit portability and scaling. Hybrid designs such as kriging-residuals + ML, LUR + fuzzy logic, or Bayesian hierarchical + DL, seek to combine physical interpretability with predictive flexibility, yet introduce added complexity (model coupling, hyperparameter burden) and can compound validation challenges if not carefully modularised.

Cross-cutting gaps and emerging good practice

Across methodological classes, four recurring issues are evident:

(i) Calibration and drift: Co-location protocols and transfer functions are not consistently described (windows for drift detection, recalibration triggers), reducing reproducibility and external validity.
(ii) Uncertainty quantification: The majority of studies report only point metrics (RMSE/MAE/R^2); calibrated UQ (e.g., coverage, CRPS) is comparatively rare, limiting decision utility.
(iii) Validation design: Spatial/temporal blocking and external-site tests remain underused relative to random k-fold, risking optimistic performance estimates for operational deployment.
(iv) Data-quality control (QC): QC pipelines (range/persistence checks, humidity/temperature compensation, interference flags) vary widely in operational sophistication; only a minority document state-of-the-art workflows with thresholds and audit trails.

Collectively, the evidence indicates that uncertainty-aware evaluation, leakage-resistant validation, explicit calibration/drift procedures, and auditable QC are the most impactful levers to improve robustness, portability, and decision value of LCS data-fusion pipelines. Probabilistic and hybrid frameworks are positioned to bridge interpretability and predictive rigour, provided that UQ and deployment-oriented validation are made standard.

5. Discussions

5.1. What are the challenges and future directions of DF for LCS?

This section addresses the third research question: *What are the key challenges and opportunities associated with the use of data fusion in LCS applications?* Our findings emphasise that, while data fusion technologies hold significant potential for integrating diverse datasets, their practical implementation in LCS systems presents notable challenges that must be tackled to advance the field. Below, we outline the most prominent challenges and opportunities from a technological perspective.

5.2. Heterogeneous data sources integration

The fusion of data from sensors with varying resolutions and formats remains a bottleneck, particularly in multi-sensor setups. LCS and regulatory-grade instruments generate data with varying formats, frequencies, and levels of accuracy. This inherent diversity, given the nature of environmental sensing data, also presents an area of opportunity for refining methodologies and tools that fuse these different heterogeneous datasets. Combining such data requires advanced preprocessing and standardisation techniques, which can be computationally expensive.

A key gap in this area is that many fusion methods (e.g., geostatistics) lack built-in support for multimodal data integration. Additionally, the emergence of unstructured data types, such as social media, which provide supplementary information correlating with human activities, introduces new challenges. Methods for learning from and integrating such data are still in their infancy, making this integration a significant hurdle [124].

Furthermore, an end-to-end integration of IoT, machine learning, data fusion, and sensor calibration solutions to improve the data quality of low-cost sensors in environmental monitoring is still lacking. High-dimensional characteristics of data from low-cost sensors remain a challenge for researchers in machine learning and data mining. Feature selection offers an effective way to address this problem by removing irrelevant and redundant data, reducing computation time, improving learning accuracy, and facilitating a better understanding of learning models or data. These areas remain open for further exploration.

In future, research should prioritise the development of standardised frameworks and protocols for calibrating and validating low-cost sensors in diverse environmental conditions. Additionally, there is a need for studies exploring the development of models, architectures, and algorithms specifically designed for LCS data. Furthermore, research on monitoring noise and light pollution appears to be at an early stage. Studies in these areas could address critical gaps in urban environmental studies. Finally, investigating the ethical and social implications of deploying low-cost sensor networks, particularly in underserved regions, would be an interesting area to explore.

5.3. Practical deployment challenges: sensor calibration, data transmission, and computational resources

The practical deployment of data fusion systems in real-world scenarios presents a set of persistent challenges that are often underestimated in controlled experimental studies but become paramount in operational contexts. Sensor calibration is perhaps the most critical, as the performance and reliability of low-cost sensors (LCS) are directly affected by calibration accuracy and stability. LCS are susceptible to numerous sources of error, including temperature and humidity dependencies, cross-sensitivity to other pollutants, ageing, and drift, all of which can degrade measurement fidelity over time [108,116]. Calibration is not a one-off activity but an ongoing process; it often requires both initial co-location with reference instruments and periodic re-calibration to maintain performance in the face of changing environmental conditions and sensor drift [113,117]. While traditional calibration relies on linear or multiple linear regression models developed during short-term co-location campaigns, recent advances incorporate machine learning techniques such as Random Forests, Artificial Neural Networks, and Support Vector Regression, which can model non-linear sensor behaviour and dynamically adjust to environmental context [109,110,112]. Nonetheless, these approaches are challenged by limited transferability across locations, as calibration models trained in one environment may underperform elsewhere, and by the logistical complexity of re-calibrating large distributed networks-especially for deployments in remote or inaccessible areas. Furthermore, the field of water quality sensing lags significantly behind air quality in calibration standardization and methodological rigour, in part due to the absence of regulatory frameworks and the scarcity of comprehensive, multi-parameter reference datasets. Data transmission reliability constitutes a second major hurdle for real-world deployment. Many LCS platforms lack built-in connectivity, necessitating manual retrieval of data via SD cards or periodic uploads, which introduces significant delays and limits the utility of near real-time data fusion applications. Even in cases where connectivity is available, networks may be constrained by bandwidth, power, or security requirements, particularly in remote or sensitive environments. Emerging wireless technologies, such as ZigBee, LoRaWAN, Wi-Fi, and cellular modules, are beginning to address these gaps by enabling remote, automated data collection; however, such advances also introduce cybersecurity vulnerabilities and operational constraints that must be carefully managed [118,119]. In certain high-security settings, cloud-based telemetry is prohibited, mandating offline operation and placing even greater demands on local data storage and processing. The rise of edge computing, where pre-processing, anomaly detection, or even partial data fusion occurs locally on the sensor node, offers a promising mitigation for bandwidth and reliability issues, but it requires careful system design and may be limited by the computational capabilities of low-power devices.

Finally, the computational resource requirements of advanced data fusion techniques, particularly geostatistical and probabilistic (Bayesian) models, pose significant implementation challenges for real-time or resource-constrained environments. While geostatistical approaches like kriging remain popular for their interpretability and integration with standard geospatial toolboxes, they are computationally intensive and scale poorly to large sensor networks or high-resolution spatial domains unless supported by distributed or parallel computing infrastructure [7,120]. Probabilistic data fusion, especially Bayesian inference, traditionally incurred prohibitive computational costs, but recent advances in scalable algorithms such as variational inference and Hamiltonian Monte Carlo, alongside the development of probabilistic programming frameworks (Stan, PyMC, TensorFlow Probability), are narrowing the gap between methodological sophistication and operational feasibility [121–123]. Cloud-based platforms have made it increasingly practical to deploy and manage these complex models in near-real-time settings, but resource constraints still limit their use in edge or embedded systems where memory and processor speed are at a premium. For such deployments, model simplification, quantisation, or the use of lightweight approximations remains essential. Altogether, addressing sensor calibration, data transmission, and computational resource challenges is fundamental for transitioning data fusion from proof-of-concept studies to robust, scalable, and trustworthy real-world systems. Sustained progress in these areas will require ongoing collaboration between sensor developers, data scientists, network engineers, and end-users, as well as careful attention to emerging best practices in both research and operational domains.

5.4. Data privacy and security in low-cost sensor data fusion

The rapid proliferation of low-cost sensors (LCS) and mobile crowdsensing in environmental monitoring has brought data privacy and security to the forefront of research and operational practice. As sensor networks increasingly capture and fuse geo-referenced, time-resolved, and potentially sensitive personal or behavioural data, the risk of individual re-identification, exposure of commuting patterns, or inference of health-related behaviours becomes significant [118]. These risks are often not systematically addressed in LCS system design or deployment, yet they represent a critical barrier to public acceptance, ethical compliance, and broad adoption of participatory sensing [125]. Recent advances in privacy-preserving data fusion focus on mitigating these risks through technical and organizational means. Privacy-preserving task allocation mechanisms, such as those based on geographic differential privacy and group-based noise addition, have been proposed to obscure precise user locations or sensing patterns while preserving data utility for fusion tasks [126]. Furthermore, group-based noise addition injects calibrated randomness into the reported location data, achieving geo-indistinguishability and limiting adversarial inference risks without degrading the value of aggregate spatial analyses [127]. Such strategies are particularly valuable in urban-scale applications, where high-resolution mobility or environmental data can otherwise reidentify individual trajectories.

In distributed and edge-enabled sensing systems, privacy is further protected through mechanisms like privacy-preserving task allocation (P2TA), which optimize the trade-off between privacy guarantees and resource utilization, often leveraging edge computing to locally preprocess or anonymize data before transmission [128]. Inference over distributed or cloud-based sensor networks raises additional confidentiality concerns, as raw or minimally processed data may be exposed during transmission or model deployment. To address this, recent works have developed secure multi-party computation and cryptographic inference protocols such as Panther, which enables secure two-party neural network inference over encrypted data, ensuring that neither data providers nor model owners are required to reveal sensitive information [129], and similarly, Lin et al.[130], efficient secure inference framework for industrial IoT, which allows multiparty model evaluation with minimal data exposure and robust protection against inference attacks. Federated learning and privacy-preserving distributed optimization are also increasingly explored, allowing model training and inference without centralizing sensitive raw data, though these approaches often entail substantial communication overhead and complex coordination among network participants. Beyond inference, LCS networks present practical security challenges due to severe hardware and connectivity constraints. Many low-cost devices lack robust encryption, authentication, or tamper detection, making them vulnerable to spoofing, data injection, and unauthorized access. These limitations make them vulnerable to spoofing, injection attacks, and unauthorised access [131,132]. Additionally, sensor networks used in air quality monitoring are prone to data integrity issues, as discussed by Lou et al.[125]

Data integrity concerns are particularly acute in environmental monitoring applications, where compromised sensor data can directly impact public health decisions and regulatory compliance. Lightweight cryptographic protocols and integrity verification schemes are being developed to secure communications without overburdening limited

sensor resources. In field deployments where wireless connectivity is infeasible or prohibited-such as in highly secure facilities or rural locations-data are often logged locally and collected manually, introducing additional concerns over completeness, timeliness, and the risk of data loss or tampering during transfer.

As the scale and complexity of LCS-based environmental monitoring expand, data privacy and security must be recognised as foundational requirements. Best practices recommend the integration of privacy-by-design principles at all levels, including anonymisation, aggregation, noise injection, secure inference, and resilient protocol design. The continued development and standardization of privacy-preserving and secure data fusion methodologies will be essential to building public trust and ensuring the long-term sustainability of participatory environmental monitoring initiatives.

5.5. Uncertainty quantification and management in data fusion

Uncertainty quantification and management are fundamental to ensuring the reliability and transparency of data fusion systems, especially when integrating heterogeneous, multi-source measurements or models typical of low-cost sensor networks. Uncertainties can arise from intrinsic sensor noise, calibration bias, environmental variability, spatial and temporal mismatches, as well as from the fusion algorithms themselves. If not explicitly modelled and managed, these uncertainties can propagate or even amplify through the fusion process, undermining the scientific value and practical utility of the final fused outputs [7,54,133]. Within geostatistical methods, uncertainty quantification is generally intrinsic to the methodology. Kriging and its variants, for example, not only provide interpolated values but also estimate the variance of predictions, offering a spatially explicit measure of confidence across the domain [48]. In universal kriging, the total prediction uncertainty combines both regression and spatial interpolation components, which is particularly useful in environmental monitoring with sparse or unevenly distributed sensors. Such spatial uncertainty maps are invaluable for risk assessment, regulatory compliance, and network design, allowing practitioners to target additional sampling or prioritise decisions based on confidence intervals. Probabilistic methods, including Bayesian inference and Gaussian Process (GP) modelling, offer a principled framework for quantifying and managing uncertainty. Bayesian models treat unknowns as probability distributions, systematically updating beliefs as new data become available [134]. This enables explicit propagation of both measurement and model uncertainties throughout the data fusion pipeline, with outcomes such as credible intervals, posterior variance, or full predictive distributions for the target variables. In addition, probabilistic scores like the Continuous Ranked Probability Score (CRPS) and coverage probabilities can be used to benchmark model calibration and reliability. These methods, however, can be computationally demanding, requiring advanced sampling (e.g., Markov Chain Monte Carlo, Hamiltonian Monte Carlo) or variational inference techniques for practical deployment in large-scale or real-time applications. In contrast, many machine learning and deep learning approaches, while powerful for prediction, often lack intrinsic mechanisms for uncertainty estimation. Common ML models such as Random Forests, Support Vector Machines, and classical neural networks typically produce point estimates, with uncertainty sometimes inferred through empirical error metrics (e.g., RMSE, MAE, cross-validation) or via post hoc bootstrapping and ensemble variance analysis. For regulatory or high-stakes decision-making, this lack of explicit uncertainty quantification can be a significant limitation. Emerging research is exploring the use of Bayesian neural networks, Monte Carlo dropout, and quantile regression forests to address this gap, enabling ML models to communicate predictive confidence more transparently. Furthermore, in cases where LUR models are used, most studies focus on maximising R^2, which lacks robustness and furthermore lacks transferability of the resulting models [136]. Hybrid and ensemble approaches are increasingly adopted to harness the strengths of both statistical and machine learning frameworks. For example, combining kriging or probabilistic outputs with machine learning predictions can produce spatially explicit, uncertainty-aware fused datasets, as shown in recent studies applying co-kriging with random forests or blending probabilistic fusion with data-driven calibration [61]. Such methods allow for uncertainty management across both spatial and temporal domains, supporting robust and context-aware environmental intelligence. Finally, explicit consideration of uncertainty is not merely a technical issue but also a communication and governance priority. Transparent reporting of predictive intervals, confidence levels, and model limitations enhances the interpretability and trustworthiness of fused data, supporting risk-informed decision-making and fostering stakeholder engagement. In summary, while advances in geostatistics and probabilistic modelling have made rigorous uncertainty quantification increasingly accessible, its practical adoption in data fusion workflows remains uneven, particularly in machine learning-centric research. Best practice recommends integrating uncertainty estimation and propagation as a core design principle, alongside thorough calibration, cross-validation, and transparent reporting, to maximize the scientific and societal value of data fusion in environmental monitoring.

5.6. Emerging trends and opportunities

5.6.1. AI-driven fusion

The integration of AI techniques, particularly ML, is an emerging trend that addresses computational challenges in data fusion. AI models capable of operating on-device offer a promising option for real-time environmental monitoring and could be a foundation for developing the next generation of sensors. Specifically, purpose-designed convolutional neural networks (CNNs) can effectively process spatial data, while recurrent neural networks (RNNs) handle temporal dynamics, thereby creating a comprehensive fusion framework. Furthermore, pre-trained models and tools like AutoML reduce the barriers to deploying AI in fusion pipelines. This is particularly worth exploring in studies utilising land-use regression (LUR) models. By leveraging AI, researchers can move beyond traditional deterministic fusion methods to develop adaptive, efficient systems capable of managing large-scale, dynamic datasets.

Designing fusion models that integrate domain knowledge from low-cost sensors is another promising area requiring further exploration. For example, probabilistic programming provides flexible approaches to account for data and model uncertainties [124]. This allows for modelling non-linear problems from a programmatic perspective. Tools supporting such implementations, like PyMC[4], are already available. Additionally, borrowing concepts from emerging fields such as urban computing [137] could further enhance fusion methodologies and broaden their applicability.

5.6.2. Cloud and edge computing

Emerging computing technologies offer robust platforms for scaling data fusion applications. Cloud platforms, such as Google Earth Engine, provide scalable environments for processing large datasets, particularly in integrating satellite imagery-derived data with modern machine learning and deep learning frameworks. This can be enhanced by edge computing architectures, where data quality assurance strategies are implemented as part of middleware. By fusing data locally before transmitting summarised insights to the cloud, latency is reduced, and bandwidth usage is optimised.

5.6.3. Hybrid fusion approaches: rationale, benefits, and challenges

Hybrid fusion approaches, which integrate two or more data fusion methodologies (such as geostatistics, machine learning, and probabilistic modelling), are gaining increasing traction as a means to leverage the complementary strengths of individual techniques while mitigating their respective limitations. The underlying rationale for hybridization

[4] https://www.PyMC.io/welcome.html

stems from the inherent complexity of environmental sensor data, which often exhibit non-stationary spatial and temporal patterns, non-linear relationships, and varying levels of uncertainty and data quality. No single fusion method is universally optimal across all scenarios: geostatistical methods excel at modelling spatial dependencies and providing uncertainty estimates, while machine learning techniques are highly effective at capturing complex, dynamic, and potentially non-linear temporal or feature relationships within the data [7,48]. By judiciously combining these methods, hybrid fusion frameworks can address the limitations of each component and deliver more robust, accurate, and context-sensitive data products [14]. A common hybrid strategy is to use geostatistics (e.g., kriging or co-kriging) for spatial interpolation, filling gaps in sensor coverage and generating uncertainty maps, while employing machine learning (e.g., deep neural networks, random forests, or recurrent neural networks) to model temporal trends, predict future states, or capture complex interactions among environmental variables [61]. For instance, in air quality mapping, spatial prediction of pollutant concentrations can be achieved using kriging based on both regulatory-grade and low-cost sensor data, while temporal dynamics and local deviations can be modelled using neural networks or ensemble ML algorithms trained on historical time series and auxiliary meteorological data. This approach has been further enhanced by integrating remote sensing products (e.g., satellite-derived land cover, aerosol optical depth, soil moisture) to inform geostatistical models or as features for machine learning models, providing broad spatial context and multi-source data fusion capabilities. The benefits of hybrid fusion are well-documented. Such approaches have demonstrated improved predictive performance, more reliable uncertainty quantification, and enhanced adaptability to heterogeneous and missing data, especially in large-scale or data-sparse settings. For example, [61] applied a fuzzy logic-based hybrid approach to combine stationary land use LUR models with low-cost sensor data, significantly improving urban air quality predictions. Similarly, [48] and [7] report the effective use of hybrid kriging-machine learning and Bayesian-ML approaches for spatio-temporal air quality (e.g., $PM_{2.5}$ and NO_2) and soil moisture estimation, yielding both high accuracy and spatially explicit uncertainty assessments. Similarly, [75] explored the combination of Gaussian Processes and machine learning to model spatio-temporal air quality data, successfully addressing the challenge of uncertainty quantification while improving predictive accuracy.

The observed 11 % share of hybrid fusion studies confirms that these methods are progressing from isolated case studies to a reproducible and emerging methodological trend, combining interpretability, uncertainty awareness, and computational adaptability within operational LCS fusion frameworks.

Despite these advantages, hybrid fusion approaches are not without challenges. Combining different methodologies introduces substantial model complexity, with increased computational requirements and intricate workflows for data preprocessing, harmonisation, and quality control. Model compatibility is a key concern; for instance, integrating probabilistic models (which yield distributions) with deterministic machine learning outputs (which provide point estimates) requires careful consideration of how uncertainties are propagated and communicated within the system. Interpretability can be reduced as hybrid systems become more complex, sometimes limiting transparency and stakeholder trust-an issue that is particularly salient in environmental management and regulatory applications. The integration process may also introduce new sources of error if assumptions, data requirements, or scale mismatches are not properly addressed. Additionally, the effective implementation of hybrid fusion demands advanced data preprocessing pipelines to manage differing data quality standards, imputation needs, and calibration errors across multiple input sources. Real-time deployment of hybrid systems, especially those involving deep learning and large geostatistical models, poses significant computational challenges, necessitating the use of distributed computing, cloud platforms, or efficient edge processing. Looking forward, the optimisation and scalability of hybrid fusion frameworks remain active areas of research. Promising directions include the development of modular, interpretable hybrid architectures that can adaptively select or weight different models based on data context; the integration of error quantification and propagation mechanisms, such as Bayesian inferences or ensemble uncertainty estimation; and the use of explainable AI techniques to improve model transparency. As environmental sensor networks and remote sensing datasets continue to expand in scale and complexity, hybrid fusion is poised to become a cornerstone of robust, reliable, and actionable environmental intelligence.

6. Conclusion and outlook

The findings from this study make several contributions to the state of the art in data fusion for low-cost sensors (LCS) in environmental monitoring. Firstly, this study acts as the first of its kind in filling in the gap of a dedicated systematic review for low-cost sensors data fusion. It has identified the key fusion methodologies currently employed in LCS applications, including geostatistical techniques and data-driven. Secondly, it also highlights the dominant areas of study, such as air quality and soil moisture monitoring, and provides a detailed analysis of the methodologies used to improve data quality. It is worth noting that despite data fusion for LCS being a relatively young field, significant advancements have been made in sensing technology and data pipelines as well as data integration models, which continues to improve addressing data quality challenges. Lastly, the study contributes to the development of a structured taxonomy of data fusion approaches, offering a useful tool for researchers to select appropriate methodologies for specific environmental monitoring contexts.

However, the study also underscores the critical challenges facing the field, including sensor limitations, data heterogeneity, and the lack of standardised frameworks for integrating LCS data with other data sources, such as satellite imagery or regulatory-grade instruments.

A limitation of this review lies in the relatively modest sample size (82 studies), which, although representative of current peer-reviewed work, may not capture the full breadth of rapidly evolving research in low-cost sensing and data-fusion methodologies.

In addition, our reliance on the term *“low-cost sensor”* in the primary search strategy may have contributed to the observed dominance of air-quality studies, as this phrase is particularly prevalent in that discipline. An exploratory supplementary search (Section 2.5; Appendix A) using broader terminology, *“affordable sensor”*, *“community monitoring”* and *“in-situ sensor”*, identified additional relevant work in water, soil, and noise domains, suggesting that future systematic reviews should employ more inclusive vocabularies to ensure comprehensive coverage. Lastly, the initial abstract and title screening was conducted by a single reviewer; we implemented independent co-author verification of all inclusions and extracted attributes. In future we hope to conduct a dual independent screening, the PRISMA gold standard, will be adopted in future updates, resources permitting.

Despite these limitations, the synthesis presented here provides a robust baseline for understanding the methodological landscape of data fusion for low-cost sensing and offers actionable guidance for advancing uncertainty-aware, hybrid, and privacy-preserving approaches in future deployments.

Looking forward, future research must focus on overcoming the limitations identified in this review, particularly in terms of sensor calibration, data standardisation, and real-time analytics. Furthermore, by addressing the gaps in sensor data quality and expanding on the integration of AI and machine learning models, the research community can better utilise LCS-driven data fusion to create more accurate, adaptive,

and scalable systems for monitoring and managing environmental sustainability.

CRediT authorship contribution statement

Gabriel Oduori: Writing – review & editing, Writing – original draft, Methodology, Conceptualization; **Chaira Cocco:** Writing – review & editing, Supervision; **Payam Sajadi:** Writing – review & editing, Methodology; **Francesco Pilla:** Writing – review & editing, Funding acquisition, Supervision.

Data availability

No data was used for the research described in the article.

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.

Acknowledgements

This research has received funding from the European Union, Horizon Europe research and innovation programme under CitiObs project, (grant agreement 101086421), and University College Dublin.

Appendix A. Summary of supplementary search records

A supplementary exploratory search was conducted using expanded terminology to assess sensitivity to search terms. The search identified 32 unique records, of which seven were directly relevant to low-cost sensing and data fusion Table A.6. These additional records were reviewed qualitatively but excluded from the quantitative synthesis to preserve methodological consistency. Their inclusion reinforces the observation that alternative terminology-particularly within citizen science and participatory sensing-captures additional, relevant studies not retrieved by the “low-cost sensor” query alone. The records are summarised below.

Table A.6
Summary of relevant studies identified through supplementary search.

Author(s)	Year	Title	Focus	DOI
Liang et al.	2005	A distributed geospatial infrastructure for Sensor Web	Decision Support	https://doi.org/10.1016/j.cageo.2004.06.014
Ali et al.	2023	A high performance-oriented AI-enabled IoT-based pest detection system using sound analytics in large agricultural field	Agriculture	https://doi.org/10.1016/j.micpro.2023.104946
Carminati et al.	2020	A self-powered wireless water quality sensing network enabling smart monitoring of biological and chemical stability in supply systems	Water Quality	https://doi.org/10.3390/s20041125
Han et al.	2005	A time-continuous land surface temperature (LST) data fusion approach based on deep learning with microwave remote sensing and high-density ground truth observations	Temperature	https://doi.org/10.1016/j.scitotenv.2024.169992
Yazdinejad et, al	2025	Advanced AI-driven methane emission detection, quantification, and localization in Canada: A hybrid multi-source fusion framework	Air Quality	https://doi.org/10.1016/j.scitotenv.2025.180142
Purwanto et al.	2022	Assessment of the dynamics of urban surface temperatures and air pollution related to COVID-19 in a densely populated City environment in East Java	Temperature	https://doi.org/10.1016/j.ecoinf.2022.101809
Yoon	2023	Building digital twinning: Data, information, and models	Digital Twin	https://doi.org/10.1016/j.jobe.2023.107021
Rojas et al.	2024	Combining multi-satellite remote and in-situ sensing for unmanned underwater vehicle state estimation	Water Quality	https://doi.org/10.1016/j.oceaneng.2024.118708
Sagl et al.	2015	Contextual sensing: Integrating contextual information with human and technical geo-sensor information for smart cities	Smart Cities	https://doi.org/10.3390/s150717013
Ferrer-Cid et al.	2022	Data reconstruction applications for IoT air pollution sensor networks using graph signal processing	Air Quality	https://doi.org/10.1016/j.jnca.2022.103434
Simone et al.	2025	Deep learning framework for cardiorespiratory disease detection using smartphone IMU sensors	Health	https://doi.org/10.1016/j.compbiomed.2025.110595
Stavropoulos et al.	2017	DemaWare2: Integrating sensors, multimedia and semantic analysis for the ambient care of dementia	Health	https://doi.org/10.1016/j.pmcj.2016.06.006
Monk et al.	2021	Detecting and mapping a CO2 plume with novel autonomous ph sensors on an underwater vehicle	Air Quality	https://doi.org/10.1016/j.ijggc.2021.103477
Xuan et al.	2022	Early diagnosis and pathogenesis monitoring of wheat powdery mildew caused by blumeria graminis using hyperspectral imaging	Agriculture	https://doi.org/10.1016/j.compag.2022.106921
Elmes et al.	2017	Effects of urban tree canopy loss on land surface temperature magnitude and timing	Smart Cities	https://doi.org/10.1016/j.isprsjprs.2017.04.011
Ibrahim et al.	2025	Embedded framework for low-cost pavement condition evaluation using microcontroller and single-board computer platforms	Smart Cities	https://doi.org/10.1016/j.autcon.2025.106442
Shin et al.	2024	Enhancing digital twin efficiency in indoor environments: Virtual sensor-driven optimization of physical sensor combinations	Digital Twin	https://doi.org/10.1016/j.autcon.2024.105326
Yoon and Koo	2023	In situ model fusion for building digital twinning	Digital Twin	https://doi.org/10.1016/j.buildenv.2023.110652
Oloo et al.	2018	Multi-dimensionality of uncertainty in big geospatial sensor data	Review	https://doi.org/10.1553/GISCIENCE2018_01_S3
Vahidi et al.	2025	Multi-Modal sensing for soil moisture mapping: Integrating drone-based ground penetrating radar and RGB-thermal imaging with deep learning	Soil	https://doi.org/10.1016/j.compag.2025.110423
Chen et al.	2023	Multi-Source Soil Moisture Data Fusion Based on Spherical Cap Harmonic Analysis and Helmert Variance Component Estimation in the Western U.S.	Soil	https://doi.org/10.3390/s23198019
Klug and Kmoch	2015	Operationalizing environmental indicators for real time multipurpose decision making and action support	Decision Support	https://doi.org/10.1016/j.ecolmodel.2014.04.009
Hobson et al.	2019	Opportunistic occupancy-count estimation using sensor fusion: A case study	Building Occupancy	https://doi.org/10.1016/j.buildenv.2019.05.032
Homan et al.	2025	Optimising multi-site sensor networks in lowland permeable catchments for comprehensive water quality monitoring and nitrogen mass balancing during baseflow conditions	Water Quality	https://doi.org/10.1016/j.watres.2025.123874
Díaz et al.	2013	Publishing sensor observations into Geospatial Information Infrastructures: A use case in fire danger assessment	Decision Support	https://doi.org/10.1016/j.envsoft.2013.06.002
Li et al.	2024	Smart Hospital Privacy Protection System Based On Cloud Computing IoT	Health	https://doi.org/10.1016/j.procs.2024.09.069
Luo Xie et al.	1994	Towards an implantable and refillable glucose sensor based on oxygen electrode principles	Health	https://doi.org/10.1016/0925-4005(94)87041-1
Reza et al.	2022	Ubiquitous GIS based outdoor evacuation assistance: An effective response to earthquake disasters	Decision Support	https://doi.org/10.1016/j.ijdrr.2022.103232
Santos-Fernández et al.	2024	Unsupervised Anomaly Detection in Spatio-Temporal Stream Network Sensor Data	Water Quality	https://doi.org/10.1029/2023WR035707
Himeur et al.	2022	Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives	Review	https://doi.org/10.1016/j.inffus.2022.06.003
Yoon et al.	2024	Virtual in-situ modelling between digital twin and BIM for advanced building operations and maintenance	Digital Twin	https://doi.org/10.1016/j.autcon.2024.105823
Olatinwo et al.	2024	Water Quality Assessment Tool for On-Site Water Quality Monitoring	Water Quality	https://doi.org/10.1109/JSEN.2024.3383887

References

[1] A. Middel, N. Nazarian, M. Demuzere, B. Bechtel, Urban climate informatics: an emerging research field, Front. Environ. Sci. 10 (2022). https://doi.org/10.3389/FENVS.2022.867434

[2] N. Castell, F.R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, A. Bartonova, et al., Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?, Environ. Int. 99 (2017) 293–302. https://doi.org/10.1016/j.envint.2016.12.007

[3] A.C. Rai, P. Kumar, F. Pilla, A.N. Skouloudis, S. Di Sabatino, C. Ratti, A. Yasar, D. Rickerby, et al., End-user perspective of low-cost sensors for outdoor air pollution monitoring, Sci. Total Environ. 607–608 (2017) 691–705. https://doi.org/10.1016/j.scitotenv.2017.06.266

[4] J. Bi, J. Stowell, E.Y.W. Seto, P.B. English, M.Z. Al-Hamdan, P.L. Kinney, F.R. Freedman, Y. Liu, Contribution of low-cost sensor measurements to the prediction of $PM_{2.5}$ levels: a case study in imperial county, California, USA, Environ. Res. 180 (2020) 108810. https://doi.org/10.1016/j.envres.2019.108810

[5] Y.-C. Lin, W.-J. Chi, Y.-Q. Lin, The improvement of spatial-temporal resolution of $PM_{2.5}$ estimation based on micro-air quality sensors by using data fusion technique, Environ. Int. 134 (2020) 105305. https://doi.org/10.1016/j.envint.2019.105305

[6] R.S. Blum, Z. Liu, Signal processing and communications, Multi-sensor Image Fusion and Its Applications, Taylor & Francis, 2006. https://books.google.ie/books?id=uVZ_tgEACAAJ.

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

[8] E.G. Snyder, T.H. Watkins, P.A. Solomon, E.D. Thoma, R.W. Williams, G.S.W. Hagler, D. Shelow, D.A. Hindin, V.J. Kilaru, P.W. Preuss, The changing paradigm of air pollution monitoring, Environ. Sci. Technol. 47 (20) (2013) 11369–11377.

[9] X. Fang, Improving Data Quality for Low-Cost Environmental Sensors, Ph.D. thesis, University of York, 2018.

[10] L. Morawska, P.K. Thai, X. Liu, A. Asumadu-Sakyi, G. Ayoko, A. Bartonova, A. Bedini, F. Chai, B. Christensen, M. Dunbabin, et al., Applications of low-cost sensing technologies for air quality monitoring and exposure assessment: how far have they gone?, Environ. Int. 116 (2018) 286–299.

[11] F.E. White, Data Fusion Lexicon, Technical Report, Defense Technical Information Center, Fort Belvoir, VA, 1991. https://doi.org/10.21236/ADA529661

[12] L.A. Klein, Sensor and Data Fusion: A Tool for Information Assessment and Decision Making, Society of Photo Optical, 2004. https://books.google.ie/books?id=-782bo4u_ogC.

[13] H. Boström, S.F. Andler, M. Brohede, R. Johansson, A. Karlsson, J. Van Laere, L. Niklasson, M. Nilsson, A. Persson, T. Ziemke, On the definition of information fusion as a field of research, 2007,

[14] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (1) (2013) 28–44. https://doi.org/10.1016/j.inffus.2011.08.001

[15] F. Castanedo, A review of data fusion techniques, Sci. World J. 2013 (2013) 1–19. https://doi.org/10.1155/2013/704504

[16] F. Alam, R. Mehmood, I. Katib, N.N. Albogami, A. Albeshri, Data fusion and IoT for smart ubiquitous environments: a survey, IEEE Access 5 (2017) 9533–9554. https://doi.org/10.1109/ACCESS.2017.2697839

[17] B.P.L. Lau, S.H. Marakkalage, Y. Zhou, N.U. Hassan, C. Yuen, M. Zhang, U.-X. Tan, A survey of data fusion in smart city applications, Inf. Fusion 52 (2019) 357–374. https://doi.org/10.1016/j.inffus.2019.05.004

[18] W. Ding, X. Jing, Z. Yan, L.T. Yang, A survey on data fusion in internet of things: towards secure and privacy-preserving fusion, Inf. Fusion 51 (2019) 129–144. https://doi.org/10.1016/j.inffus.2018.12.001

[19] R. Krishnamurthi, A. Kumar, D. Gopinathan, A. Nayyar, B. Qureshi, An overview of IoT sensor data processing, fusion, and analysis techniques, Sensors 20 (21) (2020) 6076. Number: 21 Publisher: Multidisciplinary Digital Publishing Institute, https://doi.org/10.3390/s20216076

[20] T. Meng, X. Jing, Z. Yan, W. Pedrycz, A survey on machine learning for data fusion, Inf. Fusion 57 (2020) 115–129. https://linkinghub.elsevier.com/retrieve/pii/S1566253519303902. https://doi.org/10.1016/j.inffus.2019.12.001

[21] M. Ouhami, A. Hafiane, Y. Es-Saady, M. El Hajji, R. Canals, Computer vision, IoT and data fusion for crop disease detection using machine learning: a survey and ongoing research, Remote Sens. 13 (13) (2021) 2486. https://www.mdpi.com/2072-4292/13/13/2486. https://doi.org/10.3390/rs13132486

[22] Y. Himeur, B. Rimal, A. Tiwary, A. Amira, Using artificial intelligence and data fusion for environmental monitoring: a review and future perspectives, Inf. Fusion 86–87 (2022) 44–75. https://doi.org/10.1016/j.inffus.2022.06.003

[23] A. Karagiannopoulou, A. Tsertou, G. Tsimiklis, A. Amditis, Data fusion in earth observation and the role of citizen as a sensor: a scoping review of applications, methods and future trends, Remote Sens. 14 (5) (2022) 1263. https://doi.org/10.3390/rs14051263

[24] C. Ounoughi, B.H. Sadok, Data fusion for ITS: a systematic literature review, Inf. Fusion 89 (2023) 267–291. https://www.sciencedirect.com/science/article/pii/S1566253522001087. https://doi.org/10.1016/j.inffus.2022.08.016

[25] M.A. Fadhel, A.M. Duhaim, A. Saihood, A. Sewify, M.N.A. Al-Hamadani, A.S. Albahri, L. Alzubaidi, A. Gupta, S. Mirjalili, Y. Gu, Comprehensive systematic review of information fusion methods in smart cities and urban environments, Inf. Fusion 107 (2024) 102317.

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

[27] A. Carrera-Rivera, W. Ochoa, F. Larrinaga, G. Lasa, How-to conduct a systematic literature review: a quick guide for computer science research, MethodsX 9 (2022) 101895.

[28] B. Kitchenham, S. Charters, Guidelines for performing systematic literature reviews in software engineering, 2007,

[29] C. Sohrabi, T. Franchi, G. Mathew, A. Kerwan, M. Nicola, M. Griffin, M. Agha, R. Agha, PRISMA 2020 statement: What's new and the importance of reporting guidelines, Int. J. Surgery 88 (2021) 105918. https://www.sciencedirect.com/science/article/pii/S1743919121000522. https://doi.org/10.1016/j.ijsu.2021.105918

[30] K.W. Khaw, A. Alnoor, H. Al-Abrrow, V. Tiberius, Y. Ganesan, N.A. Atshan, Reactions towards organizational change: a systematic literature review, Current Psychol. 42 (22) (2023) 19137–19160.

[31] A.P. Siddaway, A.M. Wood, L.V. Hedges, How to do a systematic review: a best practice guide for conducting and reporting narrative reviews, meta-analyses, and meta-syntheses, Annu. Rev. Psychol. 70 (2019) 747–770. https://api.semanticscholar.org/CorpusID:51941844.

[32] X. Zhu, F. Cai, J. Tian, T.K.-A. Williams, Spatiotemporal fusion of multisource remote sensing data: literature survey, taxonomy, principles, applications, and future directions, Remote Sens. 10 (4) (2018). https://www.mdpi.com/2072-4292/10/4/527. https://doi.org/10.3390/rs10040527

[33] T. Baltrušaitis, C. Ahuja, L.-P. Morency, Multimodal machine learning: a survey and taxonomy, IEEE Trans. Pattern Anal. Mach. Intell. 41 (2) (2019) 423–443. https://doi.org/10.1109/TPAMI.2018.2798607

[34] R. Gutiérrez, V. Rampérez, H. Paggi, J.A. Lara, J. Soriano, On the use of information fusion techniques to improve information quality: taxonomy, opportunities and challenges, Inf. Fusion 78 (2022) 102–137.

[35] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (7553) (2015) 436–444.

[36] I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, MIT press, 2016.

[37] M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, F. Prabhat, Deep learning and process understanding for data-driven earth system science, Nature 566 (7743) (2019) 195–204.

[38] J. Li, A.D. Heap, Spatial interpolation methods applied in the environmental sciences: a review, Environ. Model. Softw. 53 (2014) 173–189.

[39] P.L. Houtekamer, H.L. Mitchell, Ensemble kalman filtering, Quart. J. R. Meteorol. Soc. J. Atmos. Sci. Appl. Meteorol. Phys. Oceanograph. 131 (613) (2005) 3269–3289.

[40] E. Snelson, Z. Ghahramani, Sparse Gaussian processes using pseudo-inputs, Adv. Neural Inf. Process. Syst. 18 (2005).

[41] N. Cressie, Statistics for Spatial Data, John Wiley & Sons, 2015.

[42] M. Bobbia, J.-M. Poggi, B. Portier, Spatial correction of low-cost sensors observations for fusion of air quality measurements, Appl. Stoch. Models Bus. Ind. 38 (5) (2022) 766–786.

[43] S. Han, W. Kundhikanjana, P. Towashiraporn, D. Stratoulias, Interpolation-based fusion of sentinel-5P, SRTM, and regulatory-grade ground stations data for producing spatially continuous maps of $PM_{2.5}$ concentrations nationwide over Thailand, Atmosphere 13 (2) (2022) 161.

[44] M. Titsias, Variational learning of inducing variables in sparse Gaussian processes, in: D. van Dyk, M. Welling (Eds.), Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics, 5, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009, pp. 567–574. https://proceedings.mlr.press/v5/titsias09a.html.

[45] S. Wallek, M. Langner, S. Schubert, C. Schneider, Modelling hourly particulate matter (PM10) concentrations at high spatial resolution in Germany using land use regression and open data, Atmosphere 13 (8) (2022) 1282.

[46] J. Li, H. Zhang, C.-Y. Chao, C.-H. Chien, C.-Y. Wu, C.H. Luo, L.-J. Chen, P. Biswas, Integrating low-cost air quality sensor networks with fixed and satellite monitoring systems to study ground-level $PM_{2.5}$, Atmos. Environ. 223 (2020) 117293. https://linkinghub.elsevier.com/retrieve/pii/S1352231020300352. https://doi.org/10.1016/j.atmosenv.2020.117293

[47] P. Schneider, N. Castell, F.R. Dauge, M. Vogt, W.A. Lahoz, A. Bartonova, A network of low-cost air quality sensors and its use for mapping urban air quality, Mob. Inf. Syst. Leverag. Volunteer. Geograph. Inf. Earth Obser. 4 (2018) 93–110.

[48] A. Criado, J.M. Armengol, H. Petetin, D. Rodriguez-Rey, J. Benavides, M. Guevara, C. Pérez García-Pando, A. Soret, O. Jorba, Data fusion uncertainty-enabled methods to map street-scale hourly NO_2 in Barcelona: a case study with CALIOPE-Urban v1.0, Geosci. Model Dev. 16 (8) (2023) 2193–2213.

[49] D. Kibirige, E. Dobos, Soil moisture estimation using citizen observatory data, microwave satellite imagery, and environmental covariates, Water 12 (8) (2020). Publisher: MDPI AG, https://doi.org/10.3390/W12082160

[50] C.-Y. Chao, H. Zhang, M. Hammer, Y. Zhan, D. Kenney, R.V. Martin, P. Biswas, Integrating fixed monitoring systems with low-cost sensors to create high-resolution air quality maps for the Northern China Plain region, ACS Earth Space Chem. 5 (11) (2021) 3022–3035.

[51] Q. Pu, E.-H. Yoo, A gap-filling hybrid approach for hourly $PM_{2.5}$ prediction at high spatial resolution from multi-sourced AOD data, Environ. Pollut. 315 (2022) 120419.

[52] T.T. Nguyen, T.D. Pham, C.T. Nguyen, J. Delfos, R. Archibald, K.B. Dang, N.B. Hoang, W. Guo, H.H. Ngo, A novel intelligence approach based active and ensemble learning for agricultural soil organic carbon prediction using multispectral and SAR data fusion, Sci. Total Environ. 804 (2022) 150187.

[53] L. Zappa, M. Woods, D. Hemment, A. Xaver, W. Dorigo, Evaluation of remotely sensed soil moisture products using crowdsourced measurements, in: K. Themistocleous, S. Michaelides, V. Ambrosia, D.G. Hadjimitsis, G. Papadavid (Eds.), Eighth International Conference on Remote Sensing and Geoinformation of the Environ-

ment (RSCy2020), SPIE, Paphos, Cyprus, 2020, p. 88. https://doi.org/10.1117/12.2571913
[54] L. Miasayedava, J. Kaugerand, J.A. Tuhtan, Lightweight assimilation of open urban ambient air quality monitoring data and numerical simulations with unknown uncertainty, Environ. Model. Assess. 28 (6) (2023) 961–975.
[55] C. Zhang, Z. Zhu, Y. Li, E. Du, Y. Sun, Z. Liu, Pollution source detection with low-cost low-accuracy sensors through coupling forward data assimilation and inverse optimization, Water Resour. Res. 60 (11) (2024) e2023WR036834.
[56] X. Wei, Q. Liu, Y. Chen, X. Lu, B. Zhao, L. Zhang, T. Liu, Y. Zheng, J. Song, et al., Evaluation of fused multisource data of air temperature based on dropsonde and satellite observation, Sci. Total Environ. 904 (2023) 166850. https://www.sciencedirect.com/science/article/pii/S004896972305475X. https://doi.org/10.1016/j.scitotenv.2023.166850
[57] S. Wang, Y. Zhang, An attention-based cnn model integrating observational and simulation data for high-resolution spatial estimation of urban air quality, Atmos. Environ. 340 (2025) 120921.
[58] D. Tang, T. Mi, X. Zheng, M. Yang, M.L. Grieneisen, Y. Zhan, F. Yang, Harmonizing low-cost and regulatory air quality monitoring networks with interpretable semi-supervised learning: reducing exposure misclassification in underrepresented communities, J. Hazard. Mater. 491 (2025) 137893.
[59] N. Castell, P. Schneider, S. Grossberndt, M.F. Fredriksen, G. Sousa-Santos, M. Vogt, A. Bartonova, Localized real-time information on outdoor air quality at kindergartens in Oslo, Norway using low-cost sensor nodes, Environ. Res. 165 (2018) 410–419.
[60] L. Johansson, A. Karppinen, M. Kurppa, A. Kousa, J.V. Niemi, J. Kukkonen, An operational urban air quality model ENFUSER, based on dispersion modelling and data assimilation, Environ. Model. Softw. 156 (2022) 105460.
[61] R. Shafran-Nathan, Y. Etzion, D.M. Broday, Fusion of land use regression modeling output and wireless distributed sensor network measurements into a high spatiotemporally-resolved NO_2 product, Environ. Pollut. 271 (2021) 116334.
[62] L.F. Weissert, K. Alberti, G. Miskell, W. Pattinson, J.A. Salmond, G. Henshaw, D.E. Williams, Low-cost sensors and microscale land use regression: data fusion to resolve air quality variations with high spatial and temporal resolution, Atmos. Environ. 213 (2019) 285–295.
[63] T. Bush, N. Papaioannou, F. Leach, F.D. Pope, A. Singh, G.N. Thomas, B. Stacey, S. Bartington, Machine learning techniques to improve the field performance of low-cost air quality sensors, Atmos. Meas. Tech. Discuss. (2021) 1–29. MAG ID: 3210162304 S2ID: 4206264971c9e9e9b1fb76e6493ff7d8e6cbc8a3. https://doi.org/10.5194/amt-2021-282
[64] N.U. Okafor, Y. Alghorani, D.T. Delaney, et al., Improving data quality of low-cost IoT sensors in environmental monitoring networks using data fusion and machine learning approach, ICT Express 6 (3) (2020) 220–228. Publisher: Elsevier BV, https://doi.org/10.1016/j.icte.2020.06.004
[65] N.U. Okafor, D.T. Delaney, Missing data imputation on IoT sensor networks: implications for on-site sensor calibration, IEEE Sens. J. 21 (20) (2021) 22833–22845.
[66] W. Wang, J. Chen, T. Hong, Occupancy prediction through machine learning and data fusion of environmental sensing and Wi-Fi sensing in buildings, Autom. Constr. 94 (2018) 233–243.
[67] J. Fu, D. Tang, M.L. Grieneisen, F. Yang, J. Yang, G. Wu, C. Wang, Y. Zhan, A machine learning-based approach for fusing measurements from standard sites, low-cost sensors, and satellite retrievals: application to No_2 pollution hotspot identification, Atmos. Environ. 302 (2023) 119756.
[68] A. Al Yammahi, Z. Aung, Forecasting the concentration of NO_2 using statistical and machine learning methods: a case study in the UAE, Heliyon 9 (2) (2023).
[69] A. Tsanousa, C. Moschou, E. Bektsis, S. Vrochidis, I. Kompatsiaris, Fusion of environmental sensors for occupancy detection in a real construction site, Sensors 23 (23) (2023) 9596.
[70] W.-C.V. Wang, S.-C.C. Lung, C.-H. Liu, et al., Application of machine learning for the in-field correction of a $PM_{2.5}$ low-cost sensor network, Sensors 20 (17) (2020) 5002. Publisher: MDPI AG, https://doi.org/10.3390/s20175002
[71] H. Shen, M. Zhou, T. Li, C. Zeng, Integration of remote sensing and social sensing data in a deep learning framework for hourly urban $PM_{2.5}$ mapping, Int. J. Environ. Res. Public Health 16 (21) (2019) 4102.
[72] X. Lai, T. Yang, Z. Wang, P. Chen, et al., IoT Implementation of Kalman filter to improve accuracy of air quality monitoring and prediction, Appl. Sci. 9 (9) (2019) 1831. https://doi.org/10.3390/app9091831
[73] G. Mani, R. Volety, A comparative analysis of LSTM and ARIMA for enhanced real-time air pollutant levels forecasting using sensor fusion with ground station data, Cogent Eng. 8 (1) (2021) 1936886.
[74] A. Forbes, K. Jagan, J. Donlevy, J.A. e Sousa, Optimization of sensor distribution using Gaussian processes, Measur. Sensors 18 (2021) 100128.
[75] K.E. Kelly, W.W. Xing, T. Sayahi, L. Mitchell, T. Becnel, P.-E. Gaillardon, M. Meyer, R.T. Whitaker, Community-based measurements reveal unseen differences during air pollution episodes, Environ. Sci. Technol. 55 (1) (2020) 120–128.
[76] F. Leibfried, V. Dutordoir, S.T. John, N. Durrande, A tutorial on sparse Gaussian processes and variational inference, 2022, arXiv:2012.13962
[77] S. Tasnim, N. Pissinou, S.S. Iyengar, K.G. Boroojeni, K. Ahmed, RCoD: reputation-based context-aware data fusion for mobile IoT, Sensors 25 (4) (2025) 1171.
[78] M. Bauer, M. van der Wilk, C.E. Rasmussen, Understanding probabilistic sparse Gaussian process approximations, 2017, arXiv:1606.04820
[79] J. Hensman, N. Fusi, N.D. Lawrence, Gaussian processes for big data, arXiv:1309.6835 (2013).
[80] C. Lin, L.D. Labzovskii, H.W. Leung Mak, J.C.H. Fung, A.K.H. Lau, S.T. Kenea, M. Bilal, J.D. Vande Hey, X. Lu, J. Ma, Observation of $PM_{2.5}$ using a combination of satellite remote sensing and low-cost sensor network in Siberian urban areas with limited reference monitoring, Atmos. Environ. 227 (2020) 117410. https://www.sciencedirect.com/science/article/pii/S1352231020301497. https://doi.org/10.1016/j.atmosenv.2020.117410
[81] M. Anachkova, S. Domazetovska, Z. Petreski, V. Gavriloski, Design of low-cost wireless noise monitoring sensor unit based on IoT concept, J. Vibroeng. 23 (4) (2021) 1056–1064.
[82] E. İçöz, F.M. Malik, K. İçöz, High spatial resolution IoT based air PM measurement system, Environ. Ecol. Stat. 28 (4) (2021) 779–792. https://doi.org/10.1007/s10651-021-00494-4
[83] L. Pradeep, S.M.S. Nagendra, Design and development of low-cost environmental sensors for urban noise measurements, in: 2023 IEEE Applied Sensing Conference (APSCON), IEEE, 2023, pp. 1–3.
[84] B. Feenstra, A. Collier-Oxandale, V. Papapostolou, D. Cocker, A. Polidori, The airsensor open-source r-package and dataviewer web application for interpreting community data collected by low-cost sensor networks, Environ. Model. Softw. 134 (2020) 104832. https://linkinghub.elsevier.com/retrieve/pii/S1364815220308896. https://doi.org/10.1016/j.envsoft.2020.104832
[85] J. Chen, J. Yang, Maximizing coverage quality with budget constrained in mobile crowd-sensing network for environmental monitoring applications, Sensors 19 (10) (2019) 2399. https://doi.org/10.3390/s19102399
[86] Q. Jiang, A.K. Bregt, L. Kooistra, Formal and informal environmental sensing data and integration potential: perceptions of citizens and experts, Sci. Total Environ. 619–620 (2018) 1133–1142.
[87] A. Kaginalkar, S. Kumar, P. Gargava, D. Niyogi, Review of urban computing in air quality management as smart city service: an integrated IoT, AI, and cloud technology perspective, Urban Clim. 39 (2021) 100972. https://doi.org/10.1016/J.UCLIM.2021.100972
[88] A. Kaginalkar, S. Kumar, P. Gargava, D. Niyogi, Stakeholder analysis for designing an urban air quality data governance ecosystem in smart cities, Urban Clim. 48 (2023) 101403. https://linkinghub.elsevier.com/retrieve/pii/S2212095522003212. https://doi.org/10.1016/j.uclim.2022.101403
[89] P. Ferrer-Cid, J.M. Barcelo-Ordinas, J. Garcia-Vidal, A. Ripoll, M. Viana, Multisensor data fusion calibration in IoT air pollution platforms, IEEE Internet Things J. 7 (4) (2020) 3124–3132. Publisher: Institute of Electrical and Electronics Engineers (IEEE), https://doi.org/10.1109/jiot.2020.2965283
[90] P. Wei, S. Hao, Y. Shi, A. Anand, Y. Wang, M. Chu, Z. Ning, Combining Google traffic map with deep learning model to predict street-level traffic-related air pollutants in a complex urban environment, Environ. Int. 191 (2024) 108992.
[91] E. Babaeian, S. Paheding, N. Siddique, V.K. Devabhaktuni, M. Tuller, Estimation of root zone soil moisture from ground and remotely sensed soil information with multisensor data fusion and automated machine learning, Remote Sens. Environ. 260 (2021) 112434.
[92] R. Huang, R. Lal, M. Qin, Y. Hu, A.G. Russell, M.T. Odman, S. Afrin, F. Garcia-Menendez, S.M. O'Neill, Application and evaluation of a low-cost PM sensor and data fusion with CMAQ simulations to quantify the impacts of prescribed burning on air quality in Southwestern Georgia, USA, J. Air Waste Manag. Assoc. 71 (7) (2021) 815–829.
[93] S.Y. Tan, M. Jacoby, H. Saha, A. Florita, G. Henze, S. Sarkar, Multimodal sensor fusion framework for residential building occupancy detection, Energy Build. 258 (2022) 111828.
[94] J. Cukjati, D. Mongus, K.R. Žalik, B. Žalik, IoT and satellite sensor data integration for assessment of environmental variables: a case study on NO_2, Sensors 22 (15) (2022) 5660. https://doi.org/10.3390/s22155660
[95] L. Zappa, M. Forkel, A. Xaver, W. Dorigo, Deriving field scale soil moisture from satellite observations and ground measurements in a hilly agricultural region, Remote Sens. 11 (22) (2019) 2596. https://doi.org/10.3390/rs11222596
[96] R. Fehri, P. Bogaert, S. Khlifi, M. Vanclooster, Data fusion of citizen-generated smartphone discharge measurements in Tunisia, J. Hydrol. 590 (2020) 125518.
[97] F. Fiebig, S. Kochanneck, I. Mauser, H. Schmeck, Detecting occupancy in smart buildings by data fusion from low-cost sensors, in: Proceedings of the Eighth International Conference on Future Energy Systems, ACM, Shatin Hong Kong, 2017, pp. 259–261. https://doi.org/10.1145/3077839.3081675
[98] M. Zumwald, B. Knüsel, D.N. Bresch, R. Knutti, Mapping urban temperature using crowd-sensing data and machine learning, Urban Clim. 35 (2021) 100739.
[99] A. Lewis, P. Edwards, Validate personal air-pollution sensors, Nature 535 (7610) (2016) 29–31.
[100] B. Maag, Z. Zhou, L. Thiele, A survey on sensor calibration in air pollution monitoring deployments, IEEE Internet Things J. 5 (6) (2018) 4857–4870.
[101] B. Agbo, H. Al-Aqrabi, R. Hill, T. Alsboui, Missing data imputation in the internet of things sensor networks, Fut. Internet 14 (5) (2022) 143.
[102] L. Liang, J. Daniels, C. Bailey, L. Hu, R. Phillips, J. South, Integrating low-cost sensor monitoring, satellite mapping, and geospatial artificial intelligence for intra-urban air pollution predictions, Environ. Pollut. 331 (2023) 121832.
[103] M. Tarazona Alvarado, J.L. Salamanca-Coy, K. Forero-Gutièrrez, L.A. Núñez, J. Pisco-Guabave, F. Escobar-Diaz, D. Sierra-Porta, Assessing and monitoring air quality in cities and urban areas with a portable, modular and low-cost sensor station: calibration challenges, Int. J. Remote Sens. 45 (17) (2024) 5713–5736.
[104] L.M. Rivera-Muñoz, J.D. Gallego-Villada, A.F. Giraldo-Forero, J.D. Martinez-Vargas, Missing data estimation in a low-cost sensor network for measuring air quality: a case study in Aburrá valley, Water Air Soil Pollut. 232 (2021) 1–15.
[105] N.U. Okafor, D.T. Delaney, Application of machine learning techniques for the calibration of low-cost IoT sensors in environmental monitoring networks, in: 2020 IEEE 6th World Forum on Internet of Things (WF-IoT), IEEE, 2020. https://doi.org/10.1109/wf-iot48130.2020.9221246
[106] E.S. Cross, L.R. Williams, D.K. Lewis, G.R. Magoon, T.B. Onasch, M.L. Kaminsky, D.R. Worsnop, J.T. Jayne, Use of electrochemical sensors for measurement of air

pollution: correcting interference response and validating measurements, Atmos. Meas. Tech. 10 (9) (2017) 3575–3588.
[107] K. Chan, D.N. Schillereff, A.C.W. Baas, M.A. Chadwick, B. Main, M. Mulligan, F.T. O'Shea, R. Pearce, T.E.L. Smith, A. Van Soesbergen, et al., Low-cost electronic sensors for environmental research: pitfalls and opportunities, Progress Phys. Geograph. Earth Environ. 45 (3) (2021) 305–338.
[108] N.V. S.R. Nalakurthi, I. Abimbola, T. Ahmed, I. Anton, K. Riaz, Q. Ibrahim, A. Banerjee, A. Tiwari, S. Gharbia, Challenges and opportunities in calibrating low-cost environmental sensors, Sensors 24 (11) (2024) 3650.
[109] N. Zimmerman, A.A. Presto, S.P.N. Kumar, J. Gu, A. Hauryliuk, E.S. Robinson, A.L. Robinson, et al., A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring, Atmos. Meas. Tech. 11 (1) (2018) 291–313.
[110] M. Zusman, C.S. Schumacher, A.J. Gassett, E.W. Spalt, E. Austin, T.V. Larson, G. Carvlin, E. Seto, J.D. Kaufman, L. Sheppard, Calibration of low-cost particulate matter sensors: model development for a multi-city epidemiological study, Environ. Int. 134 (2020) 105329. https://www.sciencedirect.com/science/article/pii/S0160412019321920. https://doi.org/10.1016/j.envint.2019.105329
[111] M.-L. Aix, S. Schmitz, D.J. Bicout, Calibration methodology of low-cost sensors for high-quality monitoring of fine particulate matter, Sci. Total Environ. 889 (2023) 164063.
[112] L. Spinelle, M. Gerboles, M.G. Villani, M. Aleixandre, F. Bonavitacola, Field calibration of a cluster of low-cost available sensors for air quality monitoring. part a: ozone and nitrogen dioxide, Sens. Actuators B Chem 215 (2015) 249–257.
[113] C. Bachechi, F. Rollo, L. Po, HypeAIR: a novel framework for real-time low-cost sensor calibration for air quality monitoring in smart cities, Ecol. Inform. 81 (2024) 102568.
[114] S. Mahajan, P. Kumar, Evaluation of low-cost sensors for quantitative personal exposure monitoring, Sustain. Cities Soc. 57 (2020) 102076.
[115] Y.-C. Hsu, P. Dille, J. Cross, B. Dias, R. Sargent, I. Nourbakhsh, Community-empowered air quality monitoring system, in: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, 2017, pp. 1607–1619.
[116] F. Concas, J. Mineraud, E. Lagerspetz, S. Varjonen, X. Liu, K. Puolamäki, P. Nurmi, S. Tarkoma, Low-cost outdoor air quality monitoring and sensor calibration: a survey and critical analysis, ACM Trans. Sensor Netw. (TOSN) 17 (2) (2021) 1–44.
[117] D. Casado-Mansilla, A. Pujante, E.I. Fernández, S. Udina, N. Castell, N. Serrano, D. López-de Ipiña, Perspective chapter: a protocol for the use of low-cost air pollution sensors in citizen science to foster evidence-based policy making, in: C.F. Bustillo-Lecompte (Ed.), Urban Pollution - Environmental Challenges in Healthy Modern Cities, IntechOpen, London, 2025. https://doi.org/10.5772/intechopen.1009024
[118] L. Cui, G. Xie, Y. Qu, L. Gao, Y. Yang, Security and privacy in smart cities: challenges and opportunities, IEEE Access 6 (2018) 46134–46145.
[119] M.A. Hoque, C. Davidson, Design and implementation of an IoT-based smart home security system, Int. J. Netw. Distrib. Comput. 7 (2) (2019) 85–92.
[120] M. Betancourt, A conceptual introduction to Hamiltonian Monte Carlo, arXiv:1701.02434 (2017).
[121] D.M. Blei, A. Kucukelbir, J.D. McAuliffe, Variational inference: a review for statisticians, J. Am. Stat. Assoc. 112 (518) (2017) 859–877.
[122] J. Salvatier, T.V. Wiecki, C. Fonnesbeck, Probabilistic programming in python using PyMC3, PeerJ Comput. Sci. 2 (2016) e55.
[123] B. Carpenter, A. Gelman, M.D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, A. Riddell, Stan: a probabilistic programming language, J. Stat. Softw. 76 (2017) 1–32.
[124] S. Salcedo-Sanz, P. Ghamisi, M. Piles, M. Werner, L. Cuadra, A. Moreno-Martínez, E. Izquierdo-Verdiguier, J. Muñoz-Marí, A. Mosavi, G. Camps-Valls, et al., Machine learning information fusion in earth observation: a comprehensive review of methods, applications and data sources, Inf. Fusion 63 (2020) 256–272. https://doi.org/10.1016/j.inffus.2020.07.004
[125] L. Luo, Y. Zhang, B. Pearson, Z. Ling, H. Yu, X. Fu, On the security and data integrity of low-cost sensor networks for air quality monitoring, Sensors 18 (12) (2018) 4451.
[126] L. Wang, G. Qin, D. Yang, X. Han, X. Ma, Geographic differential privacy for mobile crowd coverage maximization, in: Proceedings of the AAAI Conference on Artificial Intelligence, 32, 2018.
[127] P. Zhang, X. Cheng, S. Su, N. Wang, Task allocation under geo-indistinguishability via group-based noise addition, IEEE Trans. Big Data 9 (3) (2022) 860–877.
[128] H. Shen, G. Bai, Y. Hu, T. Wang, P2TA: privacy-preserving task allocation for edge computing enhanced mobile crowdsensing, J. Syst. Archit. 97 (2019) 130–141.
[129] J. Feng, Y. Wu, H. Sun, S. Zhang, D. Liu, Panther: practical secure 2-party neural network inference, IEEE Trans. Inf. Forensics Secur. 20 (2025) 1149–1162. https://doi.org/10.1109/TIFS.2025.3526063
[130] J. Lin, Y. Miao, L. Wei, T. Leng, K.-K.R. Choo, Efficient secure inference scheme in multiparty settings for industrial internet of things, IEEE Trans. Ind. Inf. 20 (10) (2024) 11877–11886.
[131] K. Gopalakrishnan, Security vulnerabilities and issues of traditional wireless sensors networks in IoT, Principles of Internet of Things (IoT) Ecosystem: Insight Paradigm (2020) 519–549.
[132] I. Butun, P. Österberg, H. Song, Security of the internet of things: vulnerabilities, attacks, and countermeasures, IEEE Commun. Surv. Tutor. 22 (1) (2019) 616–644.
[133] C.K.I. Williams, C.E. Rasmussen, Gaussian Processes for Machine Learning, MIT press Cambridge, MA, 2006.
[134] W.A. Abdulhafiz, A. Khamis, Handling data uncertainty and inconsistency using multisensor data fusion, Adv. Artif. Intell. 2013 (1) (2013) 241260.
[135] M. Adoui, T. Herpoel, B. Frénay, Constrained tiny machine learning for predicting gas concentration with I4.0 low-cost sensors, ACM Trans. Embed. Comput. Syst., 23 (3) 2023 1–23. S2ID: c35af67d30025ca20b6e4cf901a39bb2562dc95f
[136] X. Ma, B. Zou, J. Deng, J. Gao, I. Longley, S. Xiao, B. Guo, Y. Wu, T. Xu, X. Xu, et al., A comprehensive review of the development of land use regression approaches for modeling spatiotemporal variations of ambient air pollution: a perspective from 2011 to 2023, Environ. Int. 183 (2024) 108430.
[137] X. Zou, Y. Yan, X. Hao, Y. Hu, H. Wen, E. Liu, J. Zhang, Y. Li, T. Li, Y. Zheng, Y. Liang, Deep learning for cross-domain data fusion in urban computing: taxonomy, advances, and outlook, 2024, arXiv:2402.19348.
[138] N. Bebelaar, R.C. Braggaar, C.M. Kleijwegt, R.W.E. Meulmeester, G. Michailidou, N. Salheb, S. van der Spek, N. Vaissier, E. Verbree, Monitoring urban environmental phenomena through a wireless distributed sensor network, Smart Sustain. Built Environ. 7 (1) (2018) 68–79.
[139] K. Huang, J. Bi, X. Meng, G. Geng, A. Lyapustin, K.J. Lane, D. Gu, P.L. Kinney, Y. Liu, Estimating daily $PM_{2.5}$ concentrations in New York City at the neighborhood-scale: implications for integrating non-regulatory measurements, Sci. Total Environ. 697 (2019) 134094. https://www.sciencedirect.com/science/article/pii/S0048969719340719. https://doi.org/10.1016/j.scitotenv.2019.134094
[140] A. Xaver, L. Zappa, G. Rab, I. Pfeil, M. Vreugdenhil, D. Hemment, W.A. Dorigo, Evaluating the suitability of the consumer low-cost parrot flower power soil moisture sensor for scientific environmental applications, Geosci. Instrum. Methods Data Syst. 9 (1) (2020) 117–139.
[141] E. Vidaña-Vila, J. Navarro, C. Borda-Fortuny, D. Stowell, R.M. Alsina-Pagés, Low-cost distributed acoustic sensor network for real-time urban sound monitoring, Electronics 9 (12) (2020) 2119. https://doi.org/10.3390/electronics9122119
[142] R. Carbajales, M. Iurcev, P. Diviacco, Low cost sensors and crowd-sourced data to map air pollution in urban areas, in: EGU General Assembly Conference Abstracts, 2020, p. 18946.
[143] R. Novak, I. Petridis, D. Kocman, J.A. Robinson, T. Kanduč, D. Chapizanis, S. Karakitsios, B. Flückiger, D. Vienneau, O. Mikeš, et al., Harmonization and visualization of data from a transnational multi-sensor personal exposure campaign, Int. J. Environ. Res. Public Health 18 (21) (2021) 11614.
[144] M.A. Becerra, Y. Uribe, D.H. Peluffo-Ordóñez, K.C. Álvarez-Uribe, C. Tobón, Information fusion and information quality assessment for environmental forecasting, Urban Clim. 39 (2021) 100960.
[145] Y.M. Idir, O. Orfila, V. Judalet, B. Sagot, P. Chatellier, Mapping urban air quality from mobile sensors using spatio-temporal geostatistics, Sensors 21 (14) (2021) 4717.
[146] D. Kibirige, E. Dobos, Estimation of surface soil moisture by integrating environmental data and remote-sensing satellites, Multidiszciplináris Tudományok 11 (1) (2021) 22–37.
[147] T. Veiga, A. Munch-Ellingsen, C. Papastergiopoulos, D. Tzovaras, I. Kalamaras, K. Bach, K. Votis, S. Akselsen, From a low-cost air quality sensor network to decision support services: steps towards data calibration and service development, Sensors 21 (9) (2021) 3190.
[148] P.-C. Chen, Y.-T. Lin, Exposure assessment of $PM_{2.5}$ using smart spatial interpolation on regulatory air quality stations with clustering of densely-deployed microsensors, Environ. Pollut. 292 (2022) 118401.
[149] C. Briciu-Burghina, J. Zhou, M.I. Ali, F. Regan, Demonstrating the potential of a low-cost soil moisture sensor network, Sensors 22 (3) (2022) 987.
[150] C. Corbari, N. Paciolla, I. Ben Charfi, D. Skokovic, J.A. Sobrino, M. Woods, Citizen science supporting agricultural monitoring with hundreds of low-cost sensors in comparison to remote sensing data, Eur. J. Remote Sens. 55 (1) (2022) 388–408.
[151] L. Miasayedava, J. Kaugerand, J.A. Tuhtan, Lightweight open data assimilation of pan-European urban air quality, IEEE Access 11 (2023) 84670–84688. https://doi.org/10.1109/ACCESS.2023.3302348
[152] Y. Wu, Z. Yang, Y. Liu, Internet-of-things-based multiple-sensor monitoring system for soil information diagnosis using a smartphone, Micromachines 14 (7) (2023) 1395.
[153] S. Zhu, J. Tang, X. Zhou, P. Li, Z. Liu, C. Zhang, Z. Zou, T. Li, C. Peng, Research progress, challenges, and prospects of $PM_{2.5}$ concentration estimation using satellite data, Environ. Rev. 31 (4) (2023) 605–631.
[154] H. Fritz, C. Wu, A. Novoselac, K. Kinney, Z. Nagy, Information fusion of stationary, mobile, and wearable consumer-grade sensors to confidently estimate bedroom ventilation rates, Build. Environ. 230 (2023) 109997.
[155] N. Okafor, R. Ingle, U. Matthew, M. Saunders, D. Delaney, Assessing and improving IoT sensor data quality in environmental monitoring networks: a focus on peatlands, IEEE Internet Things J. 11 (24) (2024) 40727–40742. https://doi.org/10.1109/JIOT.2024.3454241
[156] M.I. Rodríguez-García, M.G. Carrasco-García, M.d. C.R. Ribeiro, J. González-Enrique, J.J. Ruiz-Aguilar, I.J. Turias, Air pollution PM_{10} forecasting maps in the maritime area of the Bay of Algeciras (Spain), J. Mar. Sci. Eng. 12 (3) (2024) 397.
[157] S. De Vito, A. Del Giudice, G. D'Elia, E. Esposito, G. Fattoruso, S. Ferlito, F. Formisano, G. Loffredo, E. Massera, P. D'Auria, et al., Future low-cost urban air quality monitoring networks: insights from the EU's airheritage project, Atmosphere 15 (11) (2024) 1351.
[158] J. Song, X. Shi, H. Wang, X. Lv, W. Zhang, J. Wang, T. Li, W. Li, Combination of feature selection and geographical stratification increases the soil total nitrogen estimation accuracy based on vis-NIR and pXRF spectral fusion, Comput. Electron. Agric. 218 (2024) 108636.
[159] B. Tang, C.O. Stanier, G.R. Carmichael, M. Gao, Ozone, nitrogen dioxide, and $PM_{2.5}$ estimation from observation-model machine learning fusion over S. Korea: influence of observation density, chemical transport model resolution, and geostationary remotely sensed AOD, Atmos. Environ. 331 (2024) 120603.
[160] A. Das, N. Singh, S. Chakraborty, UniPreCIS: a data preprocessing solution for collocated services on shared IoT, Fut. Gen. Comput. Syst. 153 (2024) 543–557.

[161] J.-C. Gamazo-Real, R.T. Fernández, A.M. Armas, Comparison of edge computing methods in internet of things architectures for efficient estimation of indoor environmental parameters with machine learning, Eng. Appl. Artif. Intell. 126 (2023) 107149.
[162] R. Hoogerbrugge, S. van Ratingen, K. Siteur, J. Wesseling, Optimal measurement strategy for air quality combining official and low-cost measurements, Atmos. Environ. 343 (2025) 120990.
[163] H.-W. Chen, C.-Y. Chen, Y.-H. Chuang, G.-Y. Lin, Characterization of spatial distribution and source contribution of acidic/basic aerosols in microenvironment using expected value theory, Earth Syst. Environ. (2025). https://doi.org/10.1007/s41748-025-00590-6
[164] B. Choi, M.A. Hummel, Spatiotemporal air quality prediction using stochastic advection–diffusion model for multimodal data fusion, Environ. Res. Lett. 20 (1) (2025) 014065.
[165] A. Ganji, M. Lloyd, J. Xu, S. Weichenthal, M. Hatzopoulou, Traffic-related air pollution backcasting using convolutional neural network and long short-term memory approach, Sci. Total Environ. 976 (2025) 179286.
[166] S. Shetty, P.D. Hamer, K. Stebel, A. Kylling, A. Hassani, T.K. Berntsen, P. Schneider, Daily high-resolution surface $PM_{2.5}$ estimation over Europe by ML-based downscaling of the CAMS regional forecast, Environ. Res. 264 (2025) 120363.