Big Earth Data

ISSN: 2096-4471 (Print) 2574-5417 (Online) Journal homepage: www.tandfonline.com/journals/tbed20

Quality assessment and control of urban environmental sensors using physical thresholding and machine learning-based probabilities

Jangho Lee, Max Berkelhammer, Anna E. S. Vincent, Maxwell Grover, Ahram Cho, Aaron I. Packman, Bilal Kaludi, Miquel Gonzalez-Meler & Gavin McNicol

To cite this article: Jangho Lee, Max Berkelhammer, Anna E. S. Vincent, Maxwell Grover, Ahram Cho, Aaron I. Packman, Bilal Kaludi, Miquel Gonzalez-Meler & Gavin McNicol (12 Mar 2026): Quality assessment and control of urban environmental sensors using physical thresholding and machine learning-based probabilities, Big Earth Data, DOI: 10.1080/20964471.2026.2623720

To link to this article: https://doi.org/10.1080/20964471.2026.2623720

© 2026 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the International Research Center of Big Data for Sustainable Development Goals.

Published online: 12 Mar 2026.

Submit your article to this journal

Article views: 20

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=tbed20

BIG EARTH DATA
https://doi.org/10.1080/20964471.2026.2623720

RESEARCH ARTICLE

OPEN ACCESS

Check for updates

Quality assessment and control of urban environmental sensors using physical thresholding and machine learning-based probabilities

Jangho Lee[a], Max Berkelhammer[a], Anna E. S. Vincent[b,c], Maxwell Grover[d], Ahram Cho[e], Aaron I. Packman[b,c], Bilal Kaludi[a], Miquel Gonzalez-Meler[a,e] and Gavin McNicol[a]

[a]Department of Earth and Environmental Sciences, University of Illinois Chicago, Chicago, IL, USA; [b]Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA; [c]Northwestern Center for Water Research, Northwestern University, Evanston, IL, USA; [d]Environmental Research Division, Argonne National Laboratory, Lemont, IL, USA; [e]Department of Biological Sciences, University of Illinois Chicago, Chicago, IL, USA

ABSTRACT

Reliable environmental sensor data are fundamental for accurate urban climate modeling and evidence-based planning. Conventional physics-based quality control (QC) methods apply fixed thresholds to flag physically implausible values, but they often fail to detect subtle, context-dependent anomalies. This study introduces a hybrid QC framework that integrates conservative physical constraints with a probabilistic machine-learning approach based on Positive-Unlabeled XGBoost (PU-XGBoost). Using data from the CROCUS Urban Integrated Field Laboratory in Chicago, the framework generates anomaly likelihood probabilities rather than binary flags, allowing confidence-weighted data evaluation. The results demonstrate that the hybrid method effectively captures both gross and latent sensor errors overlooked by rule-based QC, while maintaining interpretability through physically informed features. Feature importance analysis highlights the dominant roles of temporal statistics, sensor type, and environmental context in anomaly detection. Overall, the proposed hybrid framework provides a scalable and interpretable foundation for self-adaptive quality assurance in next-generation urban environmental sensing networks.

ARTICLE HISTORY
Received 6 August 2025
Accepted 12 January 2026

KEYWORDS
Urban sensor; urban climate; environmental sensor; QA/QC; soil sensor; urban meteorology

1. Introduction

In modern urban environments, large networks of sensors continuously monitor meteorological and environmental conditions such as air temperature, humidity, wind speed, precipitation, and air quality. Soil sensors are increasingly included to capture below-ground variability, which is essential for understanding spatial

CONTACT Jangho Lee jholee@uic.edu Department of Earth and Environmental Sciences, University of Illinois Chicago, 845 W Taylor St, Chicago, IL 60607, USA

© 2026 The Author(s). Published by Taylor & Francis Group and Science Press on behalf of the International Society for Digital Earth, supported by the International Research Center of Big Data for Sustainable Development Goals.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

heterogeneity in urban ecosystems (Wienhold et al., 2024). These measurements form a foundational resource for informed urban greening and planning decisions and for driving high-resolution urban climate models (Chen et al., 2011; Masson et al., 2020; Pan et al., 2015). City planners and infrastructure designers rely on accurate, high-quality sensor data to capture micro-scale patterns and to design effective mitigation strategies such as enhancing canopy cover, optimizing green infrastructure, and improving ventilation to reduce urban heat (Jha et al., 2015; Lee & Berkelhammer, 2024; Muller et al., 2013; Rashid & Rehmani, 2016; Yang et al., 2023). However, the value of such dense observations is directly tied to their reliability: erroneous sensor readings can propagate through analysis and models, potentially leading to biased conclusions or suboptimal policy interventions. Therefore, ensuring the trustworthiness of urban sensor data is a critical prerequisite for any data-driven planning or climate modeling endeavors (Ching et al., 2018; Masson et al., 2020; Massoud et al., 2023).

Despite the importance and need for urban sensor networks, maintaining sensor reliability in complex urban settings is challenging (Hill, 2015; Lin & Hubbard, 2004). Urban sensors are often deployed in harsh, heterogeneous environments, which are exposed to weather extremes, heavy air pollution, and physical disturbances like vibrations from traffic, vandalism, or construction. Connectivity and power issues can further compromise data continuity and quality. Moreover, the urban landscape itself introduces significant variability: measurements can legitimately differ greatly between adjacent street canyons, open parks, and industrial and residential areas, or exhibit sudden shifts due to localized events. This variability complicates the task of distinguishing true environmental signals from sensor malfunctions. As a result, raw data collected from city sensor networks may contain a wide array of anomalies—outliers, biases, and drifted baselines—that must be identified and addressed to prevent degraded data from skewing any downstream analysis (Beele et al., 2022; Chen et al., 2021; Duarte Rocha et al., 2021; Peters et al., 2021; Sakthivel & Sengupta, 2025).

Quality control (QC) of environmental sensor data has traditionally relied on straightforward rule-based techniques, among which physical thresholding is a primary tool (Feng et al., 2025; Sun et al., 2015; Yang et al., 2021). In a threshold-based anomaly detection approach, any sensor reading that falls outside predetermined plausible bounds, or changes too abruptly beyond expected norms, is flagged as potential error. For example, a temperature measurement far above the highest recorded value for the city, or a sudden physically implausible jump in humidity, would be automatically marked as erroneous by such a rule (Kabir et al., 2022; Liu et al., 2018; Yang et al., 2008). This technique leverages domain expertise—encapsulating fundamental physical constraints and reasonable expectations for the measurements—and it is straightforward to implement in practice.

However, static threshold rules exhibit significant limitations when applied to urban datasets. Fixed thresholds or relationships cannot accommodate context or gradual shifts (von Arx et al., 2013); a rigid cutoff may misidentify a legitimate but rare environmental extreme as errors, simply because it falls outside an expected range, while a subtle sensor fault (for instance, a slow calibration drift) that keeps readings within nominal bounds can go undetected. Moreover, threshold-based decisions provide only binary outcomes with no gradation of confidence or severity. This all-or-nothing approach can lead to unnecessary data loss and fails to capture the uncertainty associated with sensor measurements, making it insufficient for reliably managing the full spectrum of data quality issues in an urban sensor network.

To address these shortcomings, more advanced anomaly detection or error detection techniques based on machine learning have been explored in recent years (Daurenbayeva et al., 2025; Fazai et al., 2018; Santoso et al., 2023; Warriach & Tei, 2013; Wellyantama & Soekirno, 2021). Machine learning models can learn complex patterns from historical sensor data and automatically detect observations that deviate from expected behavior across time and space. Unlike simple thresholds, these models can consider multiple variables and relationships simultaneously —capturing temporal trends, spatial correlations among neighboring sensors, and multivariate anomalies that would be difficult to define with manual rules. For instance, a supervised classifier, such as XGBoost (Chen et al., 2024; Henriques et al., 2020) or isolation forests (Lesouple et al., 2021; Xu et al., 2017), can be trained on past examples of faulty versus healthy sensor readings to recognize subtle signatures of sensor malfunctions, such as the gradual drift of a temperature sensor or the intermittent noise in an air quality sensor. In the context of urban environmental monitoring, such data-driven approaches have demonstrated the potential to identify anomalies that simple thresholds would miss, thereby improving detection sensitivity.

Motivated by this need for synergy, this study proposes an approach that integrates domain knowledge with machine learning in a unified probabilistic framework for sensor data quality assessment. In our framework, we apply basic physical limits as feature inputs and to set reasonable bounds, while the machine learning component evaluates each data point using both the raw observation and these domain-informed features. Importantly, the outcome of this hybrid analysis is not a rigid pass/fail judgment but a probabilistic confidence score indicating the likelihood that a given sensor reading is valid.

A key advantage of expressing data quality in probabilistic terms is the improved utility of sensor information in downstream modeling and decision-making processes. In an urban climate modeling context, probabilistic quality scores allow models to ingest sensor observations together with their associated confidence levels. Rather than outright discarding a measurement flagged as suspect, a model can use the quality probability as a weight—diminishing the influence of less reliable data points without entirely omitting them. In this study, the proposed framework is demonstrated using data from an urban sensor network in Chicago, showing that it can more effectively identify and quantify anomalies without discarding useful data.

The remainder of this paper is organized as follows. Section 2 describes the data sources and sensor network configuration. Section 3 outlines the physics-based and machine learning-based quality control methods. Section 4 presents the results and comparative analyses. Section 5 summarizes and discusses the key findings and their implications as well as limitation and future directions. Section 5 concludes the paper.

2. Data

2.1. *Waggle and MFR nodes*

CROCUS is an Urban Integrated Field Laboratory located in the Chicago metropolitan area. The CROCUS project's sensing network in Chicago consists of a distributed array of instrument nodes designed for comprehensive urban environmental monitoring (O'Brien, Tuftedal, Gala, et al., 2024b; O'Brien, Tuftedal, Wawrzyniak, et al., 2024; Pal et al., 2024). Data collection is ongoing as part of the continuous CROCUS Urban Integrated Field Laboratory operation. However, the data collection reported in this study only spans from October 2024 to April 2025, although individual sensor installation dates vary as early as May 2022. The network consists of 13 distinct sites (Figure 1), each equipped with a primary Waggle node. A Waggle node is an integrated sensor platform with on-board edge computing, developed to support AI-driven data processing in situ (Balaprakash et al., 2021). Each Waggle node hosts a suite of meteorological and atmospheric sensors, typically mounted on building rooftops or towers in the study area, with additional distributed sensors transmitting to the node via LoRaWAN. These sensors measure standard meteorological variables (air temperature, relative humidity, barometric pressure, wind speed, and direction) as well as precipitation (rainfall rate and accumulation). Waggle nodes also carry air quality sensors, as summarized in Table 1.

All sensor readings on a Waggle node are time-synchronized and processed locally by the node's computing unit, enabling initial quality checks and data reduction before transmission. Each node operates continuously, transmitting data in near-real-time via

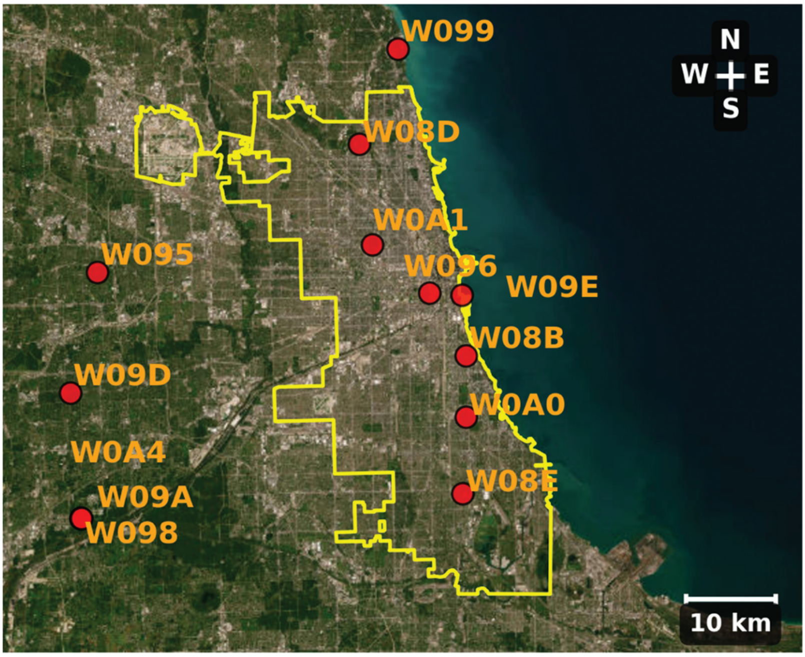

Figure 1. Locations and IDs of the 13 Waggle nodes across the Chicago metropolitan area. The map displays the city boundaries (yellow outline), with a scale bar and compass-based cardinal directions provided for spatial reference.

Table 1. Summary of sensor types and the environmental variables each sensor measure at Waggle and MFR nodes.

Sensor	Measuring Variable
BME280	Barometric pressure, relative humidity, air temperature (core)
BME680	Barometric pressure, relative humidity, air temperature (shield)
AQT530	Concentration of atmospheric pollutant & particulate matter (CO, NO, NO_2, O_3, PM1, PM2.5, PM10), barometric pressure, humidity, air temperature
WXT536	Barometric pressure, humidity, air temperature, rain accumulation, wind direction, wind speed
MFR Nodes	Air temperature, heat flux, shortwave radiation (in, out, and net), longwave radiation (in, out, and net), soil temperature (at 15 cm, 30 cm, 45 cm, and 60 cm depth), soil volumetric water content (Soil VWC, at 15 cm, 30 cm, 45 cm, and 60 cm depth), vapor pressure deficit (VPD), water conductivity, water depth, water temperature

a cloud-connected platform, yielding high-resolution insight into urban microclimates. Data from the Waggle nodes are typically recorded at fine temporal granularity (on the order of sub-seconds to minutes for most variables), which can be aggregated or summarized over longer intervals (e.g., generating hourly means or totals).

In parallel with the rooftop or elevated atmospheric sensors (Waggle), Mobile Field Research (MFR) nodes are deployed at ground level to monitor subsurface and near-surface conditions, primarily with soil and hydrologic sensors (Tables 1 and 2). MFR nodes also record near-surface atmospheric data to complement the Waggle observations. The two systems operate at the same 13 observation sites across the CROCUS network (Figure 1), but their vertical placement differs systematically. Waggle nodes are typically installed on building rooftops or elevated platforms to minimize obstruction for transmission and optimize atmospheric and air-quality measurements, while the corresponding MFR nodes are positioned at ground level within the same property boundary— usually within tens of meters of the Waggle node’s footprint. This paired but vertically separated configuration enables concurrent observations of atmospheric and subsurface conditions, maintaining spatial correspondence between the two data streams. Figure 2 illustrates these representative configurations, showing the typical rooftop—ground deployment structure used across the network.

Data communication from MFR nodes occurs via wireless (LoRaWAN) links through the nearby Waggle computing unit. Each sensor has a programmable sampling frequency,

Table 2. List of Waggle node sites and location with installed sensor types and number of associated MFR nodes (if applicable).

Waggle Node	Lat	Lon	Sensors and MFR Nodes
W0A4	41.7014	−87.9952	BME280, BME680, AQT530, WXT536
W0A1	41.9055	−87.7033	BME280, BME680, AQT530, WXT536
W0A0	41.7770	−87.6097	BME280, BME680, AQT530, WXT536, MFR (3)
W09E	41.8681	−87.6133	BME280, BME680, AQT530, WXT536
W09D	41.7952	−88.0061	BME280, BME680, AQT530, WXT536
W09A	42.0514	−87.6776	BME280, BME680, AQT530, WXT536
W099	42.0514	−87.6776	BME280, BME680, AQT530, WXT536, MFR (1)
W098	41.7013	−87.9948	BME280, BME680
W096	41.8694	−87.6458	BME280, BME680, AQT530, WXT536, MFR (2)
W095	41.8848	−87.9787	BME280, BME680
W08E	41.7198	−87.6128	BME280, BME680, AQT530, WXT536, MFR (2)
W08D	41.9805	−87.7166	BME280, BME680, AQT530, WXT536, MFR (2)
W08B	41.8229	−87.6096	BME280, BME680, AQT530, WXT536

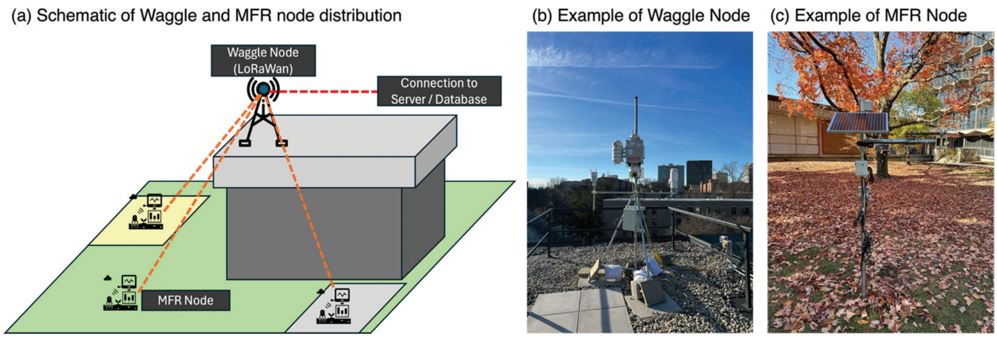
(a) Schematic of Waggle and MFR node distribution
Waggle Node (LoRaWan)
Connection to Server / Database
MFR Node
(b) Example of Waggle Node
(c) Example of MFR Node

Figure 2. Representative configuration of Waggle and MFR nodes at CROCUS observation sites. (a) Schematic illustration showing the typical spatial arrangement of atmospheric Waggle nodes (often rooftop or elevated installations) and nearby MFR nodes measuring subsurface variables at ground level within the same property boundary. (b) Rooftop Waggle node installed on the Northwestern University site (W099). (c) Ground-level MFR node deployed in Northeastern Illinois University site (W08D).

and in this study all data were aggregated into 30-minute intervals for consistency. Only variables with continuous and reliable coverage across multiple nodes were analyzed, including air temperature, relative humidity, barometric pressure, wind speed, vapor pressure deficit, soil temperature, and volumetric water content. Variables such as shortwave and longwave radiation were excluded due to limited temporal coverage and inconsistent calibration across nodes. The resulting final dataset contains roughly 10,000 observational time series of 30-minute intervals and 363 variable measurements, yielding a total of over 3.6 million individual records used for model development and evaluation. Table 3 describes the descriptive statistics for the sensor-variable combinations used in this study.

2.2. Sensor types and functional differences

The CROCUS network integrates multiple commercial environmental sensors that differ in measurement principles, internal architecture, and environmental sensitivities. Understanding these distinctions is important for interpreting potential differences in data quality and error behavior among sensor types.

Among the Bosch environmental sensors used in the network, both the BME280 and BME680 measure air temperature, relative humidity, and barometric pressure. However, the BME680 includes an additional metal-oxide gas sensing element designed to detect volatile organic compounds (VOCs) and other trace gases. This component enhances sensitivity to ambient air chemistry but also increases susceptibility to drift or contamination under high-VOC or polluted conditions. In contrast, the BME280 lacks this gas-sensing module and thus tends to exhibit greater stability under chemically variable environments, although both sensors perform similarly for temperature, humidity, and pressure measurements. The inclusion of the gas channel in the BME680 also increases power consumption and thermal load, which can influence long-term stability within enclosed housings.

Table 3. Descriptive statistics of the raw sensor readings used in this study.

Sensor	Variable	Unit	Min	Max	Mean	Median	Std
AQT	CO	ppm	0.00	6.61	0.11	0.08	0.14
AQT	Relative humidity	%	17.47	100.00	63.14	62.89	17.47
AQT	NO	ppm	0.00	1.62	0.01	0.00	0.01
AQT	NO_2	ppm	0.00	6.33	0.02	0.01	0.02
AQT	O_3	ppm	0.00	2.16	0.02	0.02	0.01
AQT	PM1	μg/m^3	0.00	1,323.83	10.25	5.53	19.42
AQT	PM10	μg/m^3	0.00	2,934.14	23.05	12.68	57.74
AQT	PM2.5	μg/m^3	0.00	2,244.30	13.47	6.87	30.09
AQT	Barometric pressure	hPa	938.75	1,018.25	992.48	993.23	8.09
AQT	Air temperature	°C	−22.69	30.18	7.04	6.66	7.86
BME280	Barometric pressure	Pa	95,570.67	101,812.16	99,257.74	99,331.66	837.25
BME280	Relative humidity	%	0.03	69.96	16.76	16.62	4.40
BME280	Air temperature	°C	−1.76	62.06	28.75	28.36	8.91
BME680	Barometric pressure	Pa	57,268.58	106,269.52	97,933.16	99,512.35	3,266.96
BME680	Relative humidity	%	0.00	99.90	60.56	61.95	18.51
BME680	Air temperature	°C	−21.03	33.91	8.87	8.50	7.77
MFR	Air temperature	°C	−21.57	31.78	5.56	5.47	7.96
MFR	Soil temperature (15 cm)	°C	−8.19	23.88	6.67	5.82	4.99
MFR	Soil temperature (30 cm)	°C	−2.69	21.44	7.02	6.22	4.55
MFR	Soil temperature (45 cm)	°C	−0.09	20.75	7.43	6.62	4.21
MFR	Soil temperature (60 cm)	°C	0.00	20.88	7.81	6.95	3.98
MFR	Vapor pressure deficit	kPa	0.00	3.66	0.35	0.22	0.36
MFR	Soil volumetric water content (15 cm)	%	2.00	43.76	26.52	27.30	4.02
MFR	Soil volumetric water content (30 cm)	%	5.27	48.26	29.18	28.97	3.67
MFR	Soil volumetric water content (45 cm)	%	15.75	42.29	27.72	27.73	1.43
MFR	Soil volumetric water content (60 cm)	%	11.19	38.22	25.80	25.92	1.32
MFR	Water conductivity	μS/cm	0.00	3,409.00	775.74	519.50	357.15
MFR	Water depth	mm	−11.00	5,507.00	1,964.65	1,877.33	267.05
MFR	Water temperature	°C	−0.60	15.60	11.69	11.62	1.20
WXT	Relative humidity	%	19.09	100.00	64.38	64.09	17.35
WXT	Barometric Pressure	hPa	955.22	1,016.95	991.32	992.10	8.22
WXT	Air Temperature	°C	−22.99	30.02	6.51	6.14	7.84
WXT	Rain accumulation	mm	0.00	17.54	0.06	0.00	0.45
WXT	Wind direction	degrees	14.00	351.40	188.88	203.34	71.34
WXT	Wind speed	m/s	0.11	14.14	2.29	2.15	1.23

For the Vaisala instruments, the AQT530 and WXT536 are designed for complementary measurement domains. The AQT530 functions primarily as an air-quality transmitter that combines electrochemical gas sensors with an optical particle counter and basic meteorological sensors, enabling simultaneous observation of gaseous and particulate pollutants. This configuration allows for rich chemical characterization of the urban atmosphere but also makes the sensor more sensitive to environmental contamination, humidity fluctuations, and calibration drift. The WXT536, in contrast, is a multi-parameter weather sensor measuring temperature, humidity, barometric pressure, wind speed and direction, and precipitation using ultrasonic anemometry and impact-based rainfall detection. Because it relies on physical rather than chemical sensing components, the WXT536 is less affected by pollutant exposure but may exhibit transient errors during strong winds, heavy rain, or condensation on transducer surfaces.

These differences in sensing mechanisms and environmental exposure help explain the distinct error signatures observed across sensor models in later sections. Recognizing these contrasts provides context for interpreting both the physics-based and machine-learning-based quality-control results presented in this study.

2.3. Sources of sensor errors

Operating a sensor network in an urban environment presents numerous data quality challenges. The raw data from the Waggle and MFR nodes often contain anomalies or errors arising from both environmental influences and instrument system limitations. The common sensor error types observed in the CROCUS urban deployment include spikes, clipping (saturation), stuck values, and consistent offset. Each of these error modes can be linked to specific causes, whether external (physical environmental factors) or internal (hardware, software, or network issues), and often they are associated with particular sensor types or conditions.

Transient spikes or outliers are brief, anomalous readings that deviate sharply from the expected range, often only for one or a few sampling intervals (Fang & Bate, 2017). In many cases, spikes are not genuine environmental events but instead reflect interference or momentary malfunctions. Spikes may also occur due to power supply fluctuations, transient electrical grounding issues, or digital communication errors (especially for MFR nodes as they communicate through wireless channels).

Clipping (saturation) errors occur when a sensor's output reaches a design or physical limit and cannot record higher (or lower) values beyond that point. In the dataset, a clipped signal appears as a flattening or plateau at a maximum or minimum reading. Clipping results when environmental conditions exceed physical limitations in the sensor, physical limitations in the response of the data-acquisition system, or programmed limitations based on either the design sensor measurement range or rules based on the signal-to-noise ratio. Data quality control for clipping involves flagging any prolonged periods at exact sensor limits.

Stuck values are error types where a sensor's output remains artificially constant (or jittering values) over a time when it should naturally vary. In dynamic urban conditions, most variables fluctuate at sampling intervals from hours to minutes. Thus, a completely flat time series for an extended duration often signals a problem. One cause of stuck values is sensor or logger failure, wherein the device stops updating but continues reporting the last known value. Another cause is software: some systems fill in missing data with the previous value or a default flag, which can appear as a constant reading if not clearly marked (Fang & Bate, 2017; Houston et al., 2019; Lugomer et al., 2017; Zhao & Zhao, 2023). Stuck value errors are typically detected by simple persistence checks: if a normally varying parameter does not change at all over a suspiciously long period, it should be flagged.

Sensor drift represents a subtle, yet significant type of sensor error characterized by a gradual shift or bias in sensor measurements away from their true values over extended periods. Such drift, if unaddressed, can result in persistent over- or underestimation of critical environmental variables, thereby undermining the integrity of data analyses. Commonly, sensor drift is driven by sensor aging processes, prolonged environmental exposure, and gradual deterioration of internal sensor components or calibration reference standards. Persistent high offset errors are characterized by prolonged periods during which sensor values consistently remain elevated relative to a typical baseline. Unlike transient spikes, these persistent deviations exceed normal variations over extended durations, suggesting sensor calibration problems, systematic bias, or local environmental anomalies such as sensor contamination, shielding damage, or obstructive elements.

3. Methods

3.1. Physics-based error detection

Physics-based error detection provides an initial, conservative assessment of data quality by leveraging explicitly defined physical constraints and empirical thresholds to identify sensor anomalies. Each measurement variable is systematically evaluated against a set of predefined criteria informed by realistic environmental conditions, inherent sensor limitations, and typical variability (Table 4). While most thresholds reflect plausible environmental ranges, some limits originate directly from sensor specifications. For example, the upper limit for particulate matter concentration (2000 µg m^{-3}) reflects the saturation range of the AQT530 sensor, and the maximum rainfall accumulation limit (200 mm per 30 min) corresponds to the acoustic detection constraint of the WXT536 sensor. These sensor-specific boundaries prevent misinterpretation of saturated signals as valid environmental observations.

The physics-based QC framework applies the following eight criteria, designed to capture distinct types of measurement errors (also in Table 4):

Table 4. Summary of physical-based QC criteria and thresholds.

Criterion	Variable	Thresholds and Conditions
Physical Range Limits	Air temperature	−40 °C to 60 °C
	Soil temperature	−20 °C to 60 °C
	Relative humidity	0% to 100%
	Barometric pressure	795 to 1,105 hPa
	Wind speed	0 to 50 m/s
	Wind direction	0° to 360°
	Rain accumulation	0 to 200 mm/30 min
	CO	0 to 150 ppm
	Ozone	0 to 0.5 ppm
	NO, NO$_2$	0 to 12 ppm
	PM1, PM2.5, PM10	0 to 2,000 µg/m^3
	Soil volumetric water content	0% to 100%
	Vapor pressure deficit	0 to 10 kPa
Step Spike	Air temperature	>5 °C change between consecutive measurements
	Relative humidity	>25% change
	Barometric pressure	>10 hPa change
	Wind speed	>15 m/s change
	Soil volumetric water content	>10% change
Flat-Line (24 h)	Air temperature	<2 °C range
	Relative humidity	<0.5% range
	Barometric pressure	<1 hPa range
	Soil volumetric water content	<0.05% range
	Soil temperature	<0.05 °C range
Monotone Ramp (6 h)	Air temperature	Std. dev. < 0.15 °C; mean Δ ≥ 0.75 °C consistently
Jitter (6 h window)	Air temperature	Std. dev. > 5 °C
	Relative humidity	Std. dev. > 15%
	Soil volumetric water content	Std. dev. > 2%
Ultra-Low Variance (7 d)	Soil temperature	Std. dev. < 0.25 °C
	Soil volumetric water content	Std. dev. < 0.15%
High-Frequency Flip (3 h)	Soil temperature	Flip rate > 90%; Δ > 0.1 °C per interval
	Soil volumetric water content	Flip rate > 90%; Δ > 0.5% per interval
Persistent High Offset	Air temperature	>10 °C above 3-month median for ≥3 days
	Relative humidity	>15% above 3-month median for ≥3 days
	Barometric pressure	>10 hPa above 3-month median for ≥3 days
	Soil volumetric water content	>7.5% above 3-month median for ≥3 days

- Physical Range Limits—Ensures all observations fall within physically meaningful bounds, capturing only clearly erroneous measurements. For instance, air temperature values outside −40 °C to 60 °C or Soil VWC exceeding 0%–100% are flagged. Similar conservative limits are defined for relative humidity, atmospheric pressure, wind speed, and pollutant concentrations (e.g., PM1/PM2.5/PM10) (Campbell et al., 2013).
- Step Spike—Identifies large, sudden changes between consecutive observations (30-min interval), such as >5°C for temperature or >10% for Soil VWC, often resulting from transient electronic or communication glitches (Taylor & Loescher, 2013).
- Flat-Line (24 h)—Flags periods of implausibly low variability that indicate sensor blockage or inactivity, applying thresholds such as a 2 °C range for air temperature or 0.05% for Soil VWC over 24 h (Foken & Wichura, 1996).
- Monotone Ramp (6 h)—Detects continuous, monotonic increases or decreases suggestive of calibration drift. This condition requires a standard deviation < 0.15 °C and a consistent average incremental change ≥ 0.75 °C per interval.
- Jitter (6 h window)—Highlights noisy or unstable data with excessive short-term variability, e.g., 6-h rolling standard deviation > 5 °C for temperature or >2% for Soil VWC (Taylor & Loescher, 2013).
- Ultra-Low Variance (7 d)—Flags suspiciously stable readings with very low long-term variability, such as 7-day standard deviation < 0.25°C for soil temperature or <0.15% for Soil VWC, often indicating malfunction or persistent bias (Campbell et al., 2013).
- High-Frequency Flip (3 h)—Detects frequent directional reversals in soil measurements (e.g., Soil VWC or soil temperature) where >90% of changes within 3 h alternate in sign with minimum magnitude differences of 0.5% or 0.1°C, indicating unstable sensor output (Taylor & Loescher, 2013).
- Persistent High Offset—Identifies prolonged deviations above expected baselines, such as air temperature > 10 °C or relative humidity > 15% above the 3-month median for ≥3 days, which may result from calibration drift, sensor contamination, or physical interference (Campbell et al., 2013).

3.2. Machine learning-based error detection

To complement the conservative physics-based quality control, we implement a machine learning (ML) approach specifically designed to detect anomalies and quantify the probability that individual sensor measurements represent errors. This probabilistic modeling framework utilizes Positive-Unlabeled (PU) learning with an XGBoost classifier (Kiryo et al., 2017; Niu et al., 2016; Timmons et al., 2020).

In the PU-learning framework, measurements flagged by the conservative physics-based thresholds serve as the positive class (label = 1), representing confirmed anomalies with high confidence. All other measurements are treated as unlabeled samples (label = 0), rather than definitively negative (non-error). This approach acknowledges that some unlabeled data points may still contain subtle or context-specific anomalies undetected by the conservative physical thresholds.

In total, the dataset used for model training comprises 3,693,888 records, of which 16,462 (≈ 0.45%) are physics-flagged positives and 3,677,426 (≈ 99.55%) are unlabeled samples. This highly imbalanced labeling reflects the typical conditions of semi-supervised environmental QC problems, where only a small fraction of anomalies is explicitly identified while the remainder of the data have uncertain status. The PU-XGBoost model therefore learns to separate high-confidence anomalies from uncertain measurements without assuming that unflagged data are error-free.

We employ an XGBoost classifier due to its effectiveness in capturing complex feature interactions, robustness to noisy and imbalanced data, and ability to model non-linear relationships (Chen et al., 2015; Zhang et al., 2022 The model uses a comprehensive set of predictors derived from all meteorological and soil variables listed in Table 1 that provide continuous and reliable records across multiple sites.

The feature set used for the PU-XGBoost model comprises three main categories: (1) Temporal-statistical features; (2) Sensor-specific categorical features, describing the type and model of each sensor (e.g., BME280, BME680, AQT530, WXT536) and its site-level ID; and (3) Measurement-variable categorical features, representing the environmental variable measured (e.g., air temperature, humidity, barometric pressure, soil temperature, or VWC). The list of statistical features, engineered to capture temporal dynamics, statistical variability, and contextual information, are summarized below:

- Temporal Lag Features—Original observations shifted over multiple lag intervals (30 min–24 h) to capture short- and medium-term temporal dependence.
- Rolling Statistical Features—Moving-window means, standard deviations, minima, maxima, and ranges computed over 3-h, 6-h, 24-h, and 7-day windows to represent evolving variability.
- Differential Features—First and absolute differences between consecutive observations to highlight abrupt changes or spikes.
- Site-Level Median Deviations—Deviation of each observation from the contemporaneous site-specific median to quantify cross-sensor inconsistencies.
- Persistent High-Offset Flags—Binary indicators of long-term deviations exceeding the physical thresholds defined in Section 3.1, linking the ML model to physics-based QC results.
- Cyclic Temporal Encodings—Hour-of-day and day-of-year transformed into sine-cosine pairs to capture diurnal and seasonal cycles.
- Sensor Metadata Encodings—Categorical variables describing sensor type (e.g., BME280, BME680, AQT530, WXT536) and measured parameter category (e.g., temperature, humidity, pressure, air quality, soil) providing contextual information to the model.

To prepare the data for the PU-XGBoost model, the raw time series were transformed into a final structured dataframe with dimensions $N \times M$. Here, $N = 3{,}693{,}888$ represents the total number of 30-minute observational records collected across all 13 sites and variable types during the study period (October 2024 to April 2025). $M = 77$ represents the total number of predictor features, which includes the engineered temporal statistics (rolling means, minima, maxima, and variances over 3 h, 6 h, 24 h, and 7d windows), lag features, and categorical encodings for sensor metadata. The target variable is a binary vector of length N, where observations flagged by the physics-based QC are labeled as 1 (positive) and all other observations are labeled as 0 (unlabeled).

To ensure rigorous model performance, hyperparameter tuning was conducted using a 5-fold time-series cross-validation (CV) procedure. This temporal CV approach preserves the chronological order of observations, preventing any leakage of future information into past training folds and thereby maintaining the validity and realism of model evaluation. Specifically, the dataset spanning October 2024 to April 2025 was sequentially divided into five continuous, non-overlapping folds. Given a total of approximately 3.6 million 30-minute samples, each CV iteration used about 0.7 million observations (20%) as the validation (test) set and the remaining 2.9 million samples (80%) for training.

The final PU-XGBoost model produces probabilistic estimates of anomaly likelihood for each measurement, rather than binary classifications. Given the conservative nature of the physics-based criteria (used as the positive class reference), these probabilistic outputs do not aim to replicate physics-based flags exactly. Instead, they deliver detailed assessments that capture context-dependent deviations. We adopt XGBoost's binary logistic objective, in which each tree contributes an additive log-odds prediction that is transformed to a probability with the sigmoid:

$$P(y = 1|x) = \frac{1}{1+\exp(-\sum_k f_k(x))} \tag{1}$$

Here, x denotes the input feature vector for a given observation (of dimension $M = 77$), and $f_k(x)$ represents the prediction score output by the k -th decision tree in the ensemble. The summation is performed over all trees in the model to compute the total log-odds. $y = 1$ represents physics-flagged positives (confirmed anomalies), and unlabeled observations are coded as $y = 0$ for PU training. The resulting probabilities (0–1) quantify anomaly likelihood and can be thresholded or ranked according to application-specific strictness; in this study we primarily analyze and report the probability distributions rather than enforcing a single hard cutoff.

To specifically address the significant class imbalance (0.45% positive vs. 99.55% unlabeled), we employed the scale_pos_weight (scale of positive samples) parameter in XGBoost. This parameter was calibrated to the ratio of the number of unlabeled samples to positive samples, balancing the positive and negative weights to ensure the model converges effectively despite the scarcity of anomaly labels.

All computations were implemented in Python 3.12 using the xgboost, scikit-learn, pandas, and NumPy libraries. Prior to training, all input predictors were standardized (z-score normalization) so that each feature had zero mean and unit variance. This scaling was applied after feature engineering and ensured consistent numerical behavior across variables with different physical units (e.g., °C, hPa, %). Although XGBoost models are inherently robust to feature scale, this normalization step improved training stability and consistency across folds in the cross-validation process. Figure 3 shows the overall flowchart of the QC framework.

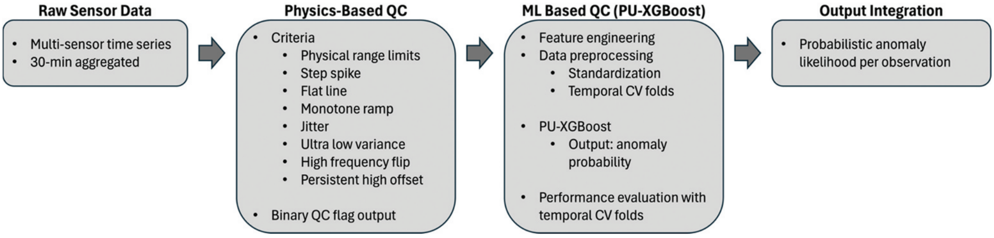

Figure 3. Flowchart summarizing the end-to-end QC framework.

For benchmarking, an unsupervised isolation forest model (Lesouple et al., 2021; Liu et al., 2008; Xu et al., 2017) was implemented under identical preprocessing and feature-engineering conditions as the PU-XGBoost model. The same 30-minute aggregated inputs and standardized predictors were used, including temporal lags, rolling statistics, and metadata encodings. This baseline comparison provided an unsupervised reference for evaluating the sensitivity and stability of the proposed hybrid QC framework.

4. Results

4.1. Physics-based error detection results

The conservative physics-based error detection approach summarized in Table 4 was first applied to the entire dataset to identify clearly anomalous sensor readings. This initial filtering identified 0.79% of all measurements as physically implausible.

The dominant source of anomalies was the 24-hour flat-line condition (42% of total error flags), indicative of prolonged sensor inactivity, communication breakdowns, or hardware malfunctions leading to constant readings. The second-largest source of errors was the persistent high offset anomaly (33%), where sensor readings consistently deviate above expected baselines for extended periods. Such persistent deviations often suggest calibration drift, sensor contamination, or physical damage to sensor shielding.

Less frequent were the 7-day low variance issues (11%), signaling extended periods of abnormally low sensor variability, potentially caused by sensor degradation, data transmission issues, or physical blockage. Similarly, the 6-hour jitter anomalies (11%), characterized by excessive short-term fluctuations, often indicate electronic noise, environmental interference, or sensor instability. The remaining errors accounted for less than 2% of total errors each.

Although most variables demonstrated low error rates (less than 0.1%), a small number of sensors and variable combination exhibited notably elevated error frequencies. For instance, barometric pressure measurements from the BME680 sensor located at node W0A0 showed a very high error rate of 32% due to 24-hour flat line, suggesting chronic sensor instability or repeated data transmission problems. Additionally, soil temperature at 60 cm depth from the one of the MFL nodes (MNLA010B) at W0A0 showed a 17% anomaly rate due to flat line and jitter, likely reflecting issues related to sensor placement or soil moisture intrusion. An overview of the physics-based QC outcomes is provided in Figure 4.

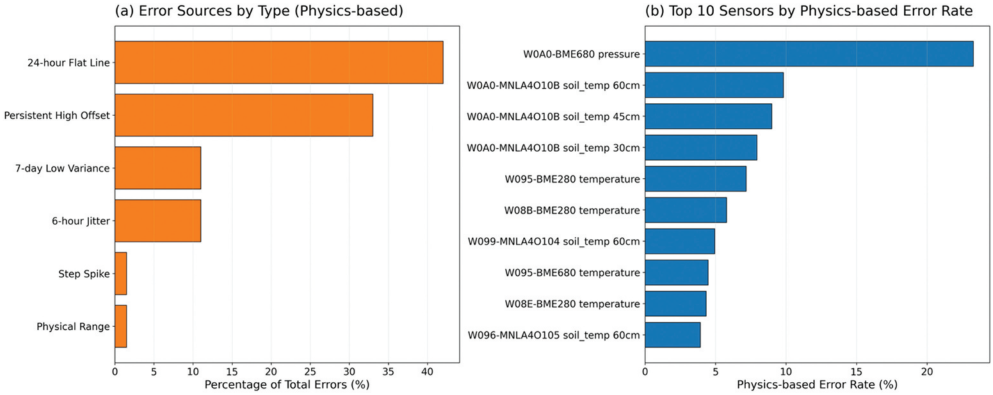

Figure 4. Overview of physics-based anomaly detection results. (a) Distribution of error sources across all sensors and variables. (b) Top 10 sensor-variable combination ranked by physics-based error rate.

Table 5. Optimized hyperparameters for the PU-XGBoost classifier.

Model	Hyperparameter	Optimized Value	Search Range
XGBoost Classifier	Max Depth	6	2, 3, 4, 5, 6, 7, 8, 9
	Learning Rate	0.01	0.001, 0.005, 0.01, 0.05, 0.1, 0.5
	Subsample	1	0.5, 0.6, 0.7, 0.8, 0.9, 1
	Colsample by Tree	0.8	0.5, 0.6, 0.7, 0.8, 0.9, 1
	Number of Estimators	300	50, 100, 150, 200, 250, 300, 350, 400, 450, 500
	Scale Positive Weight	220	Ratio of negative to positive samples

4.2. PU-XGBoost error detection results

A grid search algorithm was used to optimize the hyperparameters used for the PU-XGBoost model. The selected optimal parameters and search range are shown in Table 5. Leveraging the PU-XGBoost framework, we estimate the probability that each individual sensor measurement represents an error.

First, we evaluated the model's performance using 5-fold CV. Because the PU-XGBoost produces continuous anomaly probabilities rather than binary error flags, it was necessary to determine an optimal probability threshold for classification. Multiple thresholds were tested to maximize the Area Under the Receiver Operating Characteristic curve (AUC-ROC) , and the optimal cutoff was found to be 0.79. Based on this threshold, standard classification metrics—including accuracy, precision, recall, and F1 score—were computed and summarized in Table 6. As a baseline comparison, an isolation forest model was trained using the same dataset and evaluation procedure, with the AUC-ROC optimized threshold of its own. Its overall performance was consistently inferior to that of the PU-XGBoost model, supporting the suitability of the proposed semi-supervised approach for anomaly detection.

Table 7 compares the distribution of these anomaly probabilities between two groups: (1) measurements flagged as errors by the conservative physics-based criteria (Physics-based Error), and (2) those that were not flagged (Physics-based non-Error). This comparison illustrates how closely the probabilistic model aligns with the physics-based detection while also revealing its ability to capture additional subtle irregularities.

The results show a clear and strong correspondence between the two approaches. Nearly all measurements flagged by the physics-based QC exhibit extremely high anomaly probabilities (median > 0.99), confirming that the PU-XGBoost model effectively recognizes clear-cut, severe anomalies. Conversely, the majority of measurements that passed the physics-based check display very low probabilities, indicating high data reliability.

However, a small fraction of these unflagged observations shows moderately elevated probabilities (above the 95th percentile). These outliers highlight the model's capacity to identify context-dependent or gradual anomalies that conservative physical thresholds

Table 6. Model performance statistics from 5-fold cross-validation for the PU-XGBoost and Isolation Forest models.

Model	Dataset	Accuracy	Precision	Recall	F1 Score
PU-XGBoost	Training	0.99867	0.7415	1.0000	0.8516
	Validation	0.9979	0.6882	0.9998	0.8152
Isolation Forest	Training	0.9958	0.4963	1.0000	0.6634
	Validation	0.9946	0.4528	1.0000	0.6233

Table 7. Comparison of anomaly probability distributions derived from the PU-XGBoost model for two data groups defined by physics-based quality control: (i) measurements flagged as errors ("physics-based Error") and (ii) those not flagged ("physics-based non-Error"). The consistently high probabilities among physics-based errors confirm strong model agreement, while occasional elevated probabilities among non-error cases indicate additional subtle anomalies detected by the PU-XGBoost model.

Percentile	Physics-based Error (16,462 Samples)	Physics-based non-Error (3,677,426 Samples)
0	0.5830	0.000
0.5	0.9273	0.000
1	0.9551	0.000
5	0.9850	0.000
10	0.9910	0.000
25	0.9959	0.000
50	0.9982	0.000
75	0.9924	0.0002
90	0.9997	0.0015
95	0.9998	0.0043
99	0.9999	0.0495
99.5	0.9999	0.1825
100	0.9999	0.9992

may overlook—for instance, emerging calibration drift, environmental interference, or subtle multi-variable inconsistencies.

This clear contrast between the two distributions, as summarized in Table 7, demonstrates that the probabilistic machine learning approach effectively complements physics-based error detection by providing a graded and sensitive assessment of data quality. The probabilistic outputs can therefore inform downstream modeling and analysis by explicitly quantifying uncertainty and reducing the risk of discarding useful, yet slightly irregular, measurements.

4.3. *Feature importance in PU-XGBoost*

The importance of features used in the PU-XGBoost model is analyzed to determine which factors most strongly influence anomaly-detection performance. Feature importance is quantified using the GAIN metric provided by the XGBoost algorithm, which represents the average improvement in the model's log-loss when the feature is used for splitting. A higher GAIN value therefore indicates a larger contribution to reducing predictive uncertainty rather than a higher number of flagged anomalies. For interpretability, the GAIN values were further aggregated by feature family (e.g., temporal lags, rolling statistics, metadata encodings) to compare the relative influence of different feature groups.

In addition to numerical predictors, the PU-XGBoost incorporates categorical encodings for sensor type, measured variable, and site ID. These contextual features allow the model to capture both local behaviors specific to individual sensors and shared physical relationships across the network. As a result, the model can effectively generalize anomaly-detection rules across heterogeneous sensor configurations and locations while maintaining sensitivity to site-specific operational characteristics. The feature set analyzed comprises three distinct types:

- Statistical Features (e.g., rolling means, minimum, maximum values): These features quantify short-term temporal patterns and deviations within the sensor data over rolling windows ranging from 3 hours to 7 days.
- Sensor-Type Categorical Features (e.g., sensor type: WXT, AQT, BME280, BME680): These categorical encodings allow the model to leverage sensor-specific anomaly characteristics that might differ based on sensor design or environmental exposure.
- Measured-Variable Categorical Features (e.g., humidity, pressure, temperature): These indicate the physical variables monitored, enabling the model to capture systematic differences in error behavior across different measured parameters.

Table 8 presents the top 15 features based on their importance (GAIN). Among these, statistical features calculated over rolling temporal windows are particularly significant, underscoring the importance of temporal context for effective anomaly detection. For instance, the most influential feature (6-hour rolling minimum) suggests its critical role in distinguishing normal sensor readings from subtle or transient anomalies. Similarly, longer temporal windows (24-hour rolling minimum) and short-term rolling means (3-hour rolling mean) also emerge as vital features, reflecting the relevance of sustained deviations from typical measurement patterns.

Categorical features representing sensor types (e.g., WXT and AQT sensors) exhibit substantial importance as well, indicating that anomaly signatures differ notably by sensor hardware. Specifically, the WXT sensors has the second-highest GAIN value overall, highlighting that measurements from WXT sensors often show unique or pronounced anomaly patterns. Similarly, the BME680 sensor type, commonly used for atmospheric pressure and temperature measurements, exhibits notable GAIN, emphasizing sensor-specific influences on data quality.

Among the measured variables, humidity, pressure, and WXT-related parameters emerged as particularly informative for the PU-XGBoost model. Their high feature-importance values do not simply indicate that these sensors produce more frequent errors; rather, they reflect the strong discriminative contribution of these predictors in distinguishing anomalous from nominal conditions. WXT sensors

Table 8. Top 15 features contributing to anomaly detection in the PU-XGBoost model, ranked by the GAIN metric.

Rank	Feature Name	Type	GAIN
1	6-hour rolling minimum	Numeric rolling feature	522.5
2	Sensor type: WXT	Categorical sensor-type indicator	294.2
3	Variable type: Humidity	Categorical variable indicator	235.3
4	24-hour rolling minimum	Numeric rolling feature	231.7
5	Sensor type: AQT	Categorical sensor-type indicator	155.6
6	3-hour rolling mean	Numeric rolling feature	142.7
7	6-hour rolling maximum	Numeric rolling feature	121.9
8	Current observed value	Numeric instantaneous feature	115.3
9	3-hour rolling minimum	Numeric rolling feature	93.5
10	Variable type: Barometric Pressure	Categorical variable indicator	92.2
11	Lagged observation (30-min)	Numeric lag feature	85.6
12	Sensor type: BME680	Categorical sensor-type indicator	65.0
13	3-hour rolling maximum	Numeric rolling feature	55.1
14	Variable type: Vapor Pressure Deficit	Categorical variable indicator	50.4
15	Variable type: Water Temperature	Categorical variable indicator	47.6

record multiple atmospheric parameters (e.g., wind speed, direction, rainfall, humidity, and temperature) at a higher sampling frequency than other sensor types, resulting in a broader and more dynamic representation of environmental variability. This richer data stream allows the model to better learn subtle patterns that differentiate true physical variability from sensor anomalies, which can manifest under rapidly changing weather conditions. Similarly, humidity and pressure variables often exhibit complex, context-dependent variability that affects multiple sensing processes across the network. These parameters are inherently more sensitive to environmental noise, condensation, and calibration drift, making them effective indicators of compound or transient anomaly states. Furthermore, variations in humidity and pressure can indirectly signal environmental conditions that stress sensor reliability—such as strong pressure gradients during storms or low-humidity, high-temperature regimes during heat waves. Consequently, the elevated importance of these variables and sensor types reflects their multifaceted role as both direct indicators of sensor stress and contextual predictors that enhance the model's ability to detect anomalies across heterogeneous sensor configurations and operational environments.

Comparing across these different feature types (statistical, sensor-type, and variable-type) is possible because the XGBoost model inherently evaluates all features simultaneously based on their effectiveness in reducing prediction error. Although these feature groups represent conceptually different information—such as sensor hardware differences versus temporal trends in measurements—the model aggregates these diverse contributions into a unified measure of importance.

To assess the stability of feature-importance estimates under class imbalance, the model was retrained after randomly reducing the number of positive samples to 90%, 80%, and 50% of the original count. This procedure was conducted solely to verify the robustness of the GAIN metric against data scarcity and was not used as a data balancing strategy for the final model training. The resulting rankings were highly consistent, indicating that the GAIN metric is robust to label imbalance in this PU-learning setup. This is partly because the XGBoost performs gradient-based optimization and internal weighting at each split, normalizing contributions from minority samples.

Overall, this analysis confirms that a combination of temporal statistical metrics, sensor-specific characteristics, and measured environmental variables collectively enhances anomaly detection performance, providing detailed and context-sensitive error identification beyond simpler threshold-based approaches. Furthermore, the feature importance analysis highlights specific sensors, variables, and temporal conditions that significantly influence the model's predictions, thereby pinpointing areas that may require proactive maintenance, targeted sensor calibration, or heightened monitoring to ensure sustained data quality and system reliability.

4.4. Visual inspection of errors

Visual inspections of sensor data further illustrate the complementary strengths of physics-based and PU-XGBoost error detection methods. Figure 5 provides representative examples comparing anomalies identified by both methods across diverse sensor types and conditions.

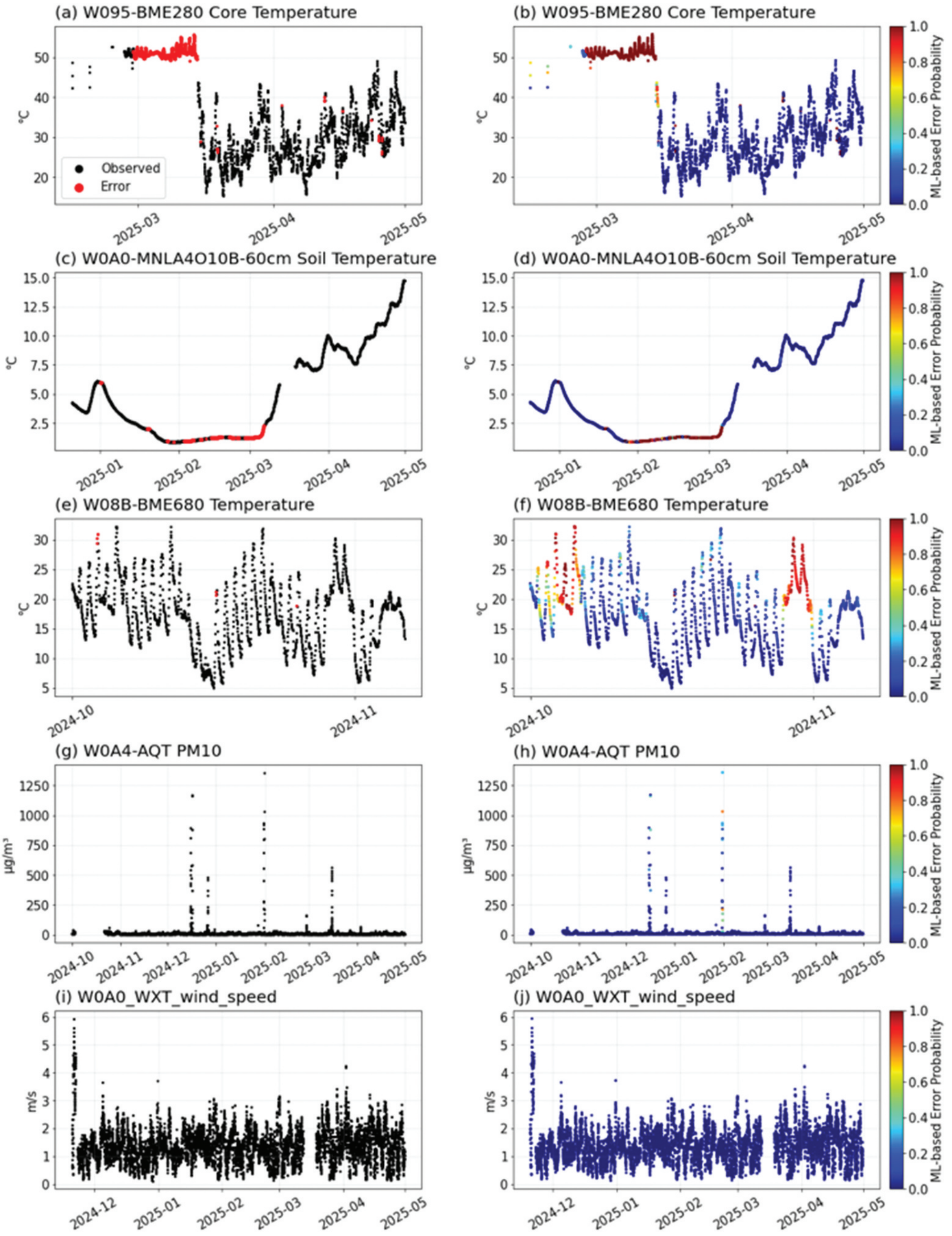

Figure 5. Visual comparison of sensor anomalies identified by physics-based and ML-based methods across diverse sensors and measurement conditions. Left-hand panels (a, c, e, g, i) show anomalies flagged by conservative physics-based criteria (red dots), while right-hand panels (b, d, f, h, j) illustrate the corresponding ML-derived probabilities of error (color bar, 0–1 scale).

Figure 5(a–d) demonstrate that the PU-XGBoost model robustly captures obvious anomalies already detected by physics-based methods, particularly errors arising from persistent offsets (Figure 5(a–b)) and prolonged flat lines (Figure 5(c–d)). These cases emphasize the strong alignment with clearly anomalous readings, as indicated quantitatively by the extremely high probabilities assigned to physics-flagged errors. The ML model, therefore, provides strong confirmation and supports these known sensor issues identified by physical criteria, effectively validating sensor maintenance needs and ensuring reliability.

In Figure 5(e–f), the complementary value of the ML-based approach becomes more evident, capturing subtler anomalies that were not flagged by the conservative physics-based method. The BME680 temperature sensor readings at W08B lack a clear and consistent diurnal cycle in certain periods (particularly in late October), suggesting potential sensor irregularities. The ML model assigns moderate-to-high error probabilities (40%–90%) to these ambiguous intervals. For the early-October window, the PU-XGBoost assigned moderate anomaly probabilities despite an apparent diurnal cycle. This behavior reflects context-dependent inconsistencies—elevated cross-sensor RMSE together with reduced diurnal amplitude and a phase shift relative to neighboring nodes—captured by the probabilistic model while remaining below binary physics-based thresholds. A Mann–Whitney U test confirmed that the mean cross-sensor RMSE during Oct 1–6 (19.0 °C) was significantly higher than during Oct 7–Nov 1 (16.7 °C, $U = 243{,}871$, $p < 10^{-52}$), supporting that the early-October anomalies correspond to statistically distinct network discrepancies rather than random variability.

Figure 5(g–j) illustrates scenarios with minimal anomalies—specifically particulate matter (PM10) from W0A4-AQT and wind speed from W0A0-WXT. Here, both physics-based and ML-based methods do not flag significant errors, reflecting confidence in the sensor data quality. However, the ML-based model assigns elevated error probabilities (around 20%–30%) to isolated, unusually high PM10 measurements (Figure 5(h)). Although these cases are not definitively anomalous, the moderate error probabilities imply uncertainty about the reliability of these data. Users can interpret these moderate error probabilities as indicators of data points requiring caution, possibly warranting closer examination or cross-validation with other data sources before use in sensitive analyses.

To evaluate the robustness of the proposed PU-learning framework, we also diagnosed the visual performance of isolation forest model. Its performance was qualitatively assessed by comparing anomaly occurrence patterns and overlap with physics-based flags. The Isolation Forest tended to over-flag transient fluctuations, consistent with result in Table 6, detecting roughly twice as many anomalies as the PU-XGBoost under comparable contamination settings.

4.5. *Sensor and variable-specific error analysis*

The analysis of site-sensor combinations ranked by the mean ML-based error probability reveals distinct spatial and sensor-specific patterns of heightened anomaly likelihood. Here, the ML-based error percentage represents the average predicted probability of being anomalous across all 30-minute observations for each site-sensor pair, rather than a thresholded (binary) error rate. This continuous formulation captures gradual differences in anomaly intensity and provides a more nuanced view of sensor reliability across the network.

Here we calculate the ML-based error percentage as the mean of all anomaly probabilities produced by the PU-XGBoost model for each sensor or variable, rather than binary classification with threshold value. This continuous metric captures the expected likelihood of anomalies without requiring a fixed probability threshold.

The highest ML-based error rate is observed for the BME680 sensor at site W0A0, exceeding 8%, suggesting persistent issues potentially related to sensor calibration error, sensor damage, or physical interference (Figure 6). Similarly, BME680 sensors at sites W08B and W096 also exhibit elevated error percentages (approximately 4%–6%), reinforcing the hypothesis of common sensor-specific vulnerabilities across multiple deployment locations. Notably, the subsurface MFR node MNLA4O10B at W0A0, also shows a relatively high anomaly rate, which aligns with earlier findings highlighting reliability challenges in subsurface data due to environmental factors like moisture intrusion, sensor placement issues, or networking issues.

Variable-level examination based on ML-based error percentages (Figure 6(b)) demonstrates the model's enhanced sensitivity to specific measurement types,

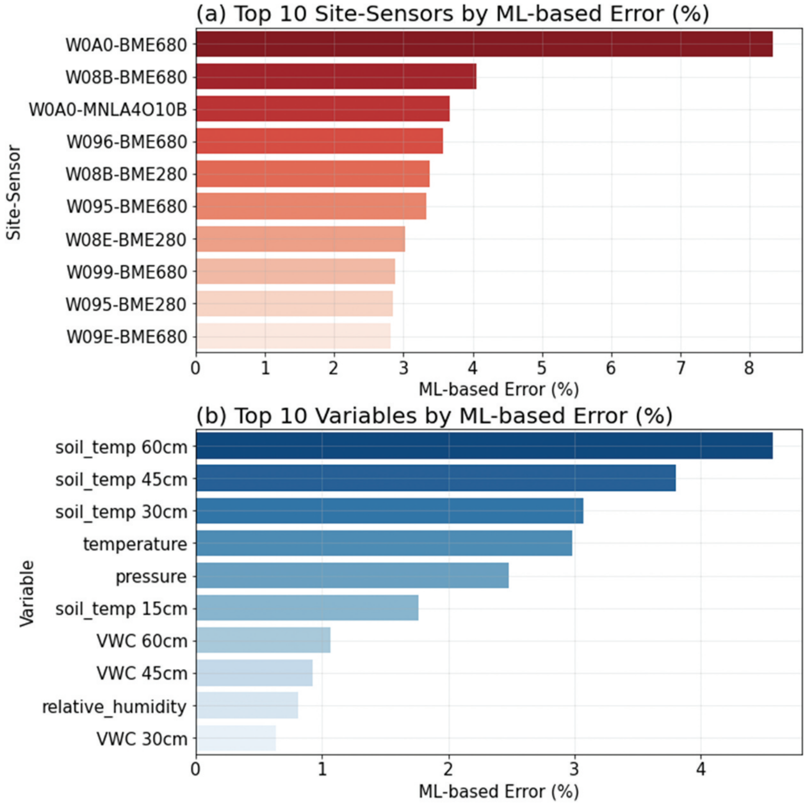

Figure 6. Sensor and variable-specific error analysis based on ML-derived anomaly probabilities. (a) Top 10 site-sensor combinations ranked by ML-based error percentage; (b) top 10 environmental variables ranked by ML-based error percentage.

notably subsurface soil temperatures. Soil temperature measurements at deeper depths (30–60 cm) consistently exhibit the highest ML-based error rates, with soil temperature at 60 cm depth leading at over 4%. These elevated error rates are likely associated with challenges inherent to subsurface measurements, such as potential sensor exposure to environmental contamination, physical disturbances (e.g., soil moisture intrusion), or transmission issues arising from reduced wireless signal strength at increased depths. Nevertheless, it's important to note that these observed error levels for deeper soil sensors are similar to error levels found for specific atmospheric sensors and notably lower than the errors documented for certain atmospheric measurements (such as the barometric pressure sensor at W0A0) described previously. Therefore, rather than broadly attributing higher error rates to all MFR or subsurface sensor deployments, these results emphasize the need to evaluate error patterns in the context of individual sensor characteristics and specific deployment locations.

5. Summary and discussion

5.1. Summary of results

This study developed and evaluated a hybrid QC framework that integrates physics-based thresholding with a probabilistic PU-XGBoost model to improve the reliability and interpretability of urban environmental sensor data. The results show that this combined approach enhances both sensitivity and flexibility in detecting anomalies compared with traditional rule-based QC methods.

The hybrid framework effectively differentiates between gross errors captured by conservative physics-based rules and more subtle, context-dependent deviations identified by the probabilistic model. The PU-XGBoost algorithm outputs anomaly probabilities rather than binary classifications, allowing users to assess confidence levels for each observation and make informed decisions regarding data inclusion or further inspection. Using the ROC–AUC criterion, the optimal probability threshold was determined to be 0.79, representing the best alignment with the physics-based model. This threshold, however, can be adjusted according to user preferences or operational requirements. Overall, the PU-XGBoost approach outperformed the Isolation Forest method in both sensitivity and consistency.

Feature-importance analysis highlights that temporal statistical features—particularly rolling minima and variance metrics over 6-hour and 24-hour windows—are the most influential predictors, indicating that many anomalies evolve gradually rather than appearing as isolated spikes. Sensor-type features (e.g., AQT, WXT, BME) and environmental variables such as humidity and pressure also play significant roles, reflecting hardware-specific sensitivities and environmental dependencies. Together, these results demonstrate that the hybrid physics-ML framework provides a robust and interpretable method for identifying both apparent and latent anomalies across urban sensor networks.

5.2. Implication and comparison with previous approaches

Traditional physics-based QC methods are simple, transparent, and yield very low false-positive rates but are limited in sensitivity, often missing gradual drifts or compound errors that fall within physical ranges. The proposed hybrid framework builds upon this foundation by using the physics-based flags as high-confidence positive labels within a positive-unlabeled (PU) learning setup, thereby extending anomaly detection to ambiguous cases without requiring exhaustive negative labels.

To benchmark performance, we compared the PU-XGBoost model against representative QC strategies. The physics-only baseline successfully eliminated clear outliers but can miss the anomalies detected by the probabilistic model. In contrast, an unsupervised Isolation Forest model trained on the same dataset exhibited unstable behavior, detecting more noise-like fluctuations while producing a noticeably higher false-positive rate, particularly for temperature and humidity variables. The PU-XGBoost framework achieved stronger detection sensitivity than the physics-based baseline and greater stability than the Isolation Forest, demonstrating a more balanced trade-off between precision and recall. This result reinforces the practical strength of the PU-learning formulation for semi-supervised QC tasks, where only positive (anomalous) samples can be confidently identified.

Compared with fully supervised classifiers, which require extensive labeled datasets rarely available in environmental monitoring, the PU-XGBoost maintains operational feasibility and interpretability while achieving comparable anomaly detection performance. Its probability-based outputs also align with operational needs, enabling flexible thresholds and integration with decision-making systems. These advantages position the hybrid approach as a bridge between traditional deterministic QC and more adaptive machine-learning methods documented in previous environmental sensing studies.

5.3. Practical implications for real-time operations

The probabilistic outputs generated by the PU-XGBoost model can be directly integrated into operational dashboards to enhance real-time sensor management within the CROCUS network. High anomaly probabilities (≥0.9) could trigger immediate maintenance alerts or automatic sensor diagnostics, while intermediate probabilities (0.6–0.9) may flag data for analyst review. This probabilistic prioritization enables adaptive maintenance scheduling that balances diagnostic accuracy and resource constraints.

A real-time CROCUS QC dashboard could visualize node-level median probabilities and short-term anomaly trends, linking these with sensor metadata such as type, age, and calibration history. Over time, accumulated probability metrics would help identify persistent degradation patterns, enabling proactive calibration or replacement before data quality deteriorates.

Beyond CROCUS, this workflow can be readily adapted for other urban or regional monitoring systems. The modular architecture—comprising physics-based flagging, feature engineering, standardization, and PU-based anomaly scoring—can be reused with minimal code or design changes. Recalibration of physical thresholds, rolling window lengths, and probability cutoffs would tailor the model to local climates and sensor characteristics. Regular retraining using newly verified anomalies would further improve adaptability and resilience across diverse environmental contexts.

5.4. Limitations and future directions

While the hybrid QC framework successfully enhances sensitivity and interpretability, it still depends on the accuracy of physics-based flags that define the positive class. Any errors systematically missed by these criteria remain underrepresented, potentially limiting detection of entirely novel failure modes. Additionally, although the approach demonstrated strong performance across the Chicago-based CROCUS deployment, generalization to other cities and sensor types requires validation under different environmental and infrastructural settings.

The current implementation also inherits the limitations of rule-based labeling, as the physics-flagged anomalies used for training may not comprehensively represent the full spectrum of error behaviors. Future work will therefore focus on expanding the training dataset to include larger, multi-city observations and developing semi-supervised learning extensions that combine labeled and unlabeled data. Such approaches could improve the model's robustness and facilitate more reliable cross-domain generalization.

Further methodological advancements could incorporate unsupervised or adaptive extensions— such as autoencoder-based residual analysis, cross-sensor consistency checks, or temporal-coherence metrics—to identify new classes of anomalies not captured by current thresholds. Integrating human-verified feedback and implementing periodic model retraining will further enhance robustness and reduce dependency on static physical labels. Over the longer term, advancing toward a self-adaptive QC system that continuously learns from operational feedback could enable autonomous, real-time monitoring across expanding urban sensor networks.

6. Conclusion

This study presents a hybrid QC framework that merges physics-based thresholds with a probabilistic PU-XGBoost model to evaluate and enhance data quality in dense urban environmental sensor networks. The framework improves sensitivity to subtle, context-dependent anomalies while preserving the interpretability and conservatism of physics-based QC methods.

By producing probabilistic anomaly likelihoods, the approach supports graded, confidence-based decisions for data management and maintenance prioritization. In benchmarking analyses, the PU-XGBoost outperformed a purely physics-based baseline and exhibited greater stability and lower false-positive behavior than an unsupervised Isolation Forest, confirming its suitability for semi-supervised QC where only positive (anomalous) labels are reliable.

The workflow can be seamlessly integrated into real-time dashboards to trigger automated alerts and facilitate proactive maintenance. Its modular and transferable design allows for application to other urban monitoring systems with modest recalibration of thresholds and retraining frequency. Future work will focus on integrating unsupervised consistency measures, adaptive retraining, and human-in-the-loop feedback to develop a scalable, interpretable, and self-improving QC system for next-generation urban sensing networks.

AI disclosure statement

The authors declare that no generative artificial intelligence (AI) or AI-assisted technologies were utilized in the preparation, writing, data analysis, or creation of figures for this manuscript. All contents, including text, data analysis, and visualizations, are the original work of the authors.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This material is based upon work supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research's Urban Integrated Field Laboratories CORCUS Project research activity, under Award Number DE-SC0023226.

ORCID

Jangho Lee http://orcid.org/0000-0002-8942-1092

Data availability statement

CROCUS sensor data used in this study is publicly available at: https://crocus.sagecontinuum.org/nodes.

References

Balaprakash, P., Collis, S., Kim, Y., Beckmann, P., Cadeddu, M., Gonzalez-Meler, M., Sullivan, R., Madireddy, S., & Kotamarthi, R. (2021). AI-enabled Modex and edge-computing over 5G for improving the predictability of water cycle extremes. https://doi.org/10.2172/1769672

Beele, E., Reyniers, M., Aerts, R., & Somers, B. (2022). Quality control and correction method for air temperature data from a citizen science weather station network in Leuven, Belgium. *Earth System Science Data*, *14*(10), 4681–4717. https://doi.org/10.5194/essd-14-4681-2022

Campbell, J. L., Rustad, L. E., Porter, J. H., Taylor, J. R., Dereszynski, E. W., Shanley, J. B., Gries, C., Henshaw, D. L., Martin, M. E., Sheldon, W. M., & Boose, E. R. (2013). Quantity is nothing without quality: Automated QA/QC for streaming environmental sensor data. *BioScience*, *63*(7), 574–585. https://doi.org/10.1525/bio.2013.63.7.10

Chen, F., Kusaka, H., Bornstein, R., Ching, J., Grimmond, C. S. B., Grossman-Clarke, S., Loridan, T., Manning, K. W., Martilli, A., Miao, S., Sailor, D., Salamanca, F. P., Taha, H., Tewari, M., Wang, X., Wyszogrodzki, A. A., & Zhang, C. (2011). The integrated WRF/urban modelling system: Development, evaluation, and applications to urban environmental problems. *International Journal of Climatology*, *31*(2), 273–288. https://doi.org/10.1002/joc.2158

Chen, J., Saunders, K., & Whan, K. (2021). Quality control and bias adjustment of crowdsourced wind speed observations. *Quarterly Journal of the Royal Meteorological Society*, *147*(740), 3647–3664. https://doi.org/10.1002/qj.4146

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., Chen, K., Mitchell, R., Cano, I., & Zhou, T. (2015). XGBoost: Extreme gradient boosting. *R Package Version 0.4-2*, *1*(4), 1–4.

Chen, Z., Li, Z., Huang, J., Liu, S., & Long, H. (2024). An effective method for anomaly detection in industrial internet of things using XGBoost and LSTM. *Scientific Reports*, *14*(1), 23969. https://doi.org/10.1038/s41598-024-74822-6

Ching, J., Mills, G., Bechtel, B., See, L., Feddema, J., Wang, X., Ren, C., Brousse, O., Martilli, A., Neophytou, M., Mouzourides, P., Stewart, I., Hanna, A., Ng, E., Foley, M., Alexander, P., Aliaga, D., Niyogi, D.Shreevastava, A. . . . Chen, F. (2018). WUDAPT: An urban weather, climate, and environmental modeling infrastructure for the anthropocene. *Bulletin of the American Meteorological Society*, *99*(9), 1907–1924. https://doi.org/10.1175/BAMS-D-16-0236.1

Daurenbayeva, N., Atymtayeva, L., Nurlanuly, A., Bykov, A., Turusbekova, U., & Shuitenov, G. (2025). A machine learning approach to microclimate monitoring and fault detection. *Applied Mathematics & Information Sciences*, *19*(2), 327–334. https://doi.org/10.18576/amis/190209

Duarte Rocha, A., Vulova, S., van der Tol, C., Förster, M., & Kleinschmit, B. (2021). Modelling hourly evapotranspiration in urban environments with scope using open remote sensing and meteorological data. *Hydrology and Earth System Sciences Discussions*, *2021*, 1–32. https://doi.org/10.5194/hess-26-1111-2022

Fang, X., & Bate, I. (2017). Issues of using wireless sensor network to monitor urban air quality. Proceedings of the First ACM International Workshop on the Engineering of Reliable, Robust, and Secure Embedded Wireless Sensing Systems, New york, NY.

Fazai, R., Ben Abdellafou, K., Said, M., & Taouali, O. (2018). Online fault detection and isolation of an air quality monitoring network based on machine learning and metaheuristic methods. *The International Journal of Advanced Manufacturing Technology*, *99*(9–12), 2789–2802. https://doi.org/10.1007/s00170-018-2674-6

Feng, Z., Zheng, L., & Ren, B. (2025). In-situ validation of embedded physics-based calibration in low-cost particulate matter sensor for urban air quality monitoring. *Urban Climate*, *59*, 102289. https://doi.org/10.1016/j.uclim.2025.102289

Foken, T., & Wichura, B. (1996). Tools for quality assessment of surface-based flux measurements. *Agricultural and Forest Meteorology*, *78*(1–2), 83–105. https://doi.org/10.1016/0168-1923(95)02248-1

Henriques, J., Caldeira, F., Cruz, T., & Simões, P. (2020). Combining k-means and XGBoost models for anomaly detection using log datasets. *Electronics*, *9*(7), 1164. https://doi.org/10.3390/electronics9071164

Hill, D. J. (2015). Assimilation of weather radar and binary ubiquitous sensor measurements for quantitative precipitation estimation. *Journal of Hydroinformatics*, *17*(4), 598–613. https://doi.org/10.2166/hydro.2015.072

Houston, L., Gabrys, J., & Pritchard, H. (2019). Breakdown in the smart city: Exploring workarounds with urban-sensing practices and technologies. *Science, Technology, & Human Values*, *44*(5), 843–870. https://doi.org/10.1177/0162243919852677

Jha, M., Marpu, P. R., Chau, C.-K., & Armstrong, P. (2015). Design of sensor network for urban micro-climate monitoring. 2015 IEEE First International Smart Cities Conference, Guadalajara, Mexico (pp. 1–4).

Kabir, S., Islam, R. U., Hossain, M. S., & Andersson, K. (2022). An integrated approach of belief rule base and convolutional neural network to monitor air quality in Shanghai. *Expert Systems With Applications*, *206*, 117905. https://doi.org/10.1016/j.eswa.2022.117905

Kiryo, R., Niu, G., Du Plessis, M. C., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *ArXiv, abs/1703.00593*. https://doi.org/10.48550/arXiv.1703.00593

Lee, J., & Berkelhammer, M. (2024). Observational constraints on the spatial effect of greenness and canopy cover on urban heat in a major midlatitude city. *Geophysical Research Letters*, *51*(21), e2024GL110847. https://doi.org/10.1029/2024GL110847

Lesouple, J., Baudoin, C., Spigai, M., & Tourneret, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, *149*, 109–119. https://doi.org/10.1016/j.patrec.2021.05.022

Lin, X., & Hubbard, K. (2004). Sensor and electronic biases/errors in air temperature measurements in common weather station networks. *Journal of Atmospheric and Oceanic Technology*, *21*(7), 1025–1032. https://doi.org/10.1175/1520-0426(2004)021<1025:SAEEIA>2.0.CO;2

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, Pisa, Italy (pp. 413–422) https://doi.org/10.1109/ICDM.2008.17.

Liu, S., Xu, L., Li, Q., Zhao, X., & Li, D. (2018). Fault diagnosis of water quality monitoring devices based on multiclass support vector machines and rule-based decision trees. *IEEE Access*, *6*, 22184–22195. https://doi.org/10.1109/ACCESS.2018.2800530

Lugomer, K., Soundararaj, B., Murcio, R., Cheshire, J., & Longley, P. (2017). Understanding sources of measurement error in the Wi-Fi sensor data in the Smart city. *Proceedings for the 25th GIS Research UK (GISRUK) Conference (GISRUK 2017)*.

Masson, V., Heldens, W., Bocher, E., Bonhomme, M., Bucher, B., Burmeister, C., de Munck, C., Esch, T., Hidalgo, J., Kanani-Sühring, F., Kwok, Y.-T., Lemonsu, A., Lévy, J.-P., Maronga, B., Pavlik, D., Petit, G., See, L., Schoetter, R.Tornay, N. . . . Votsis, A. (2020). City-descriptive input data for urban climate models: Model requirements, data sources and challenges. *Urban Climate*, *31*, 100536. https://doi.org/10.1016/j.uclim.2019.100536

Massoud, E. C., Hoffman, F., Shi, Z., Tang, J., Alhajjar, E., Barnes, M., Braghiere, R. K., Cardon, Z., Collier, N., Crompton, O., Dennedy-Frank, P. J., Gautam, S., Gonzalez-Meler, M. A., Green, J. K., Koven, C., Levine, P., MacBean, N., Mao, J.Mills, R. T. . . . Xu, C. (2023). Perspectives on artificial intelligence for predictions in ecohydrology. *Perspectives on Artificial Intelligence for Predictions in Ecohydrology. Artificial Intelligence for the Earth Systems*, *2*(4). https://doi.org/10.1175/AIES-D-23-0005.1

Muller, C. L., Chapman, L., Grimmond, C., Young, D. T., & Cai, X. (2013). Sensors and the city: A review of urban meteorological networks. *International Journal of Climatology*, *33*(7), 1585–1600. https://doi.org/10.1002/joc.3678

Niu, G., Du Plessis, M. C., Sakai, T., Ma, Y., & Sugiyama, M. (2016). Theoretical comparisons of positive-unlabeled learning against positive-negative learning. *Advances in Neural Information Processing Systems*, 29. https://proceedings.neurips.cc/paper/2016/hash/be3159ad04564bfb90db9e32851ebf9c-Abstract.html

O'Brien, J., Tuftedal, M., Wawrzyniak, E., Grover, M., Berkelhammer, M., Collis, S., Sankaran, R., Beckman, P., Ferrier, N., & Shahkarami, S. (2024). CROCUS Weather Data at University of Illinois-Chicago Tower. https://doi.org/10.15485/2482530

Pal, S., Raut, B., Muradyan, P., Tuftedal, M., O'Brien, J., Sullivan, R., Grover, M., Jackson, R., Berkelhammer, M., & Collis, S. (2024). *CROCUS high-frequency measurements of CO_2*. H_2O, Wind, and Temperature at University of Illinois Chicago.

Pan, L., Liu, Y., Liu, Y., Li, L., Jiang, Y., Cheng, W., & Roux, G. (2015). Impact of four-dimensional data assimilation (FDDA) on urban climate analysis. *Journal of Advances in Modeling Earth Systems*, *7*(4), 1997–2011. https://doi.org/10.1002/2015MS000487

Peters, D. R., Popoola, O. A., Jones, R. L., Martin, N. A., Mills, J., Fonseca, E. R., Stidworthy, A., Forsyth, E., Carruthers, D., & Dupuy-Todd, M. (2021). Evaluating uncertainty in sensor networks for urban air pollution insights. *Atmospheric Measurement Techniques Discussions*, 1–23. https://doi.org/10.5194/amt-15-321-2022

Rashid, B., & Rehmani, M. H. (2016). Applications of wireless sensor networks for urban areas: A survey. *Journal of Network and Computer Applications*, *60*, 192–219. https://doi.org/10.1016/j.jnca.2015.09.008

Sakthivel, P., & Sengupta, R. (2025). Spatial bias in placement of citizen and conventional weather stations and their impact on urban climate research: A case study of the urban heat island effect in Canada. *Urban Climate*, *59*, 102280. https://doi.org/10.1016/j.uclim.2024.102280

Santoso, B., Ryan, M., Wicaksana, H. S., Ananda, N., Budiawan, I., Mukhlish, F., & Kurniadi, D. (2023). Predictive maintenance automatic weather station sensor error detection using long short-term memory. *Ultima Computing : Jurnal Sistem Komputer*, *15*(2), 41–51. https://doi.org/10.31937/sk.v15i2.3403

Sun, X., Yan, S., Wang, B., Xia, L., Liu, Q., & Zhang, H. (2015). Air temperature error correction based on solar radiation in an economical meteorological wireless sensor network. *Sensors*, *15*(8), 18114–18139. https://doi.org/10.3390/s150818114

Taylor, J. R., & Loescher, H. L. (2013). Automated quality control methods for sensor data: A novel observatory approach. *Biogeosciences*, *10*(7), 4957–4971. https://doi.org/10.5194/bg-10-4957-2013

Timmons, C., Boskovic, A., Lakamsani, S., Gerych, W., Buquicchio, L., & Rundensteiner, E. (2020). Positive unlabeled gradient boosting. 2020 IEEE MIT Undergraduate Research Technology Conference (URTC), Cambridge, MA.

von Arx, G., Dobbertin, M., & Rebetez, M. (2013). Detecting and correcting sensor drifts in long-term weather data. *Environmental Monitoring and Assessment*, *185*(6), 4483–4489. https://doi.org/10.1007/s10661-012-2831-6

Warriach, E. U., & Tei, K. (2013). Fault detection in wireless sensor networks: A machine learning approach. 2013 IEEE 16th International Conference on Computational Science and Engineering, Washington, DC, United States.

Wellyantama, P., & Soekirno, S. (2021). Temperature, pressure, relative humidity and rainfall sensors early error detection system for automatic weather station (AWS) with artificial neural network (ANN) backpropagation. Journal of Physics: Conference Series (Vol. 1816, no. 1, p. 012056). IOP Publishing.

Wienhold, K. J., Li, D., & Fang, Z. N. (2024). Precision irrigation soil moisture mapper: A thermal inertia approach to estimating volumetric soil water content using unmanned aerial vehicles and multispectral imagery. *Remote Sensing*, *16*(10), 1660. https://doi.org/10.3390/rs16101660

Xu, D., Wang, Y., Meng, Y., & Zhang, Z. (2017). An improved data anomaly detection method based on isolation forest. 2017 10th International Symposium on Computational Intelligence and Design (ISCID), Hangzhou, China.

Yang, H., Cho, S., Tae, C.-S., & Zaheeruddin, M. (2008). Sequential rule based algorithms for temperature sensor fault detection in air handling units. *Energy Conversion and Management*, *49*(8), 2291–2306. https://doi.org/10.1016/j.enconman.2008.01.029

Yang, J., Liu, Q., Chen, G., Deng, X., & Zhang, L. (2021). Design of a temperature error correction method used for meteorology and climate research. *Atmospheric Research*, *263*, 105817. https://doi.org/10.1016/j.atmosres.2021.105817

Yang, S., Wang, L., Stathopoulos, T., & Marey, A. M. (2023). Urban microclimate and its impact on built environment—a review. *Building and Environment*, *238*, 110334. https://doi.org/10.1016/j.buildenv.2023.110334

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6), 15501329221106935. https://doi.org/10.1177/15501329221106935

Zhao, L., & Zhao, L. (2023). An algorithm for online stochastic error modeling of inertial sensors in urban cities. *Sensors*, *23*(3), 1257. https://doi.org/10.3390/s23031257