Journal of Computational Science 85 (2025) 102522

Contents lists available at ScienceDirect

Journal of Computational Science

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

A soft sensor open-source methodology for inexpensive monitoring of water quality: A case study of NO_3^- concentrations

Antonio Jesús Chaves *, Cristian Martín, Luis Llopis Torres, Manuel Díaz, Jaime Fernández-Ortega, Juan Antonio Barberá, Bartolomé Andreo

ITIS Software, University of Málaga, 29071 Málaga, Spain
Department of Geology and Center of Hydrogeology, University of Málaga (CEHIUMA), 29071 Málaga, Spain

ARTICLE INFO

Keywords:
Soft sensors
Internet of Things
Machine learning
Nitrate
Water monitoring

ABSTRACT

Nitrate (NO_3^-) concentrations in aquifers constitute a global problem affecting environmental integrity and public health. Unfortunately, deploying hardware sensors specifically for NO_3^- measurements can be expensive, thereby, limiting scalability. This research explores the integration of soft sensors with data streams through an use case to predict nitrate NO_3^- levels in real time. To achieve this objective, a methodology based on Kafka-ML is proposed, a framework designed to manage the pipeline of machine learning models using data streams. The study evaluates the effectiveness of this methodology by applying it to a real-world scenario, including the integration of low-cost sensor devices. Additionally, Kafka-ML is extended by integrating MQTT and other IoT data protocols. The methodology benefits include rapid development, enhanced control, and visualisation of soft sensors. By seamlessly integrating IoT and data analytics, the approach promotes the adoption of cost-effective solutions for managing NO_3^- pollution and improving sustainable water resource monitoring.

1. Introduction

Nitrate (NO_3^-) constitutes a common contaminant in groundwater and surface water worldwide [1]. The increasing presence of N-compounds such as ammonium, nitrite, or nitrate in natural systems is of global concern as it might pose a significant threat to both environmental quality and human health. Elevated nitrate concentrations in drinking water sources can lead to severe health issues, including methemoglobinemia in infants, commonly known as "blue baby syndrome" [2], thyroid effect, or even increase the risk of gastric cancer [3]. Additionally, an excess of N-availability in aquatic ecosystems contributes to nutrient enrichment which causes eutrophication in the form of algal blooms or aquatic plant growth, depleting oxygen levels and disrupting aquatic life [4].

In natural water systems, the background concentration of nitrate within the nitrogen cycle is less than 10 mg/L [5]. However, human activities are responsible for nitrate increasing trends as fertilisers constitute the major source of mobile *N* in the soil that might then be washed away by surface runoff into groundwater or fluvial systems. Human, animal and industrial wastewaters constitute as well important sources of nitrate contamination in freshwater systems [6]. Therefore, the World Health Organization and the European Union have set the nitrate threshold of 50 mg/L for drinking water to avoid potential health issues [7,8]. These threats linked to rapid hydrochemical variations produced in some capture points intended for drinking water supply evidence the need for high-frequency monitoring of contaminants such as nitrate. Considering the great extension of water distribution systems and the potentially supplied population (even in rural areas), the critical importance of controlling nitrate concentrations in water sources has prompted an urgent need for reliable and cost-effective monitoring solutions.

Conventional nitrate determination methods have relied on labour-intensive and time-consuming laboratory techniques, such as the cadmium reduction or ion chromatography. While these techniques provide accurate measurements, they suffer from several disadvantages, including high operational costs, long turnaround times, and an inability to provide real-time data [9]. Ion-selective electrode devices grant a reliable solution for the latter issue but present high operating budget and limitations related to the lack of sensitivity as well as problems related to electrodes maintenance. Optical sensors are widely used in municipal wastewater treatment systems and constitute a reliable technique to acquire instant and accurate measurements through UV

* Correspondence to: ITIS Software, University of Málaga, 29071 Málaga, Spain.
E-mail addresses: chaves@uma.es (A.J. Chaves), cristian@uma.es (C. Martín), lmllopis@uma.es (L. Llopis Torres), mdiaz@uma.es (M. Díaz), jaimeortega@uma.es (J. Fernández-Ortega), jabarbera@uma.es (J.A. Barberá), andreo@uma.es (B. Andreo).

https://doi.org/10.1016/j.jocs.2024.102522
Received 4 July 2024; Received in revised form 11 September 2024; Accepted 30 December 2024
Available online 3 January 2025

spectrometry [10]. However, these costly devices are prone to suffer multiple interferences in environments with high turbidity or organic matter levels. As the world becomes increasingly interconnected and data-driven, the demand for continuous, accurate, and economical process monitoring has grown exponentially. Classical configuration of monitoring stations is based on capture point control only, which severely limits the ability to respond to unexpected upstream changes. In addition, the response focuses on the controlled variable, without considering the effect on the entire process. This demand has driven the development of innovative monitoring technologies, including soft sensors [11]. The introduction of data-driven and Machine Learning (ML) based approaches in the development of control systems aims to overcome the limitations linked to the lack of flexibility of traditional methods [12].

Soft sensors provide a flexible and adaptable framework for monitoring solute concentrations in different environmental settings. Unlike traditional sensors that rely on direct chemical analysis, soft sensors utilise data-driven models to predict nitrate concentrations from a set of input variables [13]. These models can incorporate a wide range of environmental variables, including physical and chemical parameters, weather conditions, and historical data, enabling real-time monitoring and proactive management of nitrate levels, through much more economical and scalable procedures than conventional real-time NO_3^- sensors.

This research article explores the role of soft sensors for the continuous monitoring and real-time control of water resources by applying a soft-sensor development methodology [14] for predicting nitrate concentrations. The applied methodology takes advantage of the benefits of the Kafka-ML platform in connectivity with IoT devices thanks to its direct integration with data streams. IoT devices potentially send sensor measurements through data streams, meeting IoT needs. Kafka-ML [15] is an open-source framework designed to manage the lifecycle of ML/AI applications using data streams on scalable platforms through a user-friendly web interface. Thanks to this novel methodology, a machine learning model can be developed, trained, and used to obtain predictions directly from data streams with the ability to visualise them on the same platform. The methodology, now enhanced by providing support for new IoT protocols such MQTT, along with a multi-disciplinary framework for the rapid development of soft sensors, from their conception to control and visualisation, is presented. Therefore, the development of soft sensors for hydrological and environmental applications is easily achievable. In this work, a real use case of monitoring nitrate concentrations in the surface waters from Guadalhorce river and the groundwater feeding the surface flows in Málaga province (S Spain) has been carried out.

The main contributions of this study are defined as follows:

- A scalable platform for deploying data stream-based IoT soft sensors.
- An experimental case study involving the development of the soft sensor hardware prototype and the utilisation of the methodology outlined in the article.
- Analysis of the versatility of Kafka-ML for soft sensor development.
- Evaluation of the results of NO_3^- level inference from the soft sensor
- Assessment of the scalability of deploying soft sensors using the Kafka-ML-based methodology.

Hereafter, the paper structure is described: Section 2 presents the related work and the potential differences with proposed approach. Section 3 briefly describes the Kafka-ML tool and the soft sensor development methodology. Section 4 shows an application of the methodology by using a NO_3^- monitoring use case. Section 5 provides an evaluation of several key elements of the methodology. Section 6 reflects the implications of this work's findings and potential limitations of the proposed methodology. Finally, Section 7 concludes the work, including a discussion on potential future developments.

2. Related work

Many works have researched on nitrate concentration forecasting based on data techniques. In this section, the related work on nitrate measurement techniques is explored, emphasising the significance of accurate nitrate concentration data and delving into the emerging importance of soft sensors in revolutionising the field. The focus is on exploring the different methodologies employed for nitrate detection and the role that soft sensors play in enhancing the ability to monitor and manage nitrate levels effectively.

Paepae et al. [11] explored the viability of using virtual sensors for real-time monitoring of surface and groundwater quality, with a specific focus on irrigation purposes. It discusses the growing concerns of water pollution due to urbanisation, industrial development, and climate change, highlighting the need for rapid and cost-effective water quality assessment. The review also spotlights the evolution of water quality monitoring from conventional to Internet of Things-based approaches and provides insights into key parameters for irrigation water quality assessment. Finally, it explores the design principles of virtual sensors, evaluating machine learning techniques for inferential modelling and highlighting the importance of a comprehensive virtual sensing system in an IoT environment. The study concludes by pointing out the potential of deep learning techniques in improving virtual sensing for water quality assessment and emphasises the importance of future research in implementing smart monitoring solutions to complement traditional approaches.

In a separate review, Haimi et al. [16] presents a comprehensive analysis of remote sensing techniques applied in biological wastewater treatment plants, focusing on their full-scale applications. The authors underline the potential and challenges of soft sensors which have proven to be efficient and cost-effective in extracting and modelling crucial process information. They emphasise the ability to analyse hard-to-measure primary variables (e.g. concentrations of ammonia, nitrates, and total nitrogen) and process diagnostics in real-time enabling advanced control and optimisation strategies to improve environmental compliance, safety and cost-effectiveness. The researchers underscore the growing popularity of data-derived soft sensors, highlighting multivariate statistical methods, artificial neural networks, and hybrid approaches as common techniques for soft sensor development in wastewater treatment, offering valuable monitoring and backup capabilities in the face of instrument downtime.

In [17], Zare et al. introduced the use of artificial neural networks to model groundwater nitrate levels in the Arak Aquifer of Iran. The study compares artificial neural network (ANN) and linear regression methods for predicting nitrate concentrations based on various water quality indices and finds that ANN outperforms LR, offering higher accuracy with fewer parameters. Regarding its implications for sustainable water management in arid and semi-arid regions, this study highlights the importance of monitoring water quality and the potential of ANNs as a valuable tool for predicting water pollution levels.

Corona et al. [18] focused on the design and implementation of data-derived soft sensors to estimate nitrate concentrations in the post-denitrification filter unit of the Viikinmäki wastewater treatment plant in Helsinki, Finland. These soft sensors were designed to complement existing hardware analysers and provide a reliable backup system in case of sensor malfunction. The article discusses the various stages of soft sensor development, from preprocessing data to calibration and performance evaluation. The soft sensors were shown to accurately estimate nitrate concentrations and support the existing instrumentation, offering potential benefits for wastewater treatment plant monitoring and supervision. The study emphasises the feasibility and practicality of implementing these soft sensors in the plant's control system, offering a cost-effective solution for backup and validation of analytical measurements in harsh wastewater treatment environments.

In another work [19], researchers investigated the use of Random Forest and eXtreme Gradient Boosting (XGBoost) algorithms to model

nitrate concentration in surface water bodies within the context of water scarcity areas and high surface–groundwater interactions, focusing on the Júcar River Basin (RB) in Spain. The article emphasises the importance of combining data-driven methods with local knowledge to enhance model performance and explores the complex relationships between various influencing factors. These models demonstrated high accuracy in predicting nitrate concentrations, outperforming traditional hydrological models. The work highlights the potential of ML techniques for improving water quality management and identifying pollution risk zones, especially in regions facing water quality challenges.

Related to soft sensors and data streams integration, Wang et al. [20] addressed the challenge of quickly process data streams generated from industrial processes to predict quality indicator variables in real time. Traditional soft sensor models face limitations when process states change, and the data streams exhibit nonlinearity, time-variability, and label scarcity. In order to overcome these model constraints, the author propose an online-dynamic-clustering-based soft sensor (ODCSS) for semi-supervised data streams. It employs online dynamic clustering for process state identification, adaptive switching prediction to handle gradual and abrupt changes, and semi-supervised learning to expand labelled training data. The method effectively deals with nonlinearity, time variability, and label scarcity in industrial data streaming environments. The results from two case studies demonstrate the superiority of ODCSS over conventional soft sensors in semi-supervised data stream scenarios, achieving high-precision real-time predictions with limited labelled samples and pseudo-labelled data.

In a recent investigation, Online Deep Evolving Fuzzy System (ODEFS) was presented [21]. ODEFS is an adaptive soft sensor method for industrial data streams, particularly focusing on improving quality monitoring in industries such as chemical, petroleum, and steelmaking. Authors address the challenge of adapting deep learning models to data streams with evolving characteristics and limited pre-training data. The proposed ODEFS method is described, emphasising its two-layer architecture: a continuous learning feature network based on Quality-Related Stacked Autoencoder (QSAE) and a shallow prediction network with evolving fuzzy system capabilities. After an exhaustive evaluation, they confirm the effectiveness of ODEFS in handling nonlinear, time-varying features in industrial data streams and provides application results from the TE process as evidence of its performance.

Hosseinpoor et al. [22] presented an industrial virtual sensor for fault detection in induction motors, focusing on the diagnosis of broken rotor bars. The algorithm utilises an ensemble-learning soft-sensor approach with a novel drift detection mechanism to adapt to changing data distributions. The virtual sensor includes data collection, signal processing, and an ensemble classifier, which is equipped with a concept change detection mechanism to enhance fault diagnosis accuracy in dynamic environments.

Related to air pollution caused by vehicles, [23] proposed an approach that involves leveraging the OBD-II interface present in most vehicles to estimate CO_2 emissions using soft sensor techniques and edge processing hardware. The study emphasises the critical role of urban traffic in CO_2 emissions and suggests IoT-based solutions to monitor and analyse vehicular emissions for better air quality control. The innovative approach involves real-time computation of emissions and employs TinyML to enhance accuracy by handling noisy data. The experimental results show promise, affirming the feasibility and practicality of the proposed solution.

To sum up, while extensive research exists in nitrate concentration prediction and soft sensor development across various domains, no current work focuses on neural network applications for this task, particularly utilising data streams. Furthermore, there is a lack of well-defined methodologies for integrating soft sensors with data streams, except for the one proposed by [14]. Based on this methodology, the following sections describe a real application of soft sensor for nitrate concentration prediction and the challenges and solutions faced given the methodology.

3. Materials and methods

This section discusses the methodology proposed in [14] to create soft sensors with Kafka-ML, a framework designed to apply machine learning (ML) in streaming data applications. Soft sensors may predict process variables by using correlated data, providing a cost-effective alternative compared to standard sensors. Thanks to the connection with Apache Kafka, data can be processed and analysed on a large scale in real time, making it suitable for high-volume data applications.

3.1. Kafka-ML soft sensor development methodology

Kafka-ML [15] is an open source framework designed to orchestrate ML pipelines within Kubernetes infrastructures. This versatile tool facilitates the entire ML model lifecycle, comprising the model design, training, and inference. Apache Kafka is used as the primary data source for data streams in Kafka-ML. This approach not only promotes scalability and fault tolerance but also accommodates distributed models [24].

Given Kafka-ML's intrinsic compatibility with data streams for both training and inference, it emerges as a compelling solution for managing real-time data streams characteristic of IoT sensing. This adaptability extends to the seamless integration of machine learning models, including provisions for GPU-accelerated [25] or federated learning [26], thus empowering decision-making within IoT systems. In this context, Kafka-ML comes up as a formidable tool for soft sensor design and development.

As discussed in a previous study [14], there are many works about soft sensors, each applying different tools. However, it was noted that a complete and unified platform to support the development of soft sensors has not been developed. To address this gap, the proposed methodology has been used to define a soft sensor adapted to the use case of nitrate concentration monitoring from a collection of springs in the Ronda area, Málaga, Spain.

In this methodology, the development of soft sensors is achieved through a four-step process that leverages the Kafka-ML framework for real-time machine learning. It begins with data gathering, preprocessing, and integration, where the training dataset is streamed into Apache Kafka for training and validation phases. Next, the model is selected and trained, with Kafka-ML providing compatibility with popular machine learning frameworks like TensorFlow and PyTorch. The trained model is then deployed with a direct connection to the Kafka pipeline for real-time inference, enabling the soft sensor to provide continuous estimations. Finally, the methodology emphasises on continuous monitoring and maintenance by the real-time soft sensor predictions visualisation, ensuring that the measurements from the soft sensor remain accurate and reliable as possible. This comprehensive approach allows for seamless integration with IoT systems and its scalable deployment, supporting the dynamic needs of industrial applications.

Fig. 1 shows the steps of this methodology for the development of soft sensors.

3.2. IoT device prototyping

An IoT hardware prototype with low-power consumption was considered, ensuring that in cases where there is no unlimited power source, a battery will last as long as possible. Therefore, these devices are normally able to only read the sensor data, carry out some limited data preprocessing, and send the data as the device allows to do so. In this use case, the IoT device is an Arduino MKR NB 1500 as processing unit. It has connected an Adafruit temperature sensor, a DFRobot pH sensor and a DFRobot conductivity sensor. These sensors can operate at 3.3 v so they can be connected directly to the board. These sensors have been used due to their affordability and the wealth of information available on these variables in the dataset. By using components worth about 300 euros (see Table 1), the proposed solution

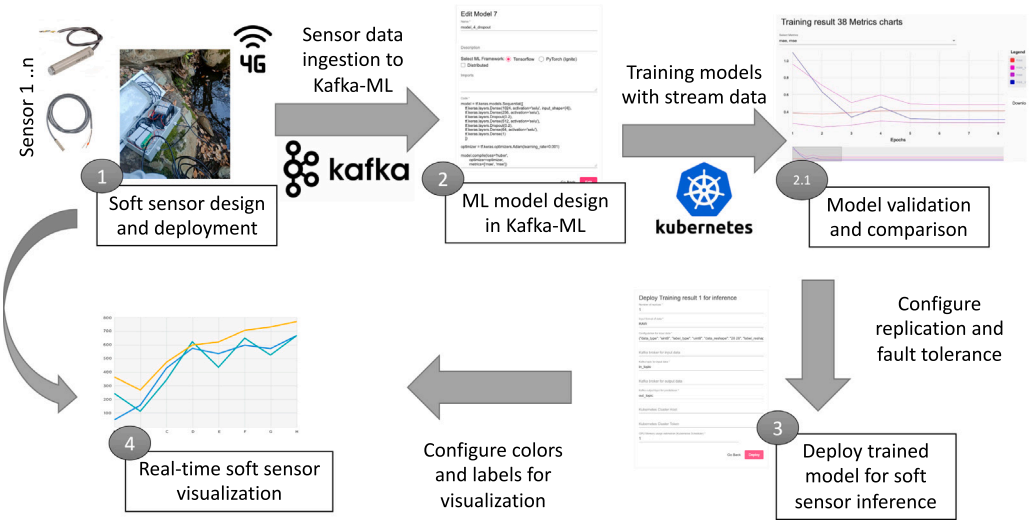

Fig. 1. Data pipeline proposed at the methodology.

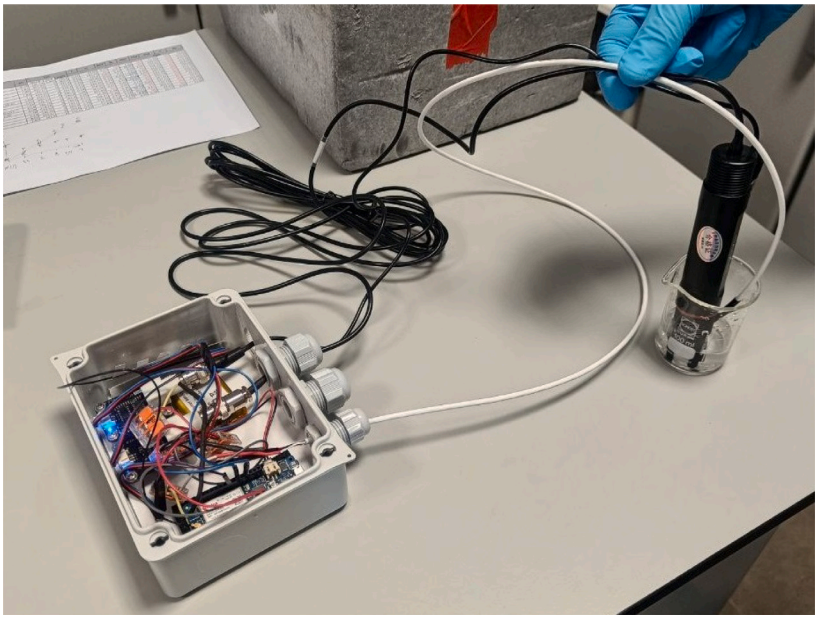

Fig. 2. Soft sensor development and test at laboratory.

Table 1
Prototype development costs.

Product	Description	Price
Arduino MKR NB 1500	Development Board	92.40 €
DFRobot EC Meter (DFR0300)	EC Sensor	65.33 €
DFRobot pH Meter Pro V2	pH Sensor	64.90 €
Adafruit 642 (DS18B20)	Temp Sensor	19.95 €
Taoglas FXUB63.07.0150C	Multiprotocol Antenna	11.72 €
Enclosure, tape, screws, etc.	Crafting stuff	37.56 €
Prototype	**Total Cost**	**291,86 €**

is affordable as the price of deploying one nitrate water probe can run into thousands of euros. This cost difference highlights the affordability and accessibility of the proposed solution, making it desirable for a variety of applications or for deploying different replicas. Fig. 2 shows an overview of the IoT prototype and sensors at the laboratory.

Section 4 describes the implementation of this methodology to monitor nitrate concentrations, serving as a clear guide for implementing soft sensors in real-world applications.

4. A methodology for the development of soft sensors with Kafka-ML

In this section, the methodology for the development of a soft sensor for NO_3^- concentration is described, from sensor selection to the real-time prediction visualisation, considering the details that differ from a theoretical application to a real one. The architecture that supports the seamless integration of soft sensors with the Kafka-ML platform is detailed.

This architecture is the backbone of this work and enables the connection of any type of soft sensor with IoT connectivity to Kafka-ML for the management of its information as data streams, allowing a scalable management of information in real-time. Fig. 3 shows the scheme of the architecture proposed for this use case, showing also the technologies that have been used for its implementation. Each applied technology and its usefulness are described in more detail in Section 4.3.

To provide a deeper understanding, the key phases of soft sensor development with Kafka-ML are described, starting with the data ingestion, followed by model validation and comparison. Afterwards, model

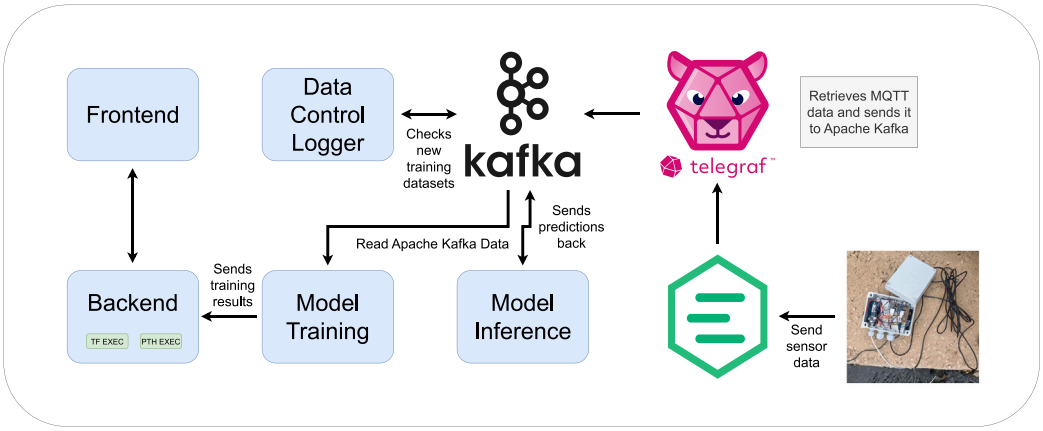

Fig. 3. Microservices architecture deployment proposal.

deployment for inference is presented, to finally show how the soft sensor predictions can be visualised, enabling real-time monitoring of the soft sensor behaviour with Kafka-ML.

4.1. Data ingestion to Kafka-ML

To detect nitrate concentrations from related variables, a data-driven model has been developed. In this case, a dataset containing 13 different physico-chemical parameters (Electrical Conductivity (EC), Temperature (T), pH, Total Organic Carbon (TOC), Alkalinity, F^-, Cl^-, SO_4^{2-}, Na^+, K^+, Ca^{2+}, Mg^{2+} and NO_3^-) from a karst spring database in the Eastern Ronda Mountains (NW Málaga province, S Spain), is used [27]. The dataset includes a complete and diverse representation of groundwater samples, concentrated in several carbonate karst aquifers in a study area of approximately 100 km^2. These parameters are particularly useful for analysing nitrate concentrations, with a focus on their natural origin. However, the dataset also includes water samples in which anthropogenic influence has contributed to enrich nitrate concentrations in groundwater, often from sources such as fertilisers, agricultural run-off, animal waste, etc. Among the variables found in the dataset, those which might be acquired through sensors available at a suitable price were selected. These are temperature, pH, and conductivity.

The data stream ingestion is done through MQTT due to the limitations of using Apache Kafka in Arduino (Kafka is not well prepared to connect multiple IoT devices). This decision required to make some adaptations to the methodology during the data transmission stage. These modifications, further detailed in Section 4.3, have proven to be successful in ensuring seamless communication between the device and Kafka-ML. The code running on the Arduino board and Kafka-ML's source code are available on GitHub.[1,2]

4.2. Model validation and comparison

The next step in the methodology is to define and train models on ingested data streams. Kafka-ML allows users to define multiple models and make them available for training in parallel. This facilitates continuous evaluation of different models, hyperparameters, and architectures. In a soft sensor design context, it is possible to create multiple models to compare their performance with data streams previously received from physical sensors.

Defining models in Kafka-ML is straightforward, users just need to insert the ML model code in the Kafka-ML Web interface. Once a set of models has been defined, a configuration (a set of models to be trained with the same data) will be created. Then, models can be deployed with corresponding training parameters such as batch size and epochs.

Fig. 4 shows how trained models are shown in Kafka-ML. Different model architectures have been trained on this data, choosing the one with the best metrics and deploying it for inference in Kafka-ML. Details of the evaluation of these architectures can be found in the evaluation section.

4.3. Model deployment for soft sensor inference

After the ML models have been trained, compared and selected the one that better fits the dataset, the next step is to deploy the models for inference. This process allows specifying the Kafka topics where data is expected to be received and where predictions are to be sent.

The objective was to use Kafka-ML because of its benefits, including scalability and fault tolerance. However, Apache Kafka (the data streaming platform used in Kafka-ML) is not designed to keep multiple connections, such as the case of IoT deployments. In this case, the MQTT protocol is a suitable candidate. To address this issue, InfluxData Telegraf[3] has been used as an interface between the devices and Apache Kafka, so all the information is similarly treated regardless of the device. InfluxData Telegraf provides a plugin-driven agent that offers multiple interfaces to establish connections to a multitude of different kinds of inputs, merging them in this case into a unified endpoint connected to Apache Kafka. This central endpoint facilitates data ingestion and enables data transmission across multiple IoT devices, regardless of their communication constraints. InfluxData Telegraf adapts seamlessly to ingest data from prevalent IoT communication protocols such as MQTT, as well as custom adapters, enhancing their versatility and interoperability within the IoT ecosystem. Therefore, in this redefined architecture InfluxData Telegraf serves as an intermediary layer between MQTT and Apache Kafka.

Until now, Kafka-ML only had support for data received from Apache Kafka, making this a bottleneck when integrating Kafka-ML with most low-performance IoT devices. Thanks to InfluxData Telegraf, data can now be received from a large range of protocols and dumped into Apache Kafka with low response times, opening up Kafka-ML to new protocols.

Once the sensor data from MQTT is available in Apache Kafka through InfluxData Telegraf, they are able to be used in Kafka-ML for model training (if the variable to be predicted is also measured), to perform inference on these data, or even to make them available as a federated dataset for a collaborative training.

[1] Arduino Soft Sensor Source Code: https://github.com/ertis-research/arduino-softsensor-kafkaml.

[2] Kafka-ML Source Code: https://github.com/ertis-research/kafka-ml.

[3] InfluxData Telegraf Website: https://www.influxdata.com/time-series-platform/telegraf/.

Training results

Filter

ID	Model	Training metrics	Validation metrics	Test metrics	Training Time	Status	Last status change	Chart	Inference	Manage	Download
1	Model 1	loss: 1.94231 mae: 2.39554 mse: 9.70203 mape: 72.58092	loss: 1.67308 mae: 2.12081 mse: 7.38284 mape: 59.1249	loss: 1.4481 mae: 1.87307 mse: 6.28555 mape: 23.98942	31.6268	✓	2023-12-18T08:43:25.582258Z				
2	Model 2	loss: 1.958 mae: 2.41204 mse: 9.7862 mape: 74.14869	loss: 1.6944 mae: 2.14328 mse: 7.57556 mape: 62.99157	loss: 1.63288 mae: 2.0897 mse: 6.98574 mape: 26.60321	32.2345	✓	2023-12-18T08:43:26.231847Z				
3	Model 3	loss: 1.96766 mae: 2.42371 mse: 9.83836 mape: 74.25732	loss: 1.69638 mae: 2.141 mse: 7.64514 mape: 63.81343	loss: 1.82124 mae: 2.28128 mse: 7.42241 mape: 27.31394	34.4811	✓	2023-12-18T08:43:28.432313Z				
4	Model 4	loss: 1.95267 mae: 2.40459 mse: 9.83035 mape: 73.05994	loss: 1.66527 mae: 2.10979 mse: 7.37536 mape: 59.37778	loss: 1.41217 mae: 1.82092 mse: 6.24327 mape: 22.81404	31.6036	✓	2023-12-18T08:43:25.466202Z				
5	Model 5	loss: 1.98444 mae: 2.44257 mse: 10.00218 mape: 75.95998	loss: 1.72589 mae: 2.17467 mse: 7.82802 mape: 66.13731	loss: 1.66371 mae: 2.1117 mse: 7.06863 mape: 26.86691	35.8356	✓	2023-12-18T08:31:48.748429Z				
6	Model 6	loss: 1.98547 mae: 2.43491 mse: 10.1005 mape: 75.36623	loss: 1.67459 mae: 2.11647 mse: 7.458 mape: 59.76976	loss: 1.35186 mae: 1.77196 mse: 5.98229 mape: 21.76204	33.1325	✓	2023-12-18T08:31:46.023343Z				

Fig. 4. Model result list in Kafka-ML.

Fig. 5. IoT prototype deployment for soft sensor in the Guadalhorce river (Málaga, S Spain).

4.4. Soft sensor visualisation

Once all the previous phases has been configured, users can deploy the soft sensor physically. Fig. 5 shows the prototype deployed during a field test at the Guadalhorce river.

An additional step for users who have a trained and deployed model is to visualise the output of a soft sensor over time. Normally, this involves creating a custom API to display predictions from the soft sensor in a web interface. Kafka-ML accommodates this need by allowing fast deployment of visualisation tools. Users can visualise the soft sensor predictions in real time, as illustrated in Fig. 6, which shows an example of visualised predictions from one of the models used in the evaluation.

To use this feature, users need to set the output topic from their inference, and can also customise some design elements like the plot colour or the number of outputs. Once connected, the results from the soft sensor will appear as shown in the example. It is important to note that this visualisation tool only displays real-time data. It has been designed for fast data visualisation, allowing users to avoid leaving Kafka-ML to use external tools like Grafana [28] for real-time visualisation.

5. Evaluation

In this section, some relevant elements of the proof of concept and the architecture designed are evaluated. Already in the previous work [14], some aspects of the architecture such as response time to multiple clients as well as communication overhead were evaluated. In this evaluation, different model architectures are trained to select the best suited to the dataset (Section 5.1). After selecting the appropriate model, a message overload test is performed on the soft sensor with real and simulated data (Section 5.2), looking for the platform's limits with InfluxData Telegraf's integration. In addition, latencies between the prototype and Kafka-ML inference are measured. Finally, some preliminary results from the usage of the soft sensor in the field and laboratory are shown (Section 5.3).

For the evaluation of the soft sensor performance in Kafka-ML, the framework has been deployed on a cluster architecture with the following computation capacities:

- **Hardware configuration**. Kafka-ML has been deployed on a cluster of 7 state-of-the-art servers. Each server has an Intel(R) Xeon(R) Gold 6230R CPU with two NVIDIA(R) Tesla(R) V100 GPUs as well as 384 GB of RAM.

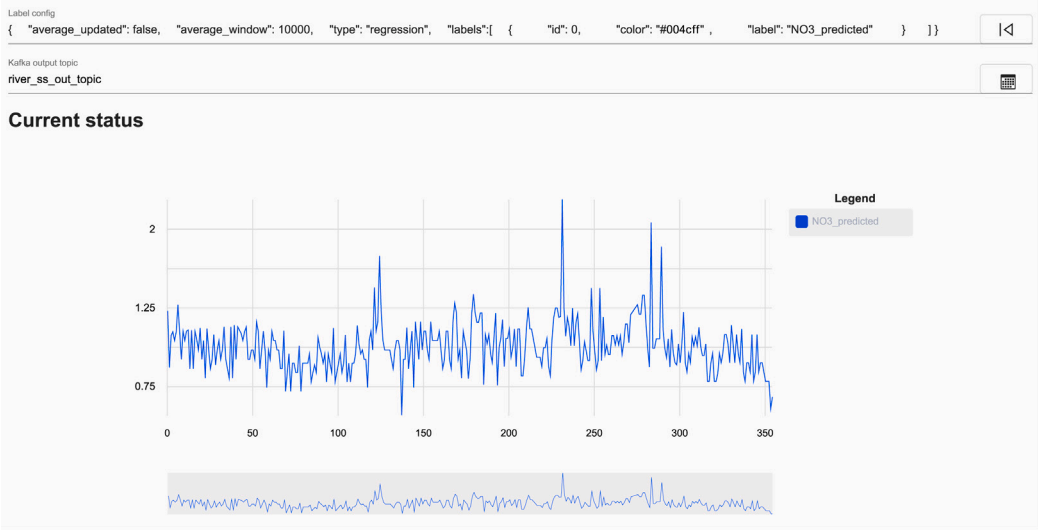

Fig. 6. Real-time NO_3^- contents soft sensor prediction visualisation in Kafka-ML.

Table 2
Models architectures used during evaluation.

#	Architecture
Model 1	512 × 128 × 64
Model 2	256 × 64 × 32
Model 3	256 × 128 × 64 × 32
Model 4	128 × 64 × 32
Model 5	512 × Dropout(0.2) × 128 × 64
Model 6	256 × Dropout(0.2) × 64 × 32

- **Software configuration**. Each node runs Kubernetes v1.21.6 and Docker 20.10.8 on Ubuntu 20.04.3 LTS. A Kubernetes master was installed on one node and the others functioned as Kubernetes secondary nodes.

5.1. Evaluation of machine learning models with Eastern Ronda Mountains dataset

Six different model architectures have been evaluated, each with a different number of layers and neurons per layer. In addition, some of these have been modified with Dropout layers. Models have been trained using Huber loss as loss function, Adam optimiser, and Scaled Exponential Linear Units (SELU) as activation functions in hidden layers. Table 2 shows the different model architectures evaluated.

The models have been defined in Kafka-ML, stacked in a configuration and deployed across the different nodes with the same training parameters. These have been trained for 32 epochs using a 8-sample batch size. Mean Absolute Error (MAE) and Mean Squared Error (MSE) have been used as metrics to benchmark their performance. Fig. 7 shows the MAE decrease for the different models and Fig. 8, the corresponding for the MSE.

As can be appreciated in the MAE and MSE metrics, all of the models have similar metrics, so `model_4` has been selected as the best given its low metrics fluctuation and its lower error rate. Once selected `model_4` as the baseline model, the inference experiments proceed using this model.

In order to assess the model's performance, MAE (1.55 mg/L) and MSE (5.60 mg/L) are compared to the dataset's statistical properties, which shows a mean concentration value of 8.15 mg/L (±6.63 mg/L). MAE of computed results is around 19% of the standard deviation and 24% of the average NO3- value, suggesting that the model reasonably forecasts nitrate concentrations over a wide range of concentrations. This is further supported by the MSE, which indicates that even at low concentrations, the model is able to handle minor variations. These findings imply that the model could be potentially applied to real-world scenarios –such as drinking water quality monitoring– given its versatility to work with different solute concentration ranges. However, despite obtaining an acceptable error for a preliminary application of the model, further development of the tool is necessary to improve its precision, especially in high ranges (>50 mg/L), and thus expand its field of application.

5.2. Scalability assessment of the proposal using real and simulated data

For evaluating inference performance, the response time of the inference service has been benchmarked through several experiments employing different configurations. These experiments have involved deploying the service with a varying number of clients transmitting sensor data for inference. In addition, the number of Kafka replicas and partitions has been manipulated.

These performance benchmarks are relevant, notably in cases where users need high availability of predictive models due to the deployment of multiple sensors in a measurement area.

Different setups have been created to control the availability and scalability of the system, controlling scalability by increasing the number of partitions and fault tolerance by increasing the number of replicas. The utilised setups are the following ones:

- One un-partitioned topic with the model replicated once.
- One topic with two partitions and the model replicated twice.
- One topic with four partitions and the model replicated four times.
- One topic with eight partitions and the model replicated eight times.

In each experiment, 512 sensor data messages have been sent and the prediction time of the model has been measured. This process has been repeated 32 times to obtain an average result. The tests have covered various inference replication configurations, Kafka topic partitions and they have been repeated with different numbers of concurrent clients (namely 1, 2, 4, 8, 16 and 32 clients), exploring the behaviour of the system under various request loads.

The standard Kafka-ML use case involves a single-issue, single-partition configuration without replicating the inference module. Fig. 9 illustrates that this configuration works adequately with a few clients, but experiences overhead and a latency increase as the number of clients increases.

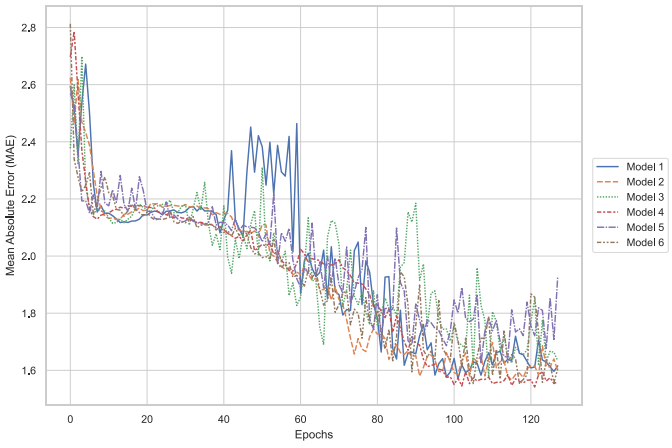

Fig. 7. Validation MAE decrease of the evaluated models during training.

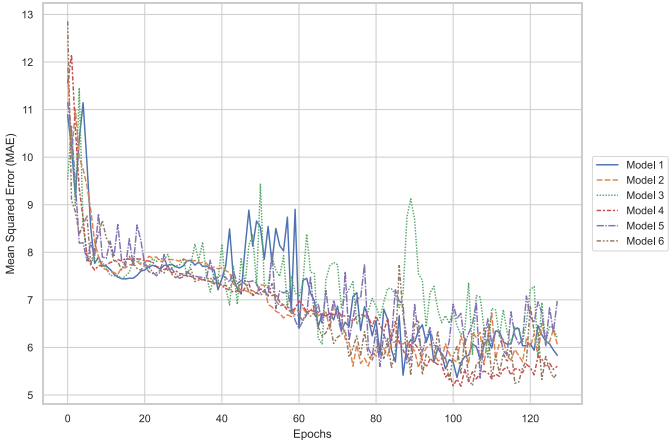

Fig. 8. Validation MSE decrease of the evaluated models during training.

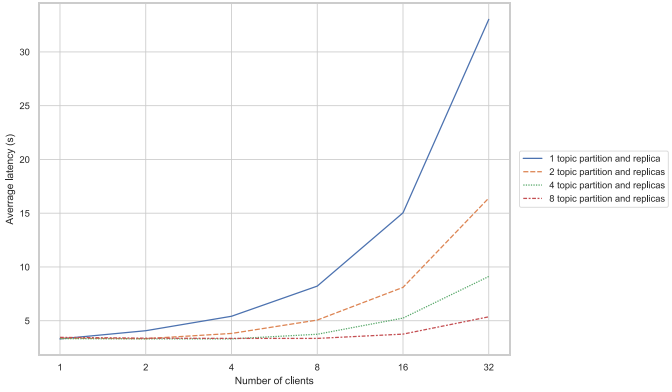

Fig. 9. Average inference latency response with different number of clients, topic partitions, and replicas.

Increasing the number of model replicas and distributing the data widely in Kafka-ML improves these results. Especially in cases requiring higher availability, lower latency is observed (except when replication and partitioning have reached their limits). Therefore, it can be concluded that increasing replication and partitioning can mitigate latency in scenarios that require higher availability.

Another experiment has been carried out to evaluate the performance of the methodology and Kafka-ML across different data transmission frequencies. In this experiment, simulated IoT devices has

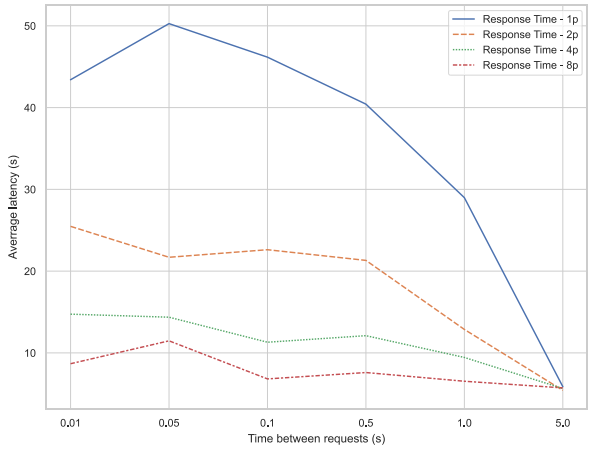

Fig. 10. Responsiveness of Kafka-ML towards data transmission at different time intervals at different scenarios.

generated data at specific time intervals in order to evaluate the responsiveness of Kafka-ML and to identify the time threshold at which the predictive capabilities are no longer limited.

In this experiment, once again four different scenarios has been considered, depending on the number of partitions and replicas deployed. This scenario has also considered different measurement times for the soft sensor with different number of replicas.

The results (depicted in Fig. 10) indicate that introducing replicas significantly improves performance, especially with lower communication latencies. With longer time intervals (5 s) it is shown that replication does not provide any improvement, but requiring higher resource allocation needs. Therefore, the final decision on replication will depend on the specific communication requirements of the resulting application.

5.3. Correlation of the soft sensors with laboratory data

Laboratory experiments and field tests with the prototype device have been also conducted. The laboratory tests are intended to validate the performance parameters obtained a in controlled environment. Thus, the accuracy of the model has been monitored under various conditions, as well as the inference time by replicating scenarios that emulated real-world deployments.

Subsequently, the evaluation has been extended to a real-world scenario by conducting tests on a river. Deploying the prototype in a real environment has allowed to assess its robustness and performance in dynamic, unpredictable conditions. During these river tests, the system has processed live sensor data acquired from the surface waters and groundwater from Guadalhorce river and associated porous aquifer, predicting the NO_3^- concentration given the data written at the MQTT broker by the prototype device.

Tables 3 and 4 show the results obtained by the sensors and the model as well as the comparison with the laboratory data.

As can be seen in Tables 3 and 4, due to the use of measurements of different nature (trained with a dataset from Eastern Ronda Mountains and tested with river water data in southern Spain) as well as the use of samples with disturbed conditions at the laboratory, the predictions obtained do not have an acceptable accuracy. This matter is discussed afterwards.

6. Discussion

The methodology adopted in this study [14] has demonstrated exceptional technical advantages in the soft sensor development

Table 3
Laboratory samples data comparison.

		T	CE	pH	NO_3^-
Milli-Q[a]	Laboratory	24.1	0	6.998	0
	Soft Sensor	23.774	0.509	7.035	9.603
Site 11	Laboratory	17	NA	7.7	5.59
	Soft Sensor	16.755	11.512	7.037	0
Site 2	Laboratory	15.9	NA	8.05	8.94
	Soft Sensor	15.333	11.809	7.032	0
Site 8	Laboratory	16	NA	7.87	15.97
	Soft Sensor	15.161	15.424	7.026	0
Site 14	Laboratory	15.8	NA	7.83	34
	Soft Sensor	15.024	10.519	7.04	0
Site 4	Laboratory	15.8	NA	7.78	56.72
	Soft Sensor	15.096	11.555	7.026	0

[a] Milli-Q water is ultra-pure water processed through a Millipore purification system, removing impurities for laboratory and research applications.

Table 4
River samples data comparison.

		T	CE	pH	NO_3^-
Site1	Laboratory	23	2.05	7.52	≈ 0
	Soft Sensor	22.813	0.484	7.026	10.277
Site2	Laboratory	23.3	0.75	7.37	≈ 0
	Soft Sensor	23.125	0.782	7.0446	5.227
Site3	Laboratory	23.3	2.97	7.96	3.21
	Soft Sensor	22.625	2.9558	7.045	3.139

domain, being acceptably applied to a NO_3^- real-time monitoring use case. Using Kafka-ML allows users to manage the pipeline of their ML models using data streams efficiently. The advantages are the rapid development of soft sensors, low response times, and scalability. Moreover, thanks to the possibility of using low-cost devices, it is possible to deploy multiple soft sensors and monitor a wider area, managing the workload with ease. The real-world application of this methodology highlights its efficacy as a feasible approach to environmental monitoring.

By combining controlled laboratory experiments with real tests on the river, efforts have been made to validate the prototype's performance under various conditions, trying to ensure its efficiency and reliability in real deployment scenarios. The results of these real tests provide valuable information about the adaptability of the system and its suitability in practical applications.

During the evaluation stage, the methodology's operation as well as the communications performance has been verified, obtaining low response times from the sensor sampling until the data is predicted on the server. Problems with the pH sensor, along with significant differences between the conditions in the used dataset and evaluated environments, raise concerns about the direct applicability of existing data to different environments. However, our main focus was to demonstrate the feasibility of Kafka-ML and the methodology in real environments by using a self-developed prototype as an example, being the main benefits clearly identified.

To address the accuracy issue, a proposal has been made to create a dedicated dataset for the surface waters and groundwater from the Guadalhorce River basin. Given the unique hydrochemical behaviour of each ecosystem, site-specific training is crucial. This dataset would be used to retrain the soft sensor through transfer learning, providing an opportunity to learn from existing data while adapting to the unique characteristics of the Guadalhorce River ecosystem.

7. Conclusions and future work

In this work, a methodology for the development of soft sensors based on data-streams is presented. Specifically, an attempt has been made to develop a PoC soft sensor for the detection of nitrate concentrations. During the course of the project, various challenges were faced while implementing the methodology, such as the incompatibility of the Arduino boards with the sending of data via Kafka, which was overcome by incorporating the InfluxData Telegraf tool, providing support to new technologies in this methodology.

To show the potential of the redesigned methodology, load tests have been carried out, simulating the connection of a multitude of devices, proving that Kafka-ML and the methodology work in harmony, serving with many devices as desired. A hardware PoC of the soft sensor has also been made, and its behaviour has been evaluated, resulting in a device on which the methodology can be applied.

It is expected that this refined methodology will be applied in future projects in collaboration with water management companies, using higher quality sensor technology as well as a larger number of input parameters, giving the model higher quality.

As future work, the following improvements to the methodology are proposed:

- Data preprocessing and postprocessing. In many machine learning application scenarios, optimising model performance often involves the need for data preprocessing or post-processing. While these tasks are usually performed before feeding the data into the model for training or inference, a valuable feature of Kafka-ML and the proposed methodology lies in the seamless integration of data processing activities and statistical analysis within the workflow. Although InfluxData Telegraf enables data processing, it is not easy to use for non-technical users, so adding a section in Kafka-ML to do this job would be of great importance.
- Pretrained model inclusion. One potential functionality for Kafka-ML is the integration of user-pretrained machine learning models. This functionality has the potential to significantly improve the versatility of the platform, allowing researchers to leverage existing models for various applications and real-time analysis.
- Concept drift detection and correction. In order to enhance the adaptability of Kafka-ML, a concept drift detection mechanism could be incorporated. This enhancement will ensure that the models remains relevant in dynamic soft-sensor environments, responding autonomously to evolving data patterns for more resilient and accurate predictions.
- Improve inference module response time. Although Kafka-ML's inference module has a good performance, it is still improvable. An enhancement would consist in optimising Kafka-ML's inference performance to reduce latency. This effort will contribute to faster and responsive predictions in multiple scenarios, consolidating the effectiveness of Kafka-ML in real-time applications.

CRediT authorship contribution statement

Antonio Jesús Chaves: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Cristian Martín:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Luis Llopis Torres:** Writing – review & editing, Writing – original draft, Supervision, Resources, Methodology, Investigation, Formal analysis, Conceptualization. **Manuel Díaz:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Investigation, Funding acquisition, Conceptualization. **Jaime Fernández-Ortega:** Writing – review & editing, Validation, Investigation, Data curation, Conceptualization. **Juan Antonio Barberá:** Writing – review & editing, Supervision, Investigation, Data curation, Conceptualization. **Bartolomé Andreo:** Writing – review & editing, Validation, Resources, Project administration, Funding acquisition, Conceptualization.

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 work is a contribution to the PRIMA funded European project KARMA (ANR-18-PRIM-0005); financed by the Ministry of Science and Innovation of Spain and FEDER funding inside the Operational Plurirregional Program of Spain 2014–2020 and the Operational Program of Smart Growing (Environmental and Biodiversity Climate Change Lab, EnBiC2-Lab; LIFEWATCH-2019-11-UMA-01-BD); funded by the project PCI2019-103675 of the International Joint Programme of the Ministry of Science, Innovation and Universities of Spain, the project P06-RNM 2161 funded by the Autonomous Government of Andalusia (Spain), the project PLSQ-00230 ('iSAT: Sistema de Alerta Temprana Inteligente') funded by the Autonomous Government of Andalusia (Spain) and to the Research Group RNM-308 funded by the Autonomous Government of Andalusia (Spain). This work is funded by the Spanish projects: Grant PID2022-141705OB-C21 ('DiTaS: A framework for agnostic compositional and cognitive digital twin services') funded by MICIU/AEI/10.13039/5011000110331 and by 'FEDER'; Grant TED2021-130167B ('GEDIER: Application of Digital Twins to more sustainable irrigated farms'), funded by MICIU/AEI/10.13039/5011000110331 and by 'European Union NextGenerationEU/PRTR'; Grant CPP2021-009032 ('ZeroVision: Enabling Zero impact wastewater treatment through Computer Vision and Federated AI') funded by MICIU/AEI/10.13039/5011000110331 and by 'European Union NextGenerationEU/PRTR'. Funding for open access charge: Universidad de Málaga / CBUA.

Data availability

The dataset used in this research has been provided by the Centre of Hydrogeology of the University of Málaga (CEHIUMA). These data are confidential and are not available for public dissemination. Any requests for data access should be directed to CEHIUMA for consideration. Instructions to deploy Kafka-ML and all its code can be found as open-source at Kafka-ML's GitHub Repository (https://github.com/ertis-research/kafka-ml). The source code running on the Arduino board is available on GitHub (https://github.com/ertis-research/arduino-softsensor-kafkaml).

References

[1] D. Postma, C. Boesen, H. Kristiansen, F. Larsen, Nitrate reduction in an unconfined sandy aquifer: water chemistry, reduction processes, and geochemical modeling, Water Resour. Res. 27 (8) (1991) 2027–2045.

[2] L. Knobeloch, B. Salna, A. Hogan, J. Postle, H. Anderson, Blue babies and nitrate-contaminated well water, Environ. Health Perspect. 108 (7) (2000) 675–678.

[3] M.H. Ward, R.R. Jones, J.D. Brender, T.M. De Kok, P.J. Weyer, B.T. Nolan, C.M. Villanueva, S.G. Van Breda, Drinking water nitrate and human health: an updated review, Int. J. Environ. Res. Public Health 15 (7) (2018) 1557.

[4] R. Valenca, H. Le, Y. Zu, T.M. Dittrich, D.C. Tsang, R. Datta, D. Sarkar, S.K. Mohanty, Nitrate removal uncertainty in stormwater control measures: Is the design or climate a culprit? Water Res. 190 (2021) 116781.

[5] S. Panno, K.C. Hackley, H. Hwang, S. Greenberg, I. Krapac, S. Landsberger, D. O'kelly, Characterization and identification of Na-Cl sources in ground water, Groundwater 44 (2) (2006) 176–187.

[6] WHO, Nitrate and nitrite in drinking-water. Background document for development of WHO guidelines for drinking-water quality, 2023, Available from: https://cdn.who.int/media/docs/default-source/wash-documents/wash-chemicals/nitrate-nitrite-background-jan17.pdf?sfvrsn=1c1e1502_4. [last accessed November 2023].

[7] W.H. Organization, et al., Guidelines for drinking-water quality: first addendum to the fourth edition, 2017.

[8] European Parliament and of the Council of the European Union, Directive (EU) 2020/2184, 2020, Directive (EU) 2020/2184 on the quality of water intended for human consumption (recast). Available from: https://eur-lex.europa.eu/eli/dir/2020/2184/oj. [last accessed November 2023].

[9] S. Singh, A.G. Anil, V. Kumar, D. Kapoor, S. Subramanian, J. Singh, P.C. Ramamurthy, Nitrates in the environment: A critical review of their distribution, sensing techniques, ecological effects and remediation, Chemosphere 287 (2022) 131996.

[10] A. Drolc, J. Vrtovšek, Nitrate and nitrite nitrogen determination in waste water using on-line UV spectrometric method, Bioresour. Technol. 101 (11) (2010) 4228–4233.

[11] T. Paepae, P.N. Bokoro, K. Kyamakya, From fully physical to virtual sensing for water quality assessment: A comprehensive review of the relevant state-of-the-art, Sensors 21 (21) (2021) 6971.

[12] M. Lowe, R. Qin, X. Mao, A review on machine learning, artificial intelligence, and smart technology in water treatment and monitoring, Water 14 (9) (2022) 1384.

[13] L.J. Stamenković, S. Mrazovac Kurilić, V. Presburger Ulniković, Prediction of nitrate concentration in Danube River water by using artificial neural networks, Water Supply 20 (6) (2020) 2119–2132.

[14] A.J. Chaves, C. Martín, L.L. Torres, E. Soler, M. Díaz, A methodology for the development of soft sensors with Kafka-ML, in: R. Sharma, G. Jeon, Y. Zhang (Eds.), Data Analytics for Internet of Things Infrastructure, Springer Nature Switzerland, Cham, 2023, pp. 307–324.

[15] C. Martín, P. Langendoerfer, P.S. Zarrin, M. Díaz, B. Rubio, Kafka-ML: Connecting the data stream with ML/AI frameworks, Future Gener. Comput. Syst. 126 (2022) 15–33.

[16] H. Haimi, M. Mulas, F. Corona, R. Vahala, Data-derived soft-sensors for biological wastewater treatment plants: An overview, Environ. Model. Softw. 47 (2013) 88–107.

[17] A. Zare, V. Bayat, A. Daneshkare, Forecasting nitrate concentration in groundwater using artificial neural network and linear regression models, Int. Agrophys. 25 (2) (2011).

[18] F. Corona, M. Mulas, H. Haimi, L. Sundell, M. Heinonen, R. Vahala, Monitoring nitrate concentrations in the denitrifying post-filtration unit of a municipal wastewater treatment plant, J. Process Control 23 (2) (2013) 158–170.

[19] D.Y. Dorado-Guerra, G. Corzo-Pérez, J. Paredes-Arquiola, M.Á. Pérez-Martín, Machine learning models to predict nitrate concentration in a river basin, Environ. Res. Commun. 4 (12) (2023) 125012.

[20] Y. Wang, H. Jin, X. Chen, B. Wang, B. Yang, B. Qian, Online-dynamic-clustering-based soft sensor for industrial semi-supervised data streams, Sensors 23 (3) (2023) 1520.

[21] Y. Gao, H. Jin, B. Wang, B. Yang, W. Yu, An adaptive soft sensor method based on online deep evolving fuzzy system for industrial process data streams, in: 2023 IEEE 12th Data Driven Control and Learning Systems Conference, DDCLS, IEEE, 2023, pp. 1799–1804.

[22] Z. Hosseinpoor, M.M. Arefi, R. Razavi-Far, N. Mozafari, S. Hazbavi, Virtual sensors for fault diagnosis: A case of induction motor broken rotor bar, IEEE Sens. J. 21 (4) (2020) 5044–5051.

[23] P. Andrade, I. Silva, M. Silva, T. Flores, J. Cassiano, D.G. Costa, A tinyml soft-sensor approach for low-cost detection and monitoring of vehicular emissions, Sensors 22 (10) (2022) 3838.

[24] A. Carnero, C. Martín, D.R. Torres, D. Garrido, M. Díaz, B. Rubio, Managing and deploying distributed and deep neural models through Kafka-ML in the cloud-to-things continuum, IEEE Access 9 (2021) 125478–125495.

[25] A.J. Chaves, C. Martín, M. Díaz, The orchestration of machine learning frameworks with data streams and GPU acceleration in Kafka-ML: A deep-learning performance comparative, Expert Syst. (2023) e13287.

[26] A.J. Chaves, C. Martín, M. Díaz, Towards flexible data stream collaboration: Federated learning in Kafka-ML, Internet Things 25 (2024) 101036.

[27] J.A. Barberá, Investigaciones Hidrogeológicas En Los Acuíferos Carbonáticos De La Serranía Oriental De Ronda, Provincia De Málaga (Ph.D. thesis), University of Málaga (Spain), p 661, 2014.

[28] Grafana website, 2024, Available online: https://grafana.com/grafana/, (Accessed 9 May 2024).

Antonio Jesús Chaves received his B.Sc. in Computer Science Engineering and his M.Sc. in Software Engineering and Artificial Intelligence from the University of Málaga, Spain, in 2021 and 2022, respectively. He is currently a Ph.D. student at the ERTIS Research Group, University of Málaga, and a member of the ITIS Software Institute, University of Málaga. His research interests include modern deep-learning techniques, such as object recognition, fault detection, federated learning, and artificial intelligence applied to the IoT field.

Cristian Martín is an Associate Professor at the University of Malaga (UMA), and he is part of the ERTIS research group and the ITIS Software. Cristian Martín obtained a Ph.D. in Computer Science in 2018 at UMA, with an extraordinary Ph.D. thesis award. His research interests are focused on the IoT, machine-learning applied, digital twins, as well as paradigms such as Fog and Edge Computing. Previously he has been working in several technology companies on RFID technology and software development. He has participated in more than 20 research projects and contracts with companies. He has carried out four international stays, one at the University of Ghent, Belgium (2016, pre-doctoral), two at the IHP research institute, Frankfurt Oder, Germany (2020–2021, post-doctoral), and the last one in Incheon, Korea (2022).

Luis Llopis Torres received the M.S. and Ph.D. degrees in computer engineering from the Universidad de Málaga, in 1995 and 2002, respectively. From 1996 to 2008, he was an Assistant Professor with the Department of Languages and Computer Science, Universidad de Málaga, where he has been an Associate Professor since 2008. He has been working in the areas of formal description techniques and real time systems. Currently, he is specially involved in the research fields of wireless sensor and actor networks, and the integration of Internet of Things and cloud computing. He has been a member of the Software Engineering Group of Universidad de Málaga (GISUM) since its foundation and recently became a member of the ITIS Software Institute, Universidad de Málaga.

Manuel Díaz is a University Full Professor in the Department of Languages and Computer Sciences at the University of Malaga, where he directs the ERTIS research group, integrated within the Software Engineering Group of the University of Malaga and from 2016 he is the CEO of Software for Critical Systems, S. L. of which he was co-founder in 2009. Between 1987 and 1991 he worked in the private sector (Olivetti Spain and R&D department of Fujitsu in Malaga). Since 1991 he belongs to the Dept. of Languages and Computer Sciences. His main lines of research focus on distributed systems, real-time embedded systems, and IoT and, more specifically, on the aspects related to middleware for this type of applications and the development of innovative applications in critical systems through the use of disruptive technologies as deep learning and distributed ledger technologies. He has been a principal researcher in 40 research contracts with private companies (Tecnatom, Abengoa, Indra, Adif, ...), 6 National Plan projects and 8 European projects (one of them as coordinator).

Jaime Fernández-Ortega graduated in Geology at the University of Oviedo and Master in Water Resources and Environment (RHYMA) at the University of Malaga. He has realised research stays at several European universities: Università degli Studi di Ferrara (Italy), Universität Freiburg – Albert-Ludwigs (Germany), Université de Montpellier (France) and Karlsruher Institut für Technologie (Germany) to acquire new knowledge and perfect the techniques related to karst groundwater. He is currently pursuing his Ph.D. research as part of the European project "KARMA," which investigates the availability and quality of karstic water resources in the Mediterranean region.

Juan Antonio Barberá received his B.Sc. in Geology from the University of Granada in 2006 and a postgraduate diploma in Groundwater Hydrology from the FCIHS in Barcelona in 2006. He earned his Ph.D. in Hydrogeology from the University of Málaga (UMA) in 2014. Since 2017, he has been part of the External Geodynamics area of the Department of Ecology and Geology at UMA. He is currently Associate Professor, engaged in teaching, research, and knowledge transfer to the public and private sectors. He teaches subjects related with Geology, Hydrology, and Hydrogeology, within the Environmental Sciences undergraduate programme and the official master's programme in Water Resources and Environment (RHYMA). His research focuses on identifying and quantifying hydrogeochemical processes in aquifer media (karstic, fractured, and detrital), studying stable and radioactive isotopes, and addressing water mixing problems (e.g., river-aquifer interactions and managed aquifer recharge). In addition, he actually is the technical responsible of the Advanced Water Analysis Laboratory at the Hydrogeology Centre of the University of Málaga.

Bartolomé Andreo is Full Professor of Hydrogeology at the University of Málaga, of which he is Vice-Chancellor of Academic Planning and Teaching Staff. He has 34 years of teaching and training on water resources, geology and environment. As Director of the CEHIUMA, he has broad national and international experience in water resources, leading or participating in over 50 projects, including 20 international projects, some with institutions such as UNESCO or the International Atomic Energy Agence of which he was advisor Successful supervision of 16 past Ph.D. students, co-author of hundreds of scientific publications, one hundred in peer review international journals. He was President of the Spanish Chapter of the International Association of Hydrogeologists (2015–2021), coordinator of the Master's degree in Water Resources and Environment at the University of Málaga (2008–2023). He was the main organiser of several scientific meetings, acting as President in 5 international congresses, including the 46th IAH Congress (world groundwater congress), with over 800 attendees. Coorganizer of many training courses on groundwater and divulgative activities (hydrogeoday, geoloday, etc.).