M2 Internship + PhD opportunity
Heterogeneous IoT network with low-cost sensors for predicting pollutant concentrations

Aymane Souani[1,2], V. Vigneron[1], H. Maaref[1]
[1]IBISC EA 4526, université Evry-Paris-Saclay, France
[2] ™ECOMESURE, Saclay, France

Oct 2025

Porteurs scientifiques : Aymane Souani, Hichem Maaref et V. Vigneron (IBISC)
Partners : IBISC (université Evry-Paris-Saclay), ™ECOMESURE
Specialized AI and Data Science : machine learning theory, high-dimensional statistics, uncertainty, information theory, generative model
Duration : 5 to 6 months, starting between January and April 2026
Funding : ECOMESURE insternship grant
Location : IBISC lab
Application domain : green tech
Mots-clés : deep learning, time series prediction, weekly supervised training, modality fusion,
Key-words :

1 Context

This internship proposal aims to develop a forecasting system for optimizing the estimation of pollutant concentrations such as $PM_{2.5}$, PM_{10}, NO_2, O_3, CO from local meteorological variables (Temp, hygrometry, Pression, WS) across ™ECOMESURE's proprietary sensor network (™Ecomzen, ™Ecomlite, ™Ecomtreck, ™Ecomsmart).

The historical data warehouse contains over 10^9 observations collected in urban, industrial, and commercial contexts.

™ECOMESURE operates an expanding network of low-cost IoT sensors capable of transmitting, in near real-time (1–5 min), measurements of $PM_{2.5}$, PM_{10}, NO_2, O_3, CO , and micro-meteorological variables to a secure SaaS platform. This dense telemetry already supports hyper-local alerting and reporting services. To transform this massive data stream into actionable intelligence, it is necessary to (*i*) maintain dynamic calibration against noise and drift, (*ii*) fuse these low-cost signals with heterogeneous data sources, and (*iii*) produce reliable multi-horizon forecasts at 24 h, 72 h, and 168

h [1]. Such hyper-local predictions will optimize building ventilation, improve citizen information, and support public policy evaluation.

Position of the problem However, operating such a dense and heterogeneous IoT network raises multiple challenges. Low-cost sensors are prone to bias, temperature–humidity sensitivity, and long-term drift, making regular calibration essential to ensure data reliability. The 1–5 min transmission interval produces high-frequency data streams subject to gaps, outliers, and synchronization issues under communication or power constraints. Moreover, pollutant concentrations exhibit strong spatio-temporal heterogeneity driven by micro-climatic and emission differences across sites, requiring adaptive, non-stationary modeling. At the system level, the secure SaaS platform must ingest and manage large volumes of multimodal telemetry while maintaining scalability and resilience. Finally, hyper-local multi-horizon forecasting under such conditions demands models that can capture complex dependencies, quantify uncertainty, and remain interpretable for decision support and regulatory use.

2 Methods/modeling approach

To address these challenges, we propose a self-supervised learning framework designed to exploit the large volumes of unlabeled data continuously produced by heterogeneous low-cost sensor (LCS) networks. The method performs pre-training on multi-source environmental datasets using masked-sequence reconstruction and contrastive representation learning, enabling the model to capture invariant temporal and cross-variable dependencies across diverse locations and device types [2]. A domain adaptation strategy is then applied to align the latent representations of the pre-trained model with the specific distribution of ™ECOMESURE sensors, minimizing the need for local calibration or labeled data. This transfer process combines adversarial feature alignment with distributional regularization to ensure consistency across pollutant and meteorological modalities. The resulting model can be fine-tuned with minimal supervision to forecast multi-horizon air-quality quantiles, achieving improved generalization under sensor drift and environmental variability. By coupling self-supervised pre-training with robust domain adaptation [3], the proposed approach aims to reduce prediction errors and maximize transferability across the expanding ™ECOMESURE network.

Data Pipeline and Calibration The dataset comprises 12 months of collocated measurements from EcomSmart sensors and Atmo-France reference stations, enabling joint calibration and validation. Raw signals underwent outlier detection, quantile normalization, and temporal fusion at 5-minute intervals to ensure data consistency. An initial neural-network calibration corrected sensor biases and environmental drifts. Subsequently, a multi-platform domain adaptation strategy aligned latent embeddings to stabilize first- and second-order statistics across heterogeneous sensor domains. The resulting forecasting model was distilled into a lightweight, edge-deployable version [4], providing multi-horizon (1–168 h) air-quality predictions across the Ecomesure network.

3 Internship supervision and scientific environment

Candidate profile We look for strongly motivated candidates (i) coming from a math, physics, computer science or engineering diploma (ii) having a strong mathematical background, notably

in linear algebra, analysis, probability and statistics, in machine learning and deep learning (iii) having good programming skills on some scientific language, preferably python.

Knowledge of sensors, particularly prolutant sensors, is not required, but is a strong plus. Knowledge of basic optimization theory is also appreciated.

Practical information The intern will be mainly hosted at the UFR science and technology (40 rue du Pelvoux) close to the city center. However, he/she may also spend some periods at ECOMESURE.

The monthly internship gratification is of about €1000.

Application procedure: send a motivation letter, a CV and your University transcript (relevé de notes) to {vincent.vigneron,hichem.maaref}@univ-evry.fr and ayamane.souani@ecomesure.fr.

What we offer: Hands-on experience with cutting-edge AI techniques for sensor control.
Tackle real-world, high-impact greentech solutions using deep learning.
Close mentorship from experienced researchers at the IBISC laboratory.
Opportunities to co-author publications and present your work at conferences.
Continuation into PhD studies

Contact {vincent.vigneron,hichem.maaref}@univ-evry.fr and ayamane.souani@ecomesure.fr.

References

[1] G. Chen, S. Chen, D. Li, and C. Chen. A hybrid deep learning air pollution prediction approach based on neighborhood selection and spatio-temporal attention. *Scientific Reports*, 15(1), 2025.

[2] C. Malings, K. E. Knowland, N. Pavlovic, J. G. Coughlin, D. King, C. Keller, S. Cohn, and R. V. Martin. Air quality estimation and forecasting via data fusion with uncertainty quantification: Theoretical framework and preliminary results. *Journal of Geophysical Research: Machine Learning and Computation*, 1(4), 2024.

[3] K. Niresi, I. Nejjar, and O. Fink. Efficient unsupervised domain adaptation regression for spatial-temporal air quality sensor fusion, 2024.

[4] P. Wang, H. Zhang, J. Liu, F. Lu, and T. Zhang. Efficient inference of large-scale air quality using a lightweight ensemble predictor. *International Journal of Geographical Information Science*, 39(4):900–924, 2025.