Veli: Unsupervised Method and Unified Benchmark for Low-Cost Air Quality Sensor Correction

Yahia Dalbah[1], **Marcel Worring**[1], **Yen-Chia Hsu**[1]
[1]University of Amsterdam
y.i.r.dalbah@uva.nl, m.worring@uva.nl, y.c.hsu@uva.nl

Abstract

Urban air pollution is a major health crisis causing millions of premature deaths annually, underscoring the urgent need for accurate and scalable monitoring of air quality (AQ). While low-cost sensors (LCS) offer a scalable alternative to expensive reference-grade stations, their readings are affected by drift, calibration errors, and environmental interference. To address these challenges, we introduce **Veli** (Reference-free **V**ariational **E**stimation via **L**atent **I**nference), an unsupervised Bayesian model that leverages variational inference to correct LCS readings without requiring co-location with reference stations, eliminating a major deployment barrier. Specifically, Veli constructs a disentangled representation of the LCS readings, effectively separating the true pollutant reading from the sensor noise. To build our model and address the lack of standardized benchmarks in AQ monitoring, we also introduce the Air Quality Sensor Data Repository (AQ-SDR). AQ-SDR is the largest AQ sensor benchmark to date, with readings from 23,737 LCS and reference stations across multiple regions. Veli demonstrates strong generalization across both in-distribution and out-of-distribution settings, effectively handling sensor drift and erratic sensor behavior. Appendices are available in the extended version.

Code — https://github.com/YahiDar/Veli
Datasets — https://github.com/YahiDar/AQ-SDR
Extended version — https://arxiv.org/abs/2508.02724

1 Introduction

The World Health Organization (WHO) estimated that over 90% of the world's population breathes air that contains pollutants above WHO guideline levels (World Health Organization 2018). These pollutants are known to cause respiratory and cardiovascular diseases, and are present in high concentrations in urban areas (Zhang et al. 2024). To meet WHO air quality standards, real-time air quality (AQ) monitoring is crucial.

Municipalities and environmental agencies rely on well-maintained, expensive monitoring stations to report pollution at the district level. The high cost of buying, installing, and maintaining these stations makes it infeasible to achieve the spatial coverage needed to capture microclimates affecting citizens. Consequently, numerous initiatives have emerged to scale up the spatial coverage of AQ sensing by using low-cost sensors (LCS). In contrast to expensive monitoring stations, LCS are affordable and accessible to the average citizen, making them suitable for crowdsourcing projects. However, LCS produce raw data that are inaccurate, noisy, and often unreliable, making it difficult to use their readings to make informed decisions.

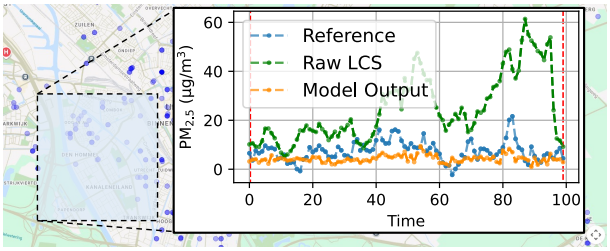

Figure 1: A snapshot from the AQ-SDR dashboard of sensors in the city of Utrecht in the Netherlands. The query area shows the results of applying our method, Veli, on hourly noisy readings from deployed LCS over four days.

To use LCS to increase the spatial coverage of AQ monitoring, reliable methods for correcting their erratic readings are necessary. Many pre-deployment calibration methods exist for LCS (Delaine, Lebental, and Rivano 2019; Hagan et al. 2018; Maag et al. 2016). However, dense deployment of LCS would require recurrent manual recalibration to prevent issues like sensor drift. To eliminate the need for manual recalibration, numerous works have explored numerical approaches for post-deployment data correction[1]. LCS data correction methods often rely on high-cost reference stations as the ground truth to train supervised machine learning models. A fundamental limitation of these models is their reliance on the co-location of LCS with high-cost stations to collect synchronized data pairs for training, which undermines the core objective of using LCS as an affordable option to increase spatial coverage (Maag, Zhou, and Thiele 2018). Moreover, these data correction models are typically

[1]To avoid confusion, we use the term 'correction' for all numerical/algorithmic approaches to data processing, and distinguish it from instrument calibration of the devices.

trained over a short period of time (often a few months), making them unreliable for long-term applications due to sensor drift and seasonal variations. Another significant but largely overlooked limitation is that these models often fail to account for real-world operational issues. For instance, deployed LCS exhibit significant bias and drift, and can experience periods of data or connectivity loss, causing their uncorrected readings to mislead end-users and public health analysts (Concas et al. 2021a). Lastly, previous studies do not use a standard benchmark or dataset for model evaluation. The lack of a common benchmark hinders reliable evaluation, as reported metrics often lack the context to compare different methods effectively.

To address these challenges, we introduce **Veli** (reference-free **V**ariational **E**stimation via **L**atent **I**nference), an unsupervised post-deployment LCS correction model. To develop and test our model, we built a standardized benchmark for AQ research, the Air Quality Sensor Data Repository (AQ-SDR). Our work makes three primary contributions:

- We propose a novel reference-free method for unsupervised data correction, eliminating the need for co-location with high-cost reference stations.
- We release the largest public benchmark for AQ monitoring, containing 23,737 sensors across diverse regions and pollution levels. This benchmark contains common sensor errors and operational failures, providing a resource suitable for modeling practical LCS deployment.
- We validate the model's real-world effectiveness and demonstrate its robustness and generalizability in both in-distribution and out-of-distribution settings.

2 Related Work

We categorize prior work into two groups: methods that rely on expensive, well-maintained reference stations for training (reference-based methods) and methods that do not use reference stations for training, and only use them for model evaluation (reference-free methods). In this work, we use the terms reference-free and unsupervised interchangeably. In the absence of established reference-free methods, we contextualize our contribution through a review of current reference-based approaches.

2.1 Reference-based Methods

Reference-based Correction Methods Reference-based correction methods use reference stations to correct inaccurate LCS readings. Given two sets, X_{LCS} and Y_{ref}, synchronized in time, a model $\mathbf{M}$ is trained to minimize the deviation between Y_{ref} and the mapping $\mathbf{M}(X_{\mathrm{LCS}})$ (e.g., using mean squared error). We assume by default that all reference ground truth data originate from accurate, well-maintained instruments. Reference-based correction methods are split into pre-deployment or post-deployment methods, depending on when the correction occurs.

A major limitation of pre-deployment reference-based correction methods is the need to co-locate target LCS units next to a reference station for an extended period to collect calibration data, making the deployment of large LCS networks impractical. Moreover, shorter co-location intervals yield models that poorly capture temporal variations such as seasonal changes. Lastly, this initial calibration does not account for long-term sensor drift, necessitating periodic recalibration. The logistical challenges of recalibrating deployed sensors mean that long-term drift often goes uncorrected in many devices. Most early studies adopted simple linear models in the pre-deployment context (e.g., ordinary least squares regression). We refer readers to (Concas et al. 2021a; Maag, Zhou, and Thiele 2018) for a comprehensive review of these methods.

The complexity of LCS errors has recently led to increased interest in non-linear post-deployment correction methods. These methods address the limitations of the traditional design paradigm, which relies on synchronized and co-located LCS-reference pairs, similar to (Ahn et al. 2025). For instance, (Cheng et al. 2019) addressed post-deployment correction using unsynchronized calibration transfer, a technique for in-field calibration via co-location with a reference station. This co-located LCS then serves as an anchor point, providing ground truth for other sensors with no co-located references in the network. While this method reports promising results, it was tested on only seven LCS during a ten-month period. Moreover, the LCS units were deployed in controlled settings, avoiding real-world issues such as missing data and extreme fluctuations. (Wang et al. 2023) proposed *CaliFormer*, a hybrid reference-based approach combining unsupervised reconstruction with supervised fine-tuning. The model is initially trained to reconstruct the LCS data in an unsupervised manner, and then fine-tuned to correct the results using ground-truth data from the reference stations.

In addition to direct correction of readings, some works have used historical data from reference stations as prior knowledge for LCS correction. Both 'RHC' (Li et al. 2020) and the Maximal Correlation Model (Li et al. 2021) leverage historical reference data to align LCS and reference readings' distributions. A significant limitation is that both approaches were evaluated on short time frames, restricting their applicability for long-term deployments.

Reference-based Interpolation Methods A different line of work bypasses LCS correction altogether, creating high-resolution AQ maps by interpolating data directly from a network of reference stations. Both MapTransfer (Cheng et al. 2020) and AirRadar (Wang et al. 2025) interpolate readings from high-cost stations to generate denser pollution maps. Despite their ability to produce high-resolution AQ maps, these approaches depend on reference stations with a sparse deployment across a region, which limits their ability to capture microclimate variations.

2.2 AQ Benchmarks

In Table 1, we compare previously published datasets and benchmarks that contain LCS data with our new dataset, AQ-SDR. We provide further details on AQ-SDR in Section 4.1 and the extended version. Previous datasets are either limited to small-scale studies on a regional level (Diez et al. 2024), or cover shorter time periods (Jiao et al. 2016). While

Dataset	# of Sensors	Period (months)	LCS & Reference
(Jiao et al. 2016)	20	10	LCS Only
(Diez et al. 2024)	49	34	LCS Only
(Van Poppel et al. 2023)	85	12	Both
(Bi et al. 2022)	109	22	Both
AQ-SDR	23737	80	Both

Table 1: Comparison between our dataset and other published AQ datasets. We disregard small-scale hyperlocal studies and datasets that have fewer than 10 sensors.

some benchmarks provide aligned LCS and reference station readings (Bi et al. 2022; Van Poppel et al. 2023), they do not provide a scale large enough to develop models that can generalize across diverse pollution levels. Our dataset is designed to serve as a unifying benchmark for LCS modeling and correction methods, capturing a wide range of failure modes, distribution shifts, sensor drift, and pollution levels to reflect real-world LCS behavior. AQ-SDR is the largest AQ sensor dataset to date, containing data from 23,737 low-cost and reference sensors across multiple global regions, collected over more than six years of deployment.

3 Reference-Free LCS Correction

3.1 Problem Formulation

A key challenge for reference-free correction is achieving robustness against the diverse failures and environmental factors seen in real-world deployments. Thus, it is essential to use a dataset that contains numerous instances of systematic drift, failures, and other erratic behaviors known to hinder LCS correction when developing and validating correction models (Concas et al. 2021b).

These combined real-world challenges often cause standard denoising and sensor fusion approaches, such as least-squares methods and Kalman filters (Kalman 1960), to fail as their state estimation becomes unreliable when readings are extremely erratic or contain missing values. To further illustrate the issue of sensor drift, we show in Figure 2 an example taken from AQ-SDR, shown as a comparison over a three-year period between a low-cost air quality sensor and a co-located, calibrated reference station in the city of Groningen in the Netherlands. LCS exhibit a noticeable shift in the distribution of their $PM_{2.5}$ readings over the years, despite having a distribution similar to that of the reference station around its initial deployment in 2019. In contrast, the well-maintained reference station shows consistent behavior, with a nearly identical data distribution over the same period.

To accurately model sensor bias, we build upon established findings that show LCS errors exhibit both nonlinear patterns and a systematic bias with heteroscedastic variance caused by environmental factors (Sharma et al. 2025; Concas et al. 2021b). These error characteristics motivated the use of techniques like Gaussian process regression to correct LCS readings (Malings et al. 2019; Li et al. 2023). Consequently, we adopt a similar rationale and propose a probabilistic sensor fusion model based on advances in Variational Autoencoders (VAEs) (Kingma and Welling 2014).

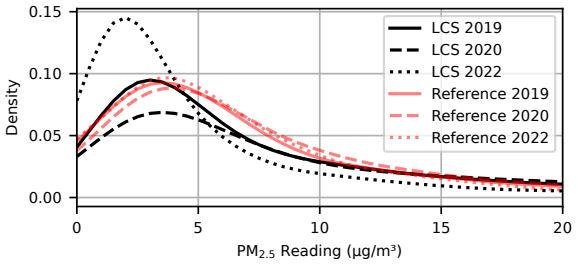

Figure 2: Probability Density Function (PDF) of $PM_{2.5}$ readings from an LCS device co-located next to a reference station over 3 years. The PDF of the LCS readings matches the reference in the first year of deployment, then shows significant drift over the next two years, unlike the well-maintained reference station that exhibits consistent behavior.

3.2 Model Overview

Our probabilistic model, shown in Figure 3, is designed to separate the true AQ readings from sensor noise. It learns a mapping from a noisy high-dimensional input stream to a low-dimensional latent variable. This latent variable represents a fused reading on a continuous manifold, which facilitates the reconstruction of a clean, corrected output. We enable the encoder to learn a robust mapping from any given reading to this manifold by training the model on a diverse range of noisy inputs.

In this implementation, we focus on correcting individual snapshots of LCS readings rather than modeling changes over time. We propose this design decision for two key reasons: First, it is difficult to obtain perfectly time-aligned data streams from multiple adjacent sensors without encountering gaps or simultaneous failures. Second, simultaneously modeling time alongside all noise patterns (e.g., spikes, missing data) compromises the model's ability to capture diverse non-temporal noise patterns. While our model processes hourly readings per pass, this snapshot-based approach does not discard the underlying temporal information. Since the correction model uses Lipschitz continuous layers, temporal signatures in the corrected output will be preserved (Virmaux and Scaman 2018).

3.3 LCS Noise Model

In this section, we provide the necessary formulation to build Veli. We refer readers to (Kingma and Welling 2014) for more insights on the foundations of VAEs, and provide a more detailed derivation in the extended version.

To model the general structure of noisy LCS readings, we start by defining a basic distribution for LCS readings:

$$x_{\text{noise}} \sim \mathcal{N}\Big(y + \mu_{\text{sens}},\ \Sigma_{\text{sens}}\Big) \tag{1}$$

where $x_{\text{noise}} \in \mathbb{R}^d$ is a noisy, raw AQ reading from d different sensors in the same vicinity and $y \in \mathbb{R}^d$ is the unobserved AQ reading if it were measured by an ideal instrument (e.g., reference station). As stated earlier, we are building a reference-free method, so y is inaccessible to us, and we replace it with predictions $\hat{y}$. $\mu_{\text{sens}} \in \mathbb{R}^d$

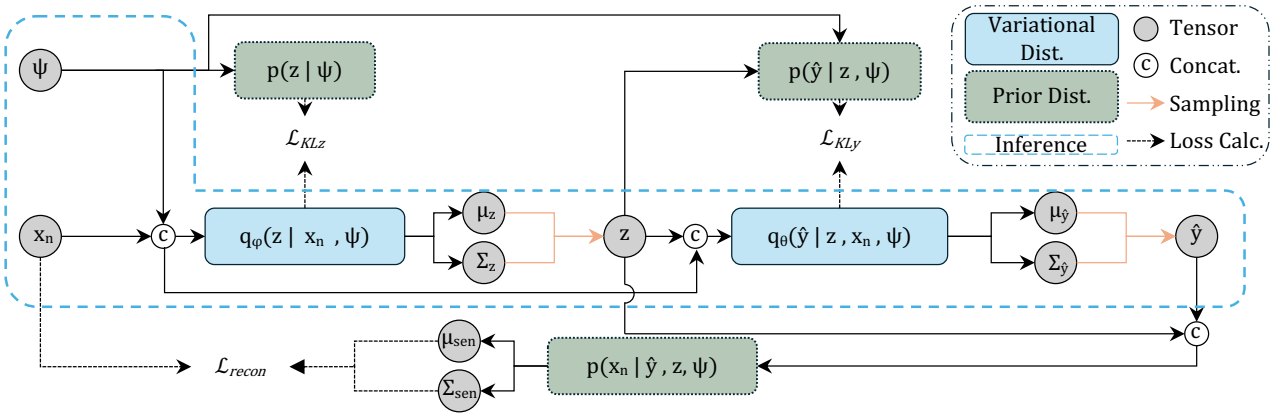

Figure 3: Veli structure following the derivation in Section 3.3. The input starts with noisy AQ readings x_n and auxiliary mask of 'NA' readings ψ on the left, propagating through the model's layers to generate a prediction of clean readings $\hat{y}$. Conditioning on ψ is omitted in some blocks for visual clarity but is implemented properly. Prior distribution blocks (green) are used in the training to estimate the variational distribution blocks (blue), which are used in the inference as indicated by the blue dashed line. All distribution blocks are modeled by two multilayer perceptron (MLP) layers followed by an MLP layer for each of the mean and variance. The losses $\mathcal{L}_{KL_z}$, $\mathcal{L}_{KL_y}$, and $\mathcal{L}_{recon}$ correspond to the three terms in eq. (6). Sampling refers to the traditional reparameterization in VAEs (Kingma and Welling 2014).

and positive definite diagonal covariance matrix $\Sigma_{\text{sens}} = \text{diag}(\sigma^2_{\text{sens},1}, \ldots, \sigma^2_{\text{sens,d}})$ are non-constant, nonlinear bias and heteroscedastic terms that affect the LCS reading. While μ_{sens} and Σ_{sens} do not model noise resulting from extreme spikes and missing data (extreme noise conditions), they can be used to produce a robust estimate of what the reading would be under normal noise conditions.

To enhance the representational capacity of the heteroscedastic terms, we introduce $z \in \mathbb{R}^r$ as a latent variable, where $r \leq d$. To model z, we condition it on an auxiliary parameter that contains additional information about the data, $\psi \in \mathbb{R}^d$. We then propose the following prior distribution:

$$p(z \mid \psi) = N(\mu(\psi), \Sigma(\psi)) \tag{2}$$

Standard VAEs typically use a standard Gaussian prior, $N(0, I)$. However, to build a more identifiable and flexible prior, we follow the approach in (Khemakhem et al. 2020) and introduce ψ as our auxiliary parameter. This approach allows the latent space to effectively learn diverse variations within the input data, which is essential in filtering erratic behavior. In the same manner, and since we are operating without LCS-reference-paired readings (x_{noise}, y), we treat y as a latent variable whose prior distribution is given as:

$$p(y \mid z, \psi) = N(\mu(z, \psi), \Sigma(z, \psi)) \tag{3}$$

3.4 Variational Approximations

We aim to reconstruct the signal by separately generating the clean and noisy components of the reading. We tackle this by maximizing a variational lower bound on $\log p(x_{\text{noise}})$ that contains z and y, conditioned on ψ, using the joint distribution factorization:

$$p(x_{\text{noise}}, y, z \mid \psi) = p(z \mid \psi)p(y \mid z, \psi)p(x_{\text{noise}} \mid y, z, \psi)$$

To estimate the distributions of y and z through the term $p(y, z \mid \psi)$, we will need to evaluate an intractable integral with no closed-form solution. Therefore, we introduce approximate variational distributions similar to (Kingma and Welling 2014), defined as:

$$q_\phi(z \mid x_{\text{noise}}, \psi) \approx p(z \mid x_{\text{noise}}, \psi)$$
$$q_\theta(y \mid z, x_{\text{noise}}, \psi) \approx p(y \mid z, x_{\text{noise}}, \psi)$$

Under the Gaussian assumption, the posterior q_ϕ becomes:

$$q_\phi\big(z \mid x_{\text{noise}}, \psi\big) = \mathcal{N}\big(\mu_z^\phi, \Sigma_z^\phi\big) \tag{4}$$

In practice, μ_z^ϕ and Σ_z^ϕ are produced by an encoder network with two-branch outputs f_ϕ, g_ϕ, respectively, such that $\mu_z^\phi = f_{\phi,\mu}(x_{\text{noise}}, \psi)$ and $\log \Sigma_z^\phi = g_\phi(x_{\text{noise}}, \psi)$. Similar to eq. (4), we can define the parameterized posterior approximation q_θ as:

$$q_\theta(y \mid z, x_{\text{noise}}, \psi) = \mathcal{N}\Big(\mu_y^\theta, \Sigma_y^\theta\Big) \tag{5}$$

and is parameterized by θ in the same manner such that $\mu_y^\theta = f_\theta(z, x_{\text{noise}}, \psi)$ and $\log \Sigma_y^\theta = g_\theta(z, x_{\text{noise}}, \psi)$. In this design, μ_y^θ is the clean reading mean estimate $\hat{y}$.

Using eqs. (5) and (4), we can approximate the intractable term $p(y, z \mid \psi)$ with a variational approximation $q_{\theta,\phi}(y, z \mid x_{\text{noise}}, \psi)$. Substituting q_ϕ and q_θ into the log-likelihood allows us to derive the Evidence Lower Bound (ELBO). Minimizing the negative ELBO sets the objective to find optimal parameters ϕ, θ for our model, such that:

$$\log p(x_{\text{noise}} \mid \psi) \geq \mathbb{E}_{q_{\theta,\phi}(y,z|x_{\text{noise}},\psi)}\big[\log p(x_{\text{noise}}, y, z \mid \psi) - \log q_{\theta,\phi}(y, z \mid x_{\text{noise}}, \psi)\big]$$

By using eq. (1) as our reconstruction goal and incorporating

z and ψ into the design, the final negative ELBO becomes:

$$\mathcal{L}(\theta,\phi) = \beta_z D_{\mathrm{KL}}\big(q_\phi(z \mid x_{\mathrm{noise}},\psi)\,\|\,p(z \mid \psi)\big) \quad (6)$$
$$+\ \beta_y D_{\mathrm{KL}}\big(q_\theta(y \mid z, x_{\mathrm{noise}},\psi)\,\|\,p(y \mid z,\psi)\big)$$
$$+\,\alpha \sum_{i=1}^{d}\left[\log\!\big(2\pi\,\sigma^2_{\mathrm{sens}}(z)_i\big) + \frac{\big(x_i - \hat{y}_i - \mu_{\mathrm{sens}}(z)_i\big)^2}{\sigma^2_{\mathrm{sens}}(z)_i}\right]$$

where $\sigma^2_{\mathrm{sens}}(z)_i$ and $\mu_{\mathrm{sens}}(z)_i$ are the non-linear bias and heteroscedastic terms in eq. (1), and $\hat{y}$ is the sampled prediction from the distribution in eq. (5) during training, but is taken as a point estimate during inference. Here $\alpha, \beta_z, \beta_y > 0$ are tunable coefficients similar to (Higgins et al. 2017).

The goal of this formulation is to minimize the reconstruction term to shrink toward zero, which happens when $\mu_{\mathrm{sens}}(z)$ absorbs the noise that creates x_{noise}, allowing q_θ to recover the underlying clean reading y. Concretely, $q_\theta(y \mid z, x_{\mathrm{noise}}, \psi)$ then learns the noise-free form of the signal such that our best estimate of the underlying clean signal, given a noisy input, becomes $\mu^\theta_y(z, x_{noise}, \psi)$.

This approach offers a clear advantage as it learns a smooth latent manifold that transforms erratic sensor readings, including missing values and spikes, into clean, continuous representations for decoding. By training on a richly varied dataset, the model captures the full spectrum of AQ conditions, producing a latent encoding for virtually any combination of noisy readings.

4 Experiments and Results

4.1 AQ-SDR Dataset

Dataset Details To build our model, we use our proposed dataset, the AQ-SDR, which aggregates the LCS data from three major citizen-science initiatives: SamenMeten, Sensor.Community, and Location Aware Sensing System (LASS) community (Chen et al. 2017). To supply reference measurements to validate our method, we provide data from four authoritative sources: LuchtMeetNet (Air Measurement Network), the Royal Netherlands Meteorological Institute, the European Environment Agency, and the Taiwanese Ministry of Environment Open Data. The majority of the deployed LCS began operating in 2019 and continue to provide measurements, with a smaller subset having been operational since before 2019. To further evaluate our model's generalizability across different pollution levels and regions, we create two partitions of the AQ-SDR dataset: an in-distribution set and an out-of-distribution set. We illustrate the difference between the in-distribution data and out-of-distribution data using two samples shown in Figure 4. The left-skewed distribution of the in-distribution data (Netherlands) reflects lower pollution levels, in contrast to the right-skewed distribution of the out-of-distribution data (Taiwan), which indicates higher pollution levels.

The first, in-distribution partition of the AQ-SDR has data from 99 sites across the Netherlands, each location hosting ten different LCS with no co-located reference station (reference-free). Its corresponding test set has five sites, chosen to provide the widest geographic coverage possible, each co-located with at least one reference station. The out-of-distribution partition consists of data from 55 heavily polluted locations in Taiwan, with no co-located reference station. The test set consists of five locations with co-located reference stations, similarly chosen to provide a wide geographic coverage. The dataset and the code to generate the partitions is publicly available and well-detailed, with further details on the dataset in the extended version. AQ-SDR will also be made accessible to the public through an interactive online dashboard as shown in Figure 1.

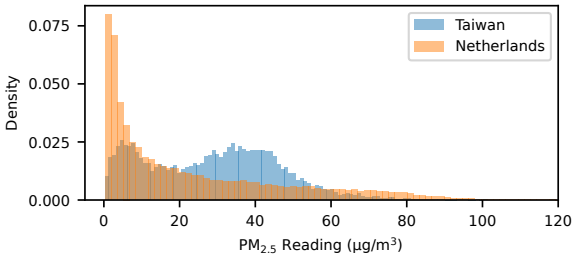

Figure 4: PDF Comparison of in-distribution and out-of-distribution data. Readings from the Netherlands are skewed to the left, indicating lower pollution levels, in contrast to the readings from Taiwan that reflect higher levels of pollution.

Dataset Processing Each LCS site hosts ten temporally aligned $PM_{2.5}$ sensors, which are sampled hourly, such that:

$$\big\{\, x_i(t) \;\big|\; i = 1,\ldots,10,\ t = t_1,\ldots,t_T \big\},$$

where $x_i(t)$ is the reading from the i-th sensor at time t. While using ten sensors was our design choice, we show in Section 4.4 implementations with fewer than ten sensors. To model missing data (referred to as 'NA' in this work) for sensor i at time t, we define an auxiliary mask ψ that is aligned in time to each location, such that $\psi_i(t) = 1$ if the data is observed, and $\psi_i(t) = 0$ if the data is missing ('NA'). This mask is essential for modeling data absence as conditional information in our model. In evaluation regions, we have aligned reference (ground truth) data $y(t)$. If there is more than one nearby reference station, we average their readings. At no point during the training was the model exposed to reference readings, keeping it completely reference-free.

4.2 Implementation Details & Evaluation Metrics

The model was implemented using PyTorch 2.3.1 and trained on an NVIDIA RTX 3090 GPU. We trained the model for 100 epochs with an ADAM optimizer, a batch size of 64, and an initial learning rate of 1×10^{-6}. Hyperparameters α, β_z, β_y in eq. (6) are set to 1, 10, 0.1, respectively, and we provide sensitivity analysis in the extended version.

All MLP layers that do not concatenate inputs use a hidden dimension of 32. For out-of-distribution fine-tuning, we froze the decoder and trained only the encoder for an additional 30 epochs on the new data distribution. Other implementation, tensor preparation, and evaluation details match previous works in time-series modeling (Liu et al. 2024).

To evaluate our model, we use Mean Absolute Error (MAE) as the standard metric from the literature to compare

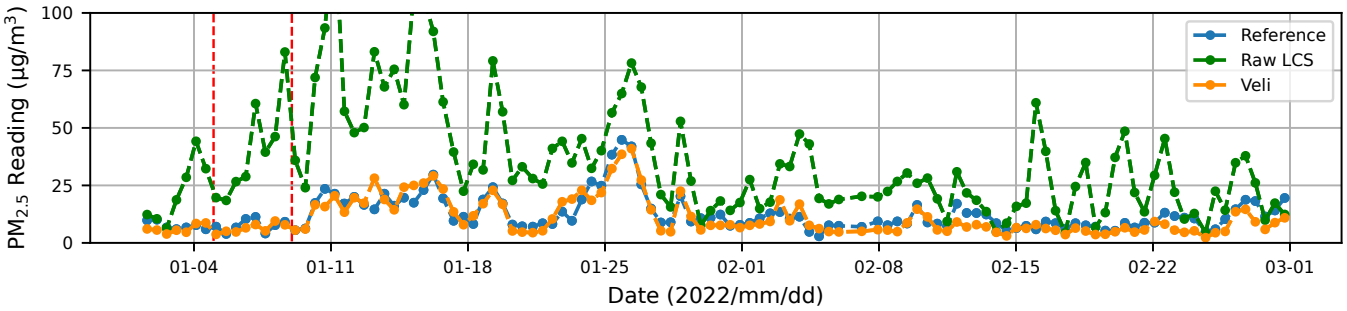

Figure 5: 12-hour-averages for Utrecht's data over two months. The readings of the raw LCS deviate significantly from the reference reading. Veli takes these readings as an input and outputs an accurate corrected measurement that matches the reference's readings. The region in the red-dashed lines is zoomed in on Figure 1.

model outputs to reference readings (Concas et al. 2021a; Maag, Zhou, and Thiele 2018). We also report the average and standard deviation of the output over five runs with different random seeds for all numerical results. In all experiments, 'raw LCS' is the input to Veli.

As this is the first work to propose a completely reference-free correction method, no other methods exist for a direct comparison. In addition, no large-scale unifying benchmark exists aside from AQ-SDR. We instead provide a comparison against traditional blind denoising techniques like Kalman Filters (KF) (Kalman 1960) and Principal Component Analysis (PCA) denoising (Weston, Schölkopf, and Bakir 2003). A KNN imputer was used to enable these two methods to run on data with missing readings.

4.3 Correction Results

In-distribution Results Table 2 presents the model's performance across five locations in five different cities in the Netherlands. The MAE decreased substantially compared to the raw LCS readings in Amsterdam, Rotterdam, and Utrecht. We also show the LCS units from IJmuiden and Nijmegen providing accurate readings that do not require correction. Veli introduces minimal stochastic noise due to sampling, and we expand on this in the extended version.

City	MAE ($\mu g/m^3$)			
	LCS_m	PCA	KF	Veli
Amsterdam	11.34	10.45	9.77	3.73±0.15
Rotterdam	21.27	22.31	11.57	3.36±0.37
Utrecht	24.77	13.72	15.95	5.25±0.26
IJmuiden	4.02	3.93	4.36	3.44±0.20
Nijmegen	2.82	2.82	2.96	3.06±0.18

Table 2: MAE comparison for in-distribution raw LCS, PCA denoising, KF denoising, and Veli's output. LCS_m is the average of raw LCS readings. Veli's results show mean ± standard deviation across five random seeds.

Figure 5 presents a 12-hour-average time series over a two-month period in Utrecht, which has the worst raw LCS accuracy among the selected regions. The region in the graph between the red-dashed lines highlights a four-day window, whose hourly sampling was shown earlier in Figure 1. Veli successfully captures both short- and long-term trends and spikes, despite our model being completely reference-free. While our model performs best when the raw LCS data exhibits an underlying trend (i.e., is not completely random), it is also designed to handle a common failure mode in which sensors produce random readings. We demonstrate this robustness in Section 4.4.

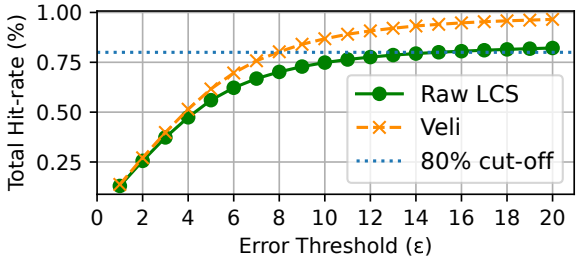

Figure 6: Percentage of data points that are within a threshold ϵ of the reference readings.

To obtain a holistic view of Veli's correction, we measure the hit rate, defined as the percentage of individual readings whose value is within a given threshold (ϵ) from the reference value. We count all measurements that have MAE $\leq \epsilon$, and plot this percentage of total data in Figure 6. For the raw LCS readings, we need to relax the MAE margin to be up to 20 to capture 80% of all readings, in contrast to our model, for which the required margin is reduced to only 7.34.

Out-of-distribution Results We further evaluate our model on out-of-distribution data from five locations in five different cities in Taiwan, shown in Table 3. "Veli zero-shot" denotes applying the weights trained on in-distribution data directly. For the fine-tuning variant, we froze the decoder and trained the encoder for 30 additional epochs on the Taiwanese LCS subset (reference-free). While the model shows strong average performance in a zero-shot setting, the results are inconsistent, illustrated by the high standard deviation across experiments. After fine-tuning, the model becomes significantly more reliable on out-of-distribution data.

City	LCS_m	PCA	KF	Veli	
				Zero-shot	Fine-tuned
Taichung	10.01	10.01	9.98	7.78±1.22	7.65±0.03
Tainan	14.09	14.25	13.28	8.59±1.48	7.83±0.27
Taoyuan	9.22	9.11	9.04	5.79±0.10	5.64±0.06
Taipei	7.52	7.49	7.58	6.51±0.98	6.43±0.03
Puzi	13.75	13.70	13.80	9.10±1.27	9.04±0.09

Table 3: MAE ($\mu g/m^3$) comparison for out-of-distribution raw LCS, PCA denoising, KF denoising, and Veli's output. LCS_m is the average of raw LCS readings. Veli's results show mean ± standard deviation across five random seeds.

4.4 Model Analysis and Discussion

Temporal Analysis $PM_{2.5}$ readings typically exhibit strong autocorrelation that gradually decays over time, primarily driven by the underlying pollutant concentrations (Zaini et al. 2022). In addition to this inherent structure, noise from LCS can also introduce significant autocorrelation, often persisting for up to 48 hours, as illustrated in Figure 7. The raw LCS readings remain highly autocorrelated for an extended period of time, in contrast to the trends seen in the reference stations. Our model eliminates this noise, producing outputs that closely match the reference time series. This behavior is consistent with the discussion in Section 3.1, showing that our model corrects the readings without compromising temporal information.

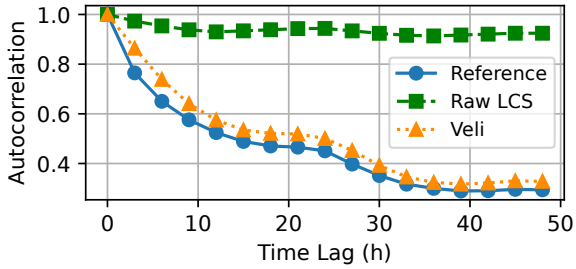

Figure 7: Comparison of autocorrelation over 48 hours. Correcting the raw LCS with Veli yields a behavior that is similar to a well-maintained reference station.

Ablation Studies & Limitations To simulate adversarial sensor failure (channel dropout), we took the original test data and randomly replaced a fixed number of the 10 sensor readings, n, with 'NA' values for each hourly sample. We then evaluated the model's performance for different values of n, from 1 to 9. Figure 8 shows how these injected failures degrade the correction's performance, but remains within an acceptable range of accuracy (MAE < 10). For sensors that are already accurate (e.g., Nijmegen's LCS), using a variation of Veli with fewer channels would be beneficial, which we show in the next subsection and in the extended version.

Minimum Number of Viable Sensors As established previously, our configuration uses a collection of 10 LCS per region. To evaluate Veli's flexibility, we tested its correction performance on subsets containing only 3, 5, and 7 sensors. For every sample, we ensured that at least half the sensors had a non-NA reading (rounded down). As Figure 9 shows, reducing the number of sensors does not significantly affect Veli's performance. However, using only three sensors increases the risk of connectivity loss, which can result in data gaps. Therefore, we retain 10 sensors as our standard configuration to maximize connectivity and data availability, but show that Veli remains effective using as few as 3 sensors.

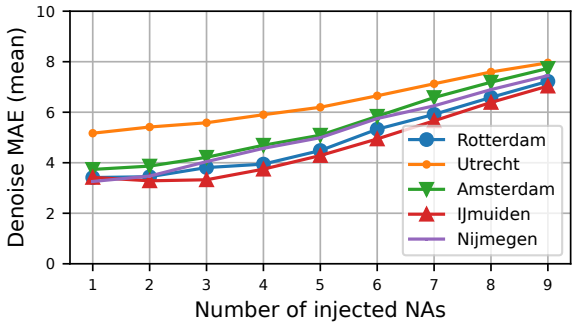

Figure 8: Effect of modeling sensor failure by injecting 'NA' readings into available LCS readings.

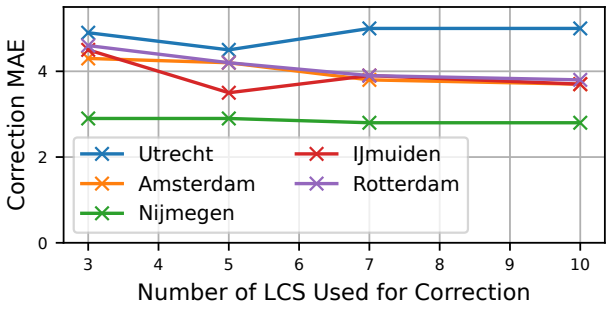

Figure 9: MAE of applying Veli on LCS readings when trained and tested on subsets with 3, 5, 7, and the default 10 sensors. The results demonstrate that performance is not significantly impacted by a reduction in sensor count.

5 Conclusion

In this work, we present Veli, an unsupervised Bayesian correction method for low-cost AQ sensors that does not require high-cost reference stations, lowering the barrier for deploying dense monitoring networks. To develop and evaluate our model, and to build a unifying benchmark for LCS, we also present the largest AQ benchmark to date, AQ-SDR, which contains data from 23,737 sensors distributed across multiple regions, capturing a diverse set of sensor errors and failure modes. Our evaluation demonstrates that Veli provides robust correction across varying pollution levels and data distributions and is resilient against common sensor failures, such as erratic spikes and complete sensor blackouts. We envision this work serving both as a practical solution for long-term LCS deployment and as a foundational benchmark for future research in AQ monitoring.

Acknowledgements

We sincerely thank the Dutch government for supporting this research with the starter grant (startersbeurzen). We also thank the organizations and researchers who provide the open data to enable this research, including the Dutch National Institute for Public Health and the Environment (RIVM), the Dutch Royal Netherlands Meteorological Institute (KNMI), Dr. Ling-Jyh Chen in Taiwan Academia Sinica for the AirBox project, the Taiwan Ministry of Environment, the Sensor.Community platform, and the European Environmental Agency (EEA). We also thank the GGD Amsterdam and RIVM for providing information about how air quality sensor stations work in the Netherlands. We also thank the CREATE Lab at the Robotics Institute at Carnegie Mellon University for the technical support in building the air quality dashboard.

References

Ahn, S.; Kim, H.; Shin, S.; and Seo, Y.-D. 2025. Real-Time Calibration Model for Low-Cost Sensor in Fine-Grained Time Series. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(1): 3–11.

Bi, J.; Carmona, N.; Blanco, M. N.; Gassett, A. J.; Seto, E.; Szpiro, A. A.; Larson, T. V.; Sampson, P. D.; Kaufman, J. D.; and Sheppard, L. 2022. Publicly available low-cost sensor measurements for PM exposure modeling: Guidance for monitor deployment and data selection. *Environment International*, 158: 106897.

Chen, L.-J.; Ho, Y.-H.; Lee, H.-C.; Wu, H.-C.; Liu, H.-M.; Hsieh, H.-H.; Huang, Y.-T.; and Lung, S.-C. C. 2017. An Open Framework for Participatory PM2.5 Monitoring in Smart Cities. *IEEE Access*, 5: 14441–14454.

Cheng, Y.; He, X.; Zhou, Z.; and Thiele, L. 2019. ICT: In-field Calibration Transfer for Air Quality Sensor Deployments. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 3(1).

Cheng, Y.; He, X.; Zhou, Z.; and Thiele, L. 2020. MapTransfer: Urban Air Quality Map Generation for Downscaled Sensor Deployments. In *Proceedings of the 2020 IEEE/ACM Fifth International Conference on Internet-of-Things Design and Implementation (IoTDI '20)*, 14–26. Sydney, Australia: IEEE/ACM.

Concas, F.; Mineraud, J.; Lagerspetz, E.; Varjonen, S.; Liu, X.; Puolamäki, K.; Nurmi, P.; and Tarkoma, S. 2021a. Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration: A Survey and Critical Analysis. *ACM Transactions on Sensor Networks*, 17(2): 1–44.

Concas, F.; Mineraud, J.; Lagerspetz, E.; Varjonen, S.; Liu, X.; Puolamäki, K.; Nurmi, P.; and Tarkoma, S. 2021b. Low-Cost Outdoor Air Quality Monitoring and Sensor Calibration: A Survey and Critical Analysis. *ACM Trans. Sen. Netw.*, 17(2).

Delaine, F.; Lebental, B.; and Rivano, H. 2019. In Situ Calibration Algorithms for Environmental Sensor Networks: A Review. *IEEE Sensors Journal*, 19(15): 5968–5978.

Diez, S.; Lacy, S.; Urquiza, J.; and Edwards, P. 2024. QUANT: a long-term multi-city commercial air sensor dataset for performance evaluation. *Scientific Data*, 11(1): 904.

Hagan, D. H.; Isaacman-VanWertz, G.; Franklin, J. P.; Wallace, L. M.; Kocar, B. D.; Heald, C. L.; and Kroll, J. H. 2018. Calibration and assessment of electrochemical air quality sensors by co-location with regulatory-grade instruments. *Atmospheric Measurement Techniques*, 11(1): 315–328.

Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; and Lerchner, A. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations*.

Jiao, W.; Hagler, G. S. W.; Williams, R.; Sharpe, R.; Brown, R.; Garver, D.; Judge, R.; Caudill, M.; Rickard, J.; Davis, M.; Weinstock, L.; Zimmer-Dauphinee, S.; and Buckley, K. 2016. Community Air Sensor Network (CAIRSENSE) project: Evaluation of low-cost sensor performance in a suburban environment in the southeastern United States. *Atmospheric Measurement Techniques*, 9(11): 5281–5292.

Kalman, R. E. 1960. A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D): 35–45.

Khemakhem, I.; Kingma, D.; Monti, R.; and Hyvarinen, A. 2020. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Chiappa, S.; and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2207–2217. PMLR.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Li, G.; Ma, R.; Liu, X.; Wang, Y.; and Zhang, L. 2020. RCH: robust calibration based on historical data for low-cost air quality sensor deployments. In *Adjunct Proceedings of the 2020 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2020 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '20 Adjunct, 650–656. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380768.

Li, G.; Wu, Z.; Liu, N.; Liu, X.; Wang, Y.; and Zhang, L. 2021. Blind Calibration by Maximizing Correlation. In *Adjunct Proceedings of the 2021 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2021 ACM International Symposium on Wearable Computers*, UbiComp/ISWC '21 Adjunct, 637–642. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384612.

Li, G.; Wu, Z.; Liu, N.; Liu, X.; Wang, Y.; and Zhang, L. 2023. A Variational Bayesian Blind Calibration Approach for Air Quality Sensor Deployments. *IEEE Sensors Journal*, 23(7): 7129–7141.

Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; and Long, M. 2024. iTransformer: Inverted Transformers Are

Effective for Time Series Forecasting. In *The Twelfth International Conference on Learning Representations*.

Maag, B.; Saukh, O.; Hasenfratz, D.; and Thiele, L. 2016. Pre-Deployment Testing, Augmentation and Calibration of Cross-Sensitive Sensors. In *Proceedings of the 2016 European Conference on Wireless Sensor Networks (EWSN)*, 169–180. Junction Publishing, Canada / ACM.

Maag, B.; Zhou, Z.; and Thiele, L. 2018. A Survey on Sensor Calibration in Air Pollution Monitoring Deployments. *IEEE Internet of Things Journal*, 5(6): 4857–4870.

Malings, C.; Tanzer, R.; Hauryliuk, A.; Kumar, S. P. N.; Zimmerman, N.; Kara, L. B.; Presto, A. A.; and Subramanian, R. 2019. Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring. *Atmospheric Measurement Techniques*, 12(2): 903–920.

Sharma, R.; Razakamanantsoa, A.; Kumar, A.; Thajudeen, T.; and Jullien, A. 2025. Performance and Applicability of Low-Cost PM Sensors to assess Global Pollution Variability through Machine Learning Techniques. *Atmospheric Environment: X*, 100331.

Van Poppel, M.; Schneider, P.; Peters, J.; Yatkin, S.; Gerboles, M.; Matheeussen, C.; Bartonova, A.; Davila, S.; Signorini, M.; Vogt, M.; Dauge, F. R.; Skaar, J. S.; and Haugen, R. 2023. SensEURCity: A multi-city air quality dataset collected for 2020/2021 using open low-cost sensor systems. *Scientific Data*, 10(1): 322.

Virmaux, A.; and Scaman, K. 2018. Lipschitz regularity of deep neural networks: analysis and efficient estimation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Wang, H.; Liu, Y.; Zhao, C.; He, J.; Ding, W.; and Chen, X. 2023. CaliFormer: Leveraging Unlabeled Measurements to Calibrate Sensors with Self-Supervised Learning. In *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing (UbiComp/ISWC '23 Adjunct)*, 743–748. Cancún, Quintana Roo, Mexico: ACM.

Wang, Q.; Xia, Y.; Zhong, S.; Li, W.; Wu, Y.; Cheng, S.; Zhang, J.; Zheng, Y.; and Liang, Y. 2025. AirRadar: Inferring Nationwide Air Quality in China with Deep Neural Networks. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence*, volume 39, 28467–28475.

Weston, J.; Schölkopf, B.; and Bakir, G. 2003. Learning to Find Pre-Images. In Thrun, S.; Saul, L.; and Schölkopf, B., eds., *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

World Health Organization. 2018. Health risks.

Zaini, N.; Ean, L. W.; Ahmed, A. N.; Abdul Malek, M.; and Chow, M. F. 2022. PM2.5 forecasting for an urban area based on deep learning and decomposition method. *Scientific Reports*, 12(1): 17565.

Zhang, Z.; An, R.; Guo, H.; and Yang, X. 2024. Effects of PM2.5 exposure and air temperature on risk of cardiovascular disease: evidence from a prospective cohort study. *Frontiers in Public Health*, 12.