

Machine Learning Approaches for Calibration of Low-Cost PM2.5 Sensors: From Linear Regression to Quantum Models

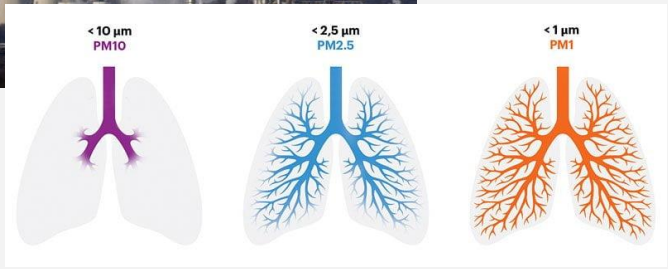
Alexander Uzhinskiy[a], Dušan Topalović[b], Mariia Khvostova[c]

[a] Meshcheryakov Laboratory of Information Technologies (JINR, Russia)
[b] Vinča Institute of Nuclear Sciences, University of Belgrade (Serbia)
[c] Lomonosov Moscow State University, Dubna Branch (Russia)

Relevance



Industrial and natural
PM air pollution



Diseases and premature
deaths in humans

Continuous monitoring required

Challenge

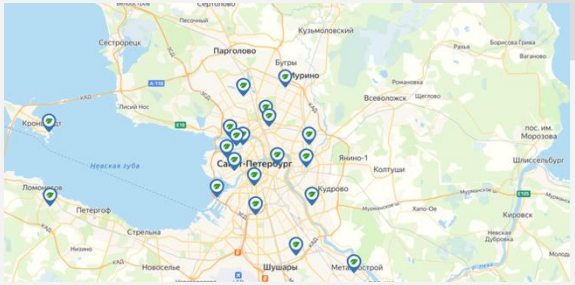


Reference station

Expensive to install and maintain
Limited in spatial coverage



SPB monitoring network

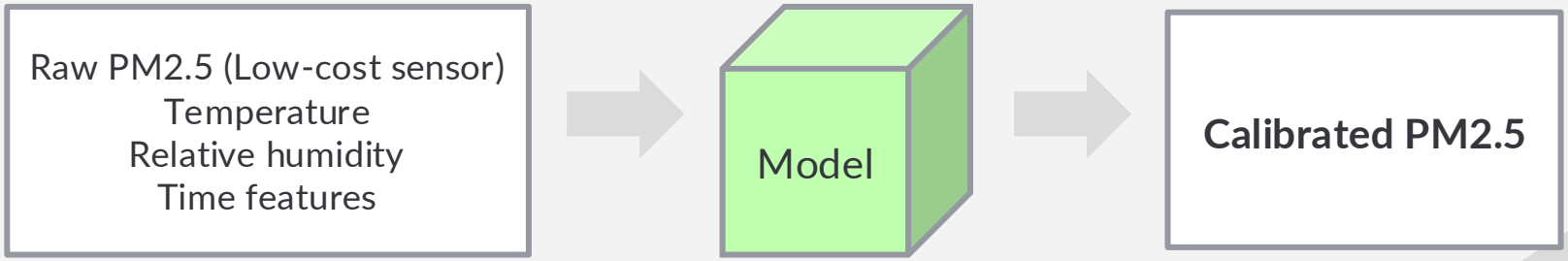


Portable sensor

Affordable
Potential for **high-density urban monitoring networks**
Sensitive to environmental conditions

Goal

Minimize deviation between low-cost PM2.5 sensor readings and reference measurements through machine learning-based calibration



Data*

Reference station: SEPA
Low-cost sensors: M1, M3, N48
Hourly measurements August 2025 – January 2026
Readings: temperature (T), relative humidity (rH), PM2.5 concentration

Date	PM10_SEPA	T_M1	rH_M1	PM25_M1	T_M3	rH_M3	PM25_M3	T_N48	rH_N48	PM25_N48
8.19.25 12:00	11,13	23,61	59,60	3,87	23,52	58,64	4,36	24,12	56,06	3,42
8.19.25 13:00	11,53	24,17	56,94	3,57	24,50	54,83	4,02	25,32	51,05	3,11
8.19.25 14:00	10,09	24,85	55,00	2,92	25,37	52,53	3,29	25,78	49,61	2,58
8.19.25 15:00	14,85	24,68	54,11	3,25	25,12	51,66	3,58	25,56	48,78	2,87
8.19.25 16:00	11,42	24,34	54,32	3,60	24,56	52,37	3,92	24,87	49,79	3,20
8.19.25 17:00	10,58	23,75	55,58	3,85	23,93	53,62	4,09	24,14	51,32	3,48

Extract from original dataset

**Courtesy of Serbian Environmental Protection Agency;*
Vinča Institute of Nuclear Sciences, University of Belgrade

Data Preprocessing

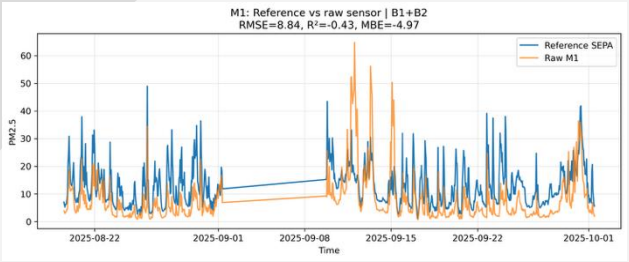
Time block selection

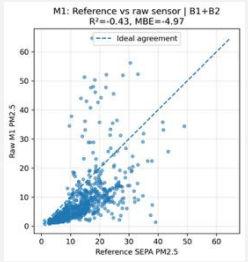
	Period	Duration
B1	Aug 19 – Sep 1, 2025	309 hours (~13 days)
B2	Sep 9 – Oct 1, 2025	522 hours (~22 days)

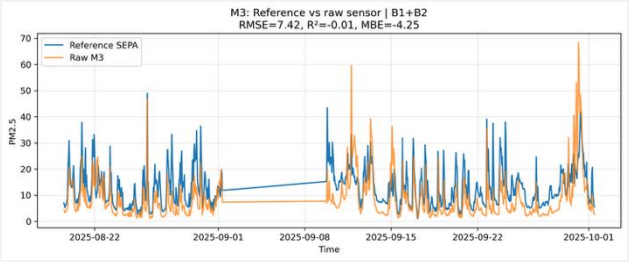
831 hours

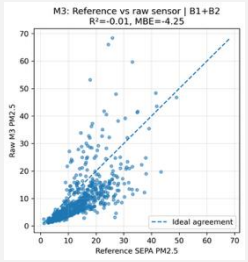
✓ Temporal synchronization
✓ No gaps

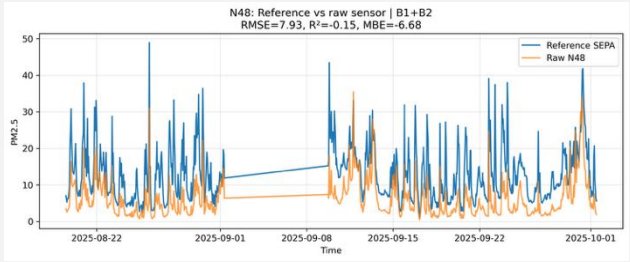
Sensors

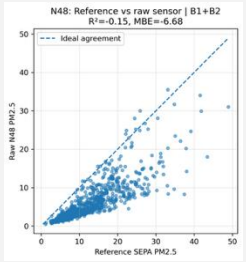














- ✓ **M1** – pronounced temporal variability
- ✓ **M3** – highest agreement with the reference
- ✓ **N48** – consistent systematic bias

RMSE – overall error, R^2 – agreement with reference, MBE – systematic bias

Problem Statement

Given:

$x(t) = [T(t), RH(t), PM_{2.5}^{sensor}(t)] \in \mathbb{R}^3$ – sensor readings (temperature, relative humidity, raw PM2.5).

$y(t) = PM_{2.5}^{ref}(t) \in \mathbb{R}$ – reference PM2.5 measurements (SEPA).

$X(t) = [x(t - W + 1), x(t - W + 2), \dots, x(t)] \in \mathbb{R}^{W \times 3}$, where W – temporal window size.

E.g.: Temporal window = {1,6}

PM (3:00)	PM (4:00)	PM (5:00)	PM (6:00)	PM (7:00)	PM (8:00)
R (3:00)	R (4:00)	R (5:00)	R (6:00)	R (7:00)	R (8:00)
T (3:00)	T (4:00)	T (5:00)	T (6:00)	T (7:00)	T (8:00)
$x(t-5)$	$x(t-4)$	$x(t-3)$	$x(t-2)$	$x(t-1)$	$x(t)$

Evaluation

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2} \quad \downarrow$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i| \quad \downarrow$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{\sum_{i=1}^{n}(\bar{y} - y_i)^2} \quad \rightarrow 1$$

$$MBE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i) \quad \rightarrow 1$$

Experimental Setup

✓ 5-fold walk-forward

	Train	Val	Test
1	50%	15%	10%
2	55%	15%	10%
3	60%	15%	10%
4	65%	15%	10%
5	70%	15%	10%

✓ 7 initializations

Models

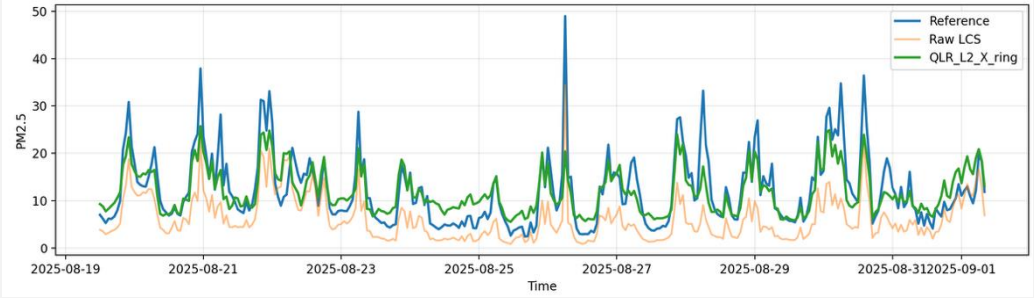
Group	Models
Classical	Random Forest (RF), XGBoost (XGB), Linear Regression (LR)
Temporal	GRU, LSTM, Transformer
Quantum	Quantum Linear Regression (QLR)

Sensor-Wise Results

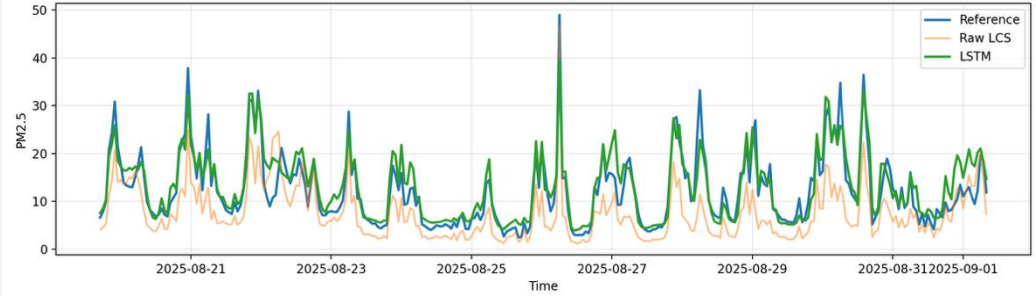
Sensor	Rank	Family	Model	Window	RMSE ± std	R^2 ± std
M1	**1**	**Quantum**	**QLR**	**1**	**4.942 ± 1.443**	**0.363 ± 0.043**
	2	Temporal	Transformer	6	5.127 ± 1.374	0.331 ± 0.085
	3	Temporal	GRU	6	5.133 ± 1.433	0.335 ± 0.072
M3	**1**	**Temporal**	**LSTM**	**6**	**2.964 ± 0.886**	**0.778 ± 0.047**
	2	Temporal	GRU	6	3.093 ± 0.924	0.757 ± 0.058
	3	Classical	RF	6	3.223 ± 0.886	0.728 ± 0.068
N48	**1**	**Temporal**	**Transformer**	**6**	**2.893 ± 0.393**	**0.756 ± 0.105**
	2	Temporal	LSTM	6	2.911 ± 0.494	0.748 ± 0.134
	3	Temporal	GRU	6	2.918 ± 0.448	0.744 ± 0.132

Model-Wise Results

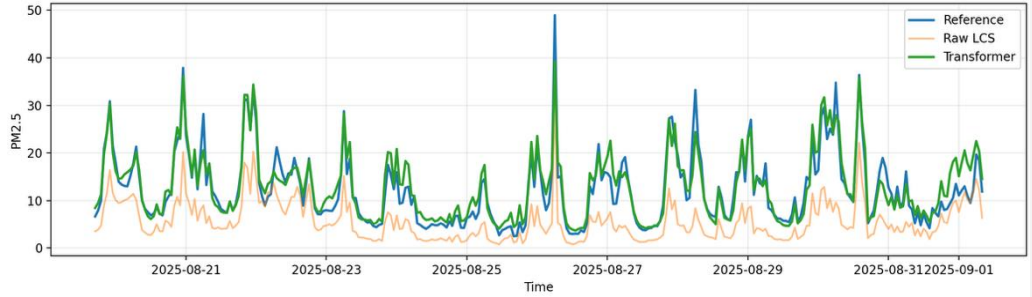
M1



M3



N48



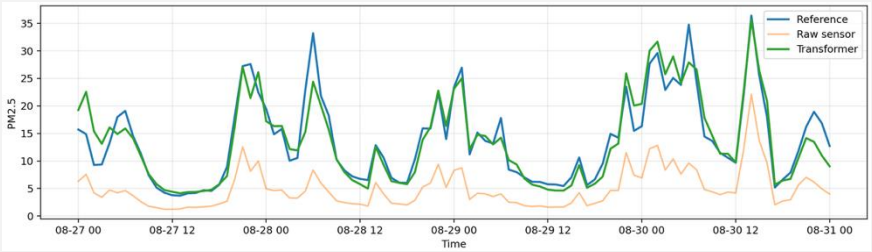
Model-Wise Results

Overall Rank	Model	Family	Top-1	Top-3	Mean Rank
1	**LSTM**	**Temporal**	**1**	**2**	**2.67**
2	**Transformer**	**Temporal**	**1**	**2**	**2.67**
3	**QLR**	Quantum	**1**	**1**	**4.33**
4	GRU	**Temporal**	0	2	3.00
5	RF	Classical	0	1	5.33
6	XGB	Classical	0	0	5.33
7	LR	Classical	0	0	9.67

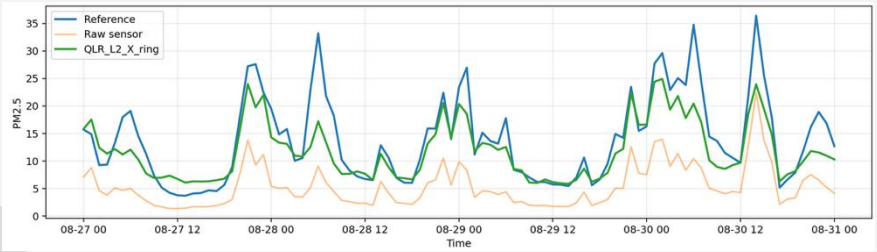
Calibration Effectiveness and Peak* Performance

Sensor	Best Model	Window	RMSE Improvement	MAE Improvement	Peak Improvement
M1	QLR	1	43.1%	56.2%	33.8%
M3	LSTM	6	52.5%	59.6%	55.3%
N48	Transformer	6	64.4%	70.5%	68.5%

N48



M1



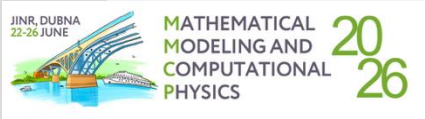
- ✓ Substantial **improvement for all sensors.**
- ✓ **Largest improvement for N48**, indicating predictable error pattern.
- ✓ **Stochastic M1 achieves 43% RMSE reduction** with QLR.

*Peak = high PM2.5 episodes **(≥ 90th percentile)**

Conclusion

- ✓ Machine learning enables **effective calibration of low-cost PM2.5 sensors**: RMSE reduction of 43–64% across all test devices.
- ✓ The optimal model **depends on sensor type and error characteristics**.
- ✓ **Temporal models** provide the **best overall performance**, while **QLR** is most effective for **stochastic sensors**.

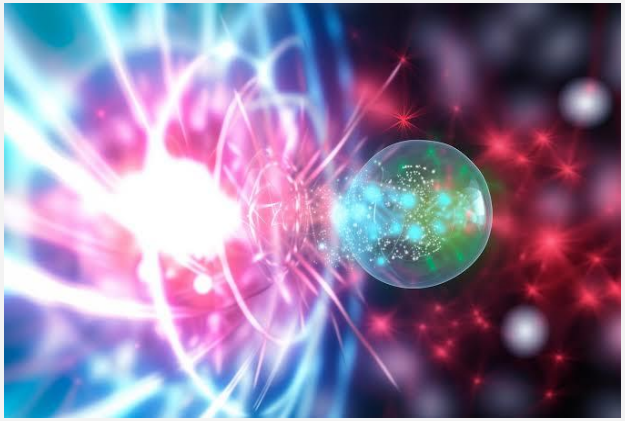
Thank you!


JINR, DUBNA
22-26 JUNE
MATHEMATICAL
MODELING AND
COMPUTATIONAL
PHYSICS
2026

Training Parameters

Model	Window	Scaler	Optimizer	Epochs	Hyperparameters
RF	1, 6, 8	StandardScaler	Built-in	— (convergence)	300 trees, max_depth=None
XGB					max_depth=4, lr=0.05
LR			Closed-form	— (analytical)	—
LSTM	6, 8		AdamW	**120** + early stopping	hidden=64, layers=2, dropout=0.2
GRU					hidden=64, layers=2, dropout=0.2
Transformer					d_model=32, nhead=4, layers=2
QLR	1		AdamOptimizer	**300** iterations	qubits=3, layers=2, stepsize=0.02

QLR Phenomenon



Hypothesized mechanisms

- ✓ Compact nonlinear representation via angle-based feature encoding
- ✓ Implicit regularization induced by constrained quantum state space
- ✓ Increased sensitivity to local signal variations while maintaining robustness to noise