Use of Low-Cost Sensors for Air Quality Monitoring in South Asia:
Virtual Training Series

Air quality data management and analysis

Dr. Emil Varghese
Senior Associate, CSTEP
emil.varghese@cstep.in

CSTEP
Center for Study of Science, Technology & Policy

Center for Study of Science, Technology and Policy

Mission

CSTEP is one of India's leading think tanks, with a mission to enrich policymaking with innovative approaches using science and technology for a sustainable, secure, and inclusive society.

- Clean Energy Transition
- Clean Air for All
- Sustainable Secure Future
- Digital Transformation
- Strategic Initiatives

20+
Years of Impactful Research

150+
Research publications guiding sustainable development

15+
Tools powering action nationwide

Air Quality (AQ)

The Air Quality (AQ) sector work closely with state and central pollution control boards to identify pollution sources through advanced emissions inventory modelling and source apportionment studies.

Low-cost sensor networks	PM chemical characterization
Air Quality modelling	Capacity building

Artificial Intelligence & Digital Platforms (AIDP)

Source apportionment tools	Air Quality Data Management System
Rooftop solar explorer	Climate change risk assessment

The Artificial Intelligence and Digital Platforms (AIDP) sector leads the development of decision-making tools, information portals, productivity tools, data visualization, analytics and AI/ML to drive efficiencies and increase value by amplifying outcomes and impact.

India Sensor Evaluation and Training (Indi-SET) facility: an independent, standardized testing for air sensors in India

- Reference-grade instruments for $PM_{2.5}$, PM_{10}, NO_2, O_3, CO and SO_2 + weather parameters + BC + Real-time chemical characterization + metal characterization

- Phase 1 sensor evaluation: Jan 20, 2024 to Jan 15, 2025
- Phase 2 sensor evaluation: Sept 24, 2025 to Jan 15, 2026 (AIRLAB Microsensor Challenge 2025)

General steps in the sensor data handling

Purchase of sensor nodes

Collocation with reference instruments

- Performance evaluation
- Development of localised calibration models

Deployment of sensor nodes

Sensor network data analysis

Sensor data processing levels

Schneider et al. (2019)

level	name	definition	example: gas sensors	example: particle sensors
Level-0	raw measurements	original measurand produced by sensor system	voltage corresponding to measured quantity, such as current for electrochemical and infrared sensors, resistance/conductance for metal-oxide sensors	voltage corresponding to current due to light scattered in nephelometers, or to binned counts for optical particle-counters
Level-1	intermediate geophysical quantities	estimate derived from corresponding level-0 data, using basic physical principles or simple calibration equations, and no compensation schemes.	for electrochemical sensors, NO_2 concentration in $\mu g/m^3$ or ppb, using only Level-0 data from the NO_2 sensor itself with no additional corrections beyond factory calibration ("raw data in concentration units")	binned particle-counts or PM mass in $\mu g/m^3$ derived from Level-0 data using simple calibration/assumed particle-density
Level-2A	standard geophysical quantities	estimate using sensor plus other onboard sensors demonstrated as appropriate for artifact correction and directly related to measurement principle[b]	NO_2 concentration in $\mu g/m^3$ or ppb, derived from onboard NO_2/NO/O_3 sensors, corrected for interferences based on the measurement principle and/or T/RH effects using onboard data	PM concentration in $\mu g/m^3$, corrected for T/RH effects with onboard-measured T/RH
Level-2B	standard geophysical quantities-extended	as Level-2A but using external data demonstrated as appropriate for artifact correction and directly related to measurement principle[b]	as Level-2A but using external data from nearby station related to correcting for interferences based on the measurement principle (e.g. O_3, T/RH)	as Level-2A but using external T/RH from nearby station
			Measurement/Prediction Boundary	
Level-3	advanced geophysical quantities	estimate using sensor plus internal/external inputs, not constrained to data proven as causes of measurement bias or related to measurement principle[b]	NO_2 concentration in $\mu g/m^3$ or ppb, derived from Level-2A or Level-2B data, further corrected by proxies known to be correlated with NO_2, e.g. emissions or modeled NO_2	PM concentration in $\mu g/m^3$, derived from Level-2A or Level-2B data, further corrected by proxies known to be correlated with PM, e.g. emissions or modeled PM
Level-4	spatially continuous geophysical quantities	spatially continuous maps derived from network of sensor systems	map of NO_2 concentrations in $\mu g/m^3$ or ppb, e.g. by assimilation of sensor network data into a physical model	map of $PM_{2.5}$ concentrations in $\mu g/m^3$, e.g. by assimilation of sensor network data into a physical model

[a]T/RH stands for temperature and relative humidity. The spatial support of all Levels except Level-4 is point measurements at single locations or for entire networks. [b]See Hagler et al. (2018).[2]

Sensor Evaluation Methodology

Data collection → Data standardization → Data cleaning → Data processing → Performance evaluation

Development of calibration models

Collocate sensor nodes with reference instruments → Split the data: Training and test data set → Develop correction model using training data → Test the model performance using test data → Periodically check the model and sensor performance

Data collection from sensor nodes

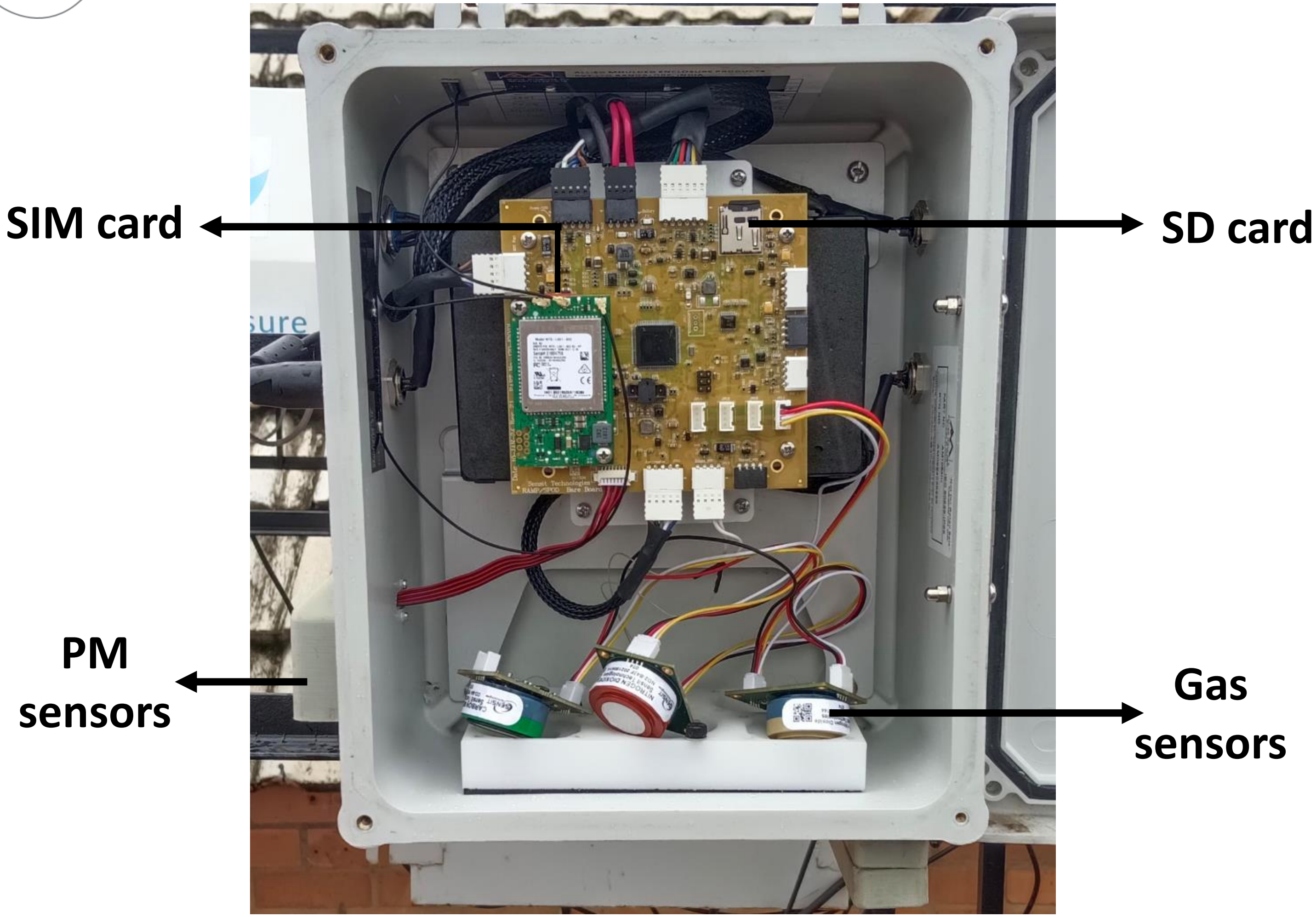
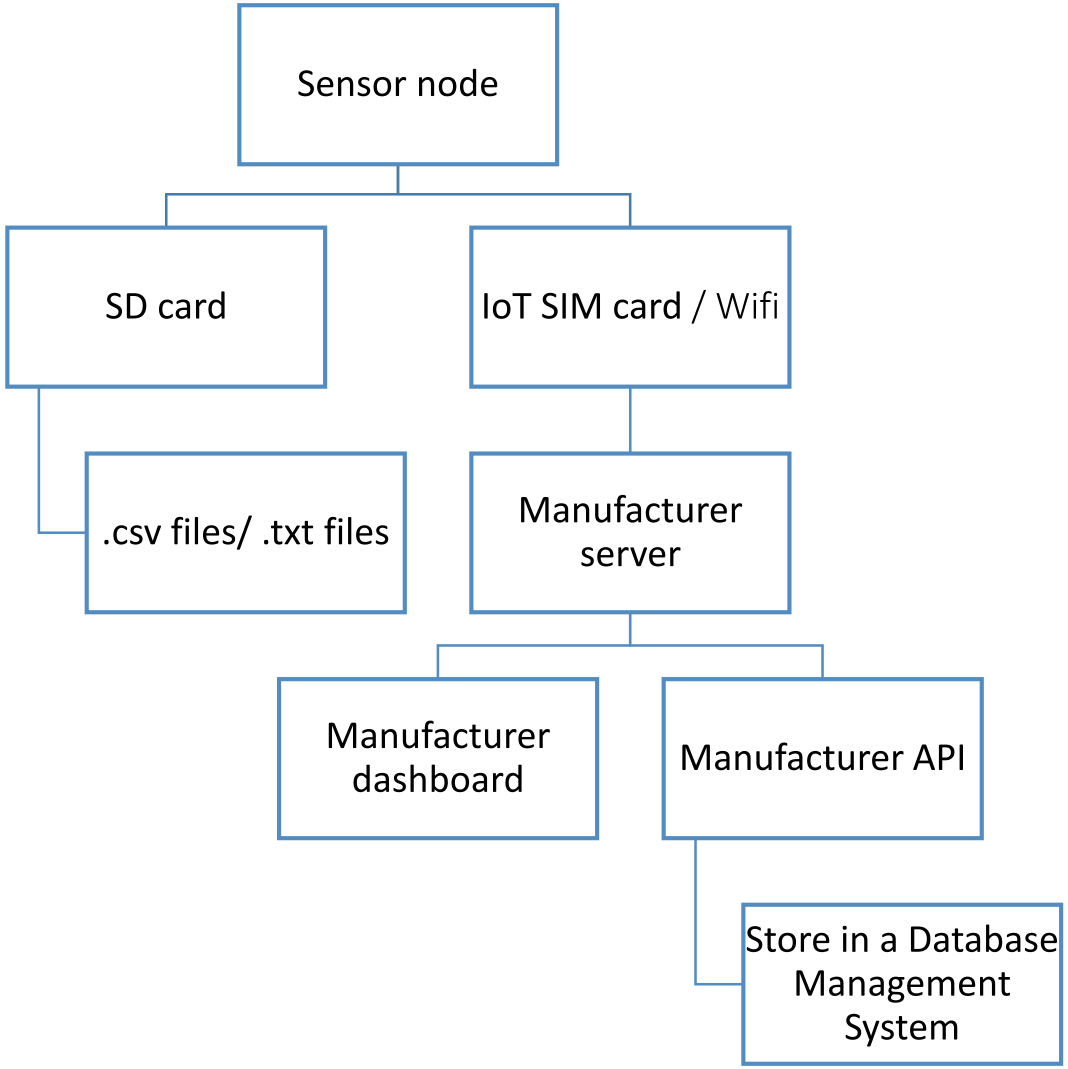


Photo credits: Vinod S.



Data standardization, cleaning and processing

Data standardization

- Convert to IST
- Assign standard units and parameters
- Aggregate to 1-minute interval

Data cleaning

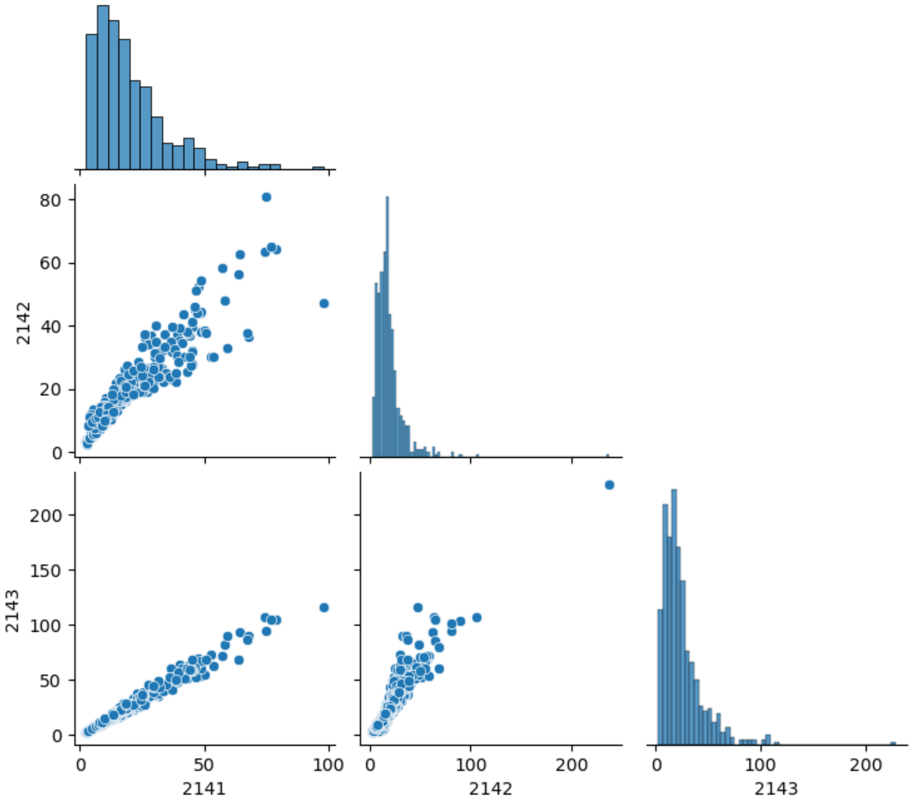
- Remove days with <75% data availability
- Remove negative values
- Remove out of threshold values

Data processing

- 15-minute averaging for gases
- 1-hour averaging for PM
- Consider 75% data availability while aggregating
- For averaging data 10:00 hrs to 11:00 hrs, assign to 10:00 hrs (optional)
- Merge sensor data with reference data based on 'DateTime' for analysis

Inter-unit precision

- Pearson correlation coefficient (r)
- Coefficient of determination (R^2)
- Standard deviation (SD)
- Coefficient of Variation (CV)



$$SD = \sqrt{\frac{1}{(N \times M) - 1} \sum_{j=1}^{M} \left[\sum_{h=1}^{N} (x_{hj} - \bar{x}_h)^2 \right]}$$

$$CV = \frac{SD}{\bar{x}} \times 100$$

where (for PM):

SD = Standard deviation of 1-hour averaged sensor PM concentration measurements (μg/m^3)

N = Number of 1-hour periods during which all identical instruments are operating and returning valid averages

M = Number of identical sensors operated simultaneously during the field test

x_{hj} = 1-hour averaged sensor PM concentration for hour h and sensor j (μg/m^3)

$\bar{x}_h$ = 1-hour averaged PM concentration for hour h across all M sensors (μg/m^3)

$\bar{x}$ = deployment average PM concentration for a field test (μg/m^3)

Comparison of sensor to reference instruments

- Linearity: r, R^2
- Bias: Slope, Intercept

For a regression of the form:

$$y = \beta_0 + \beta_1 x$$

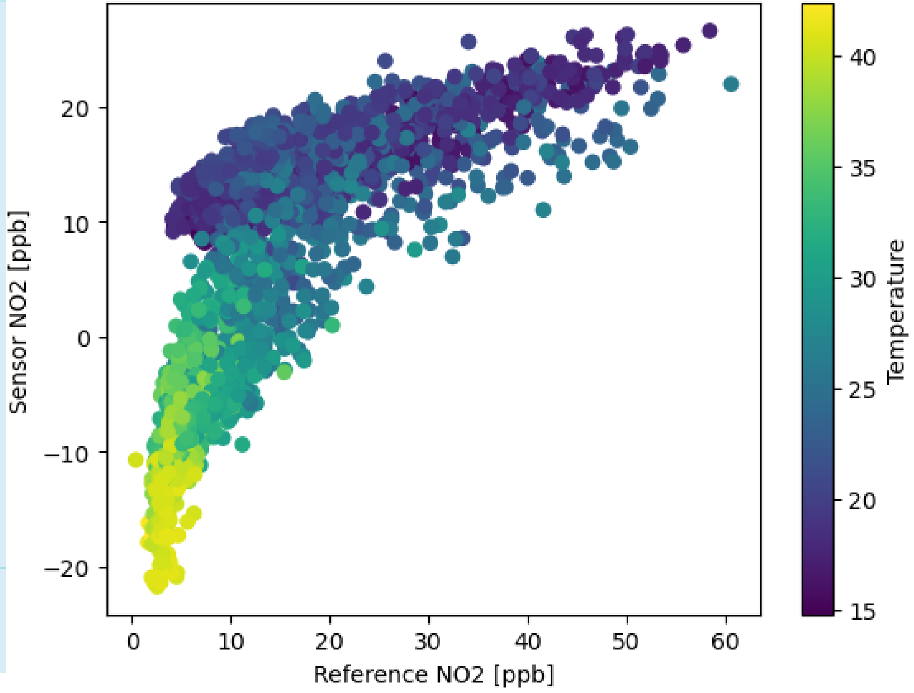
the slope (β_0) and intercept (β_1) are obtained from the least squares formulas:

$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}, \qquad \beta_0 = \bar{y} - \beta_1 \bar{x}$$

where:

- x_i = reference concentration values
- y_i = sensor concentration values
- $\bar{x}$ = mean of reference concentration values
- $\bar{y}$ = mean of sensor concentration values

In Python, the same result is obtained using NumPy's **polyfit**



Comparison of sensor to reference instruments

- Error: RMSE, NRMSE, CvMAE

$$RMSE = \sqrt{\frac{1}{N}\sum_{h=1}^{N}(x_h - R_h)^2}$$

N = Number of 1-hour periods during which all identical instruments are returning valid averages
x_h = 1-hour averaged sensor PM concentration for hour h
R_h = 1-hour averaged reference monitor concentration for hour h
$RMSE$ = Root Mean Square Error

$$NRMSE = \frac{RMSE}{\bar{R}_d} \times 100$$

NRMSE = Normalized Root Mean Square Error (%)
$\bar{R}_d$ = valid 1-hour averaged reference concentration over the entire testing period

$$CvMAE = \frac{\sum_{i=1}^{N} |x_h - n_{bias} - R_h|}{\sum_{i=1}^{N} R_h} \times 100$$

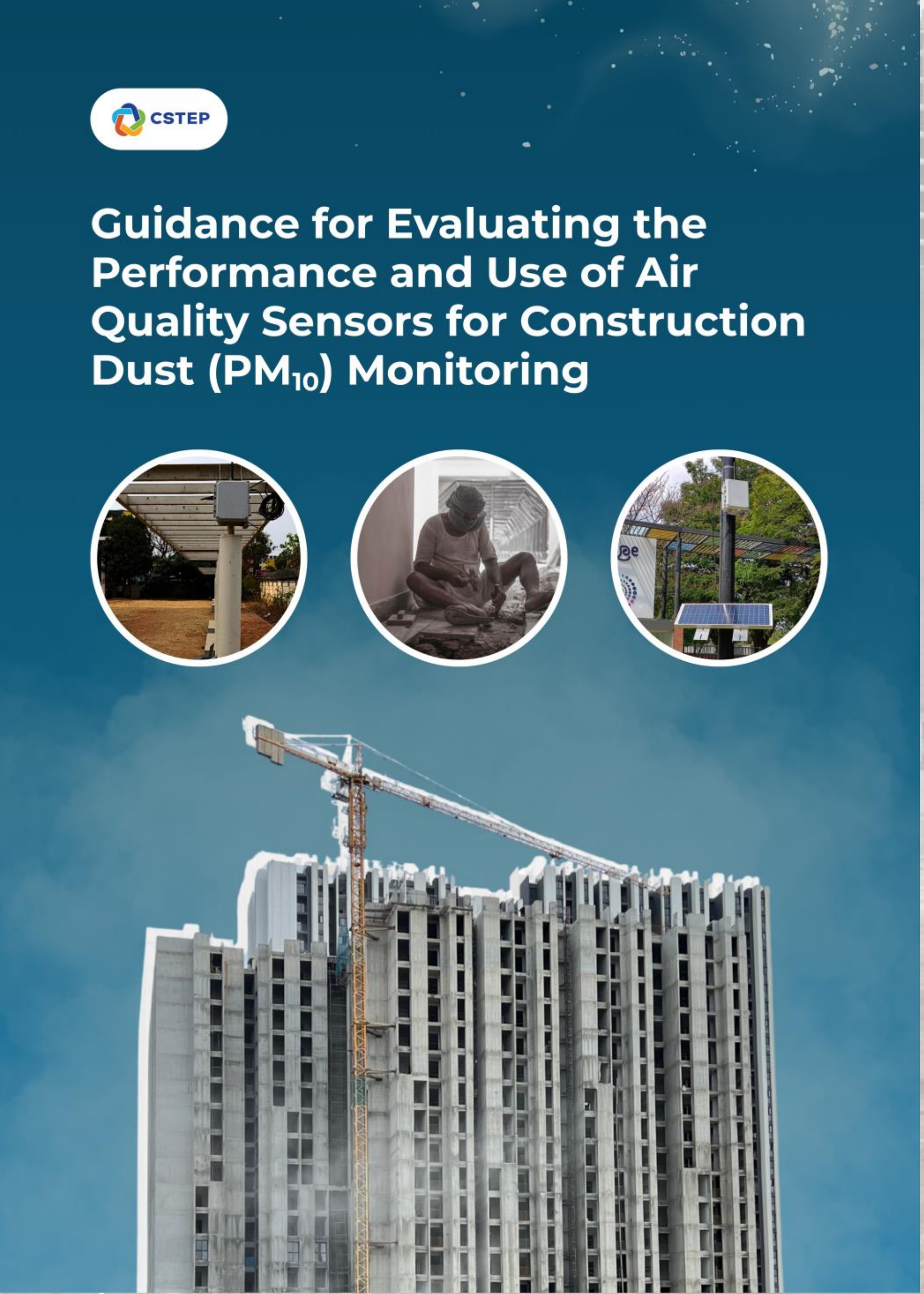
where:

$$n_{bias} = \frac{1}{N}\sum_{i=1}^{N}(x_h - R_h)$$

x_h = 1-hour averaged sensor concentration for hour h
R_h = 1-hour averaged reference monitor concentration for hour h
N = Number of 1-hour periods during which all identical instruments are returning valid averages
$CvMAE$ = Coefficient of Variation in Mean Absolute Error

Performance Evaluation Target Values (adapated from US-EPA)

Performance metrics	$PM_{2.5}$	PM_{10}	O_3	NO_2	CO	SO_2
Precision						
Pearson correlation coefficient (r)	≥ 0.9	≥ 0.9	≥ 0.9	≥ 0.9	≥ 0.9	≥ 0.9
Coefficient of determination (R^2)	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8
Standard Deviation (SD)	≤ 5 µg/m³	≤ 5 µg/m³	≤ 5 ppb	≤ 5 ppb	≤ 0.02 ppm	≤ 5 ppb
Coefficient of Variation (CV)	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %
Bias						
Slope (m)	1 ± 0.35	1 ± 0.35	1 ± 0.2	1 ± 0.35	1 ± 0.2	1 ± 0.35
Intercept (c)	-5 ≤ c ≤ 5 µg/m³	-10 ≤ c ≤ 10 µg/m³	-5 ≤ c ≤ 5 ppb	-5 ≤ c ≤ 5 ppb	-0.05 ≤ c ≤ 0.05 ppm	-5 ≤ c ≤ 5 ppb
Linearity						
Pearson correlation coefficient (r)	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8	≥ 0.8
Coefficient of determination (R^2)	≥ 0.7	≥ 0.7	≥ 0.8	≥ 0.7	≥ 0.8	≥ 0.7
Error						
Root Mean Square Error (RMSE)	≤ 7 µg/m³	≤ 14 µg/m³	≤ 5 ppb	≤ 15 ppb	≤ 0.15 ppm	≤ 15 ppb
Normalized Root Mean Square Error (NRMSE)	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %
Bias-corrected Coefficient of Variation in Mean Absolute Error (CvMAE)	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %	≤ 30 %



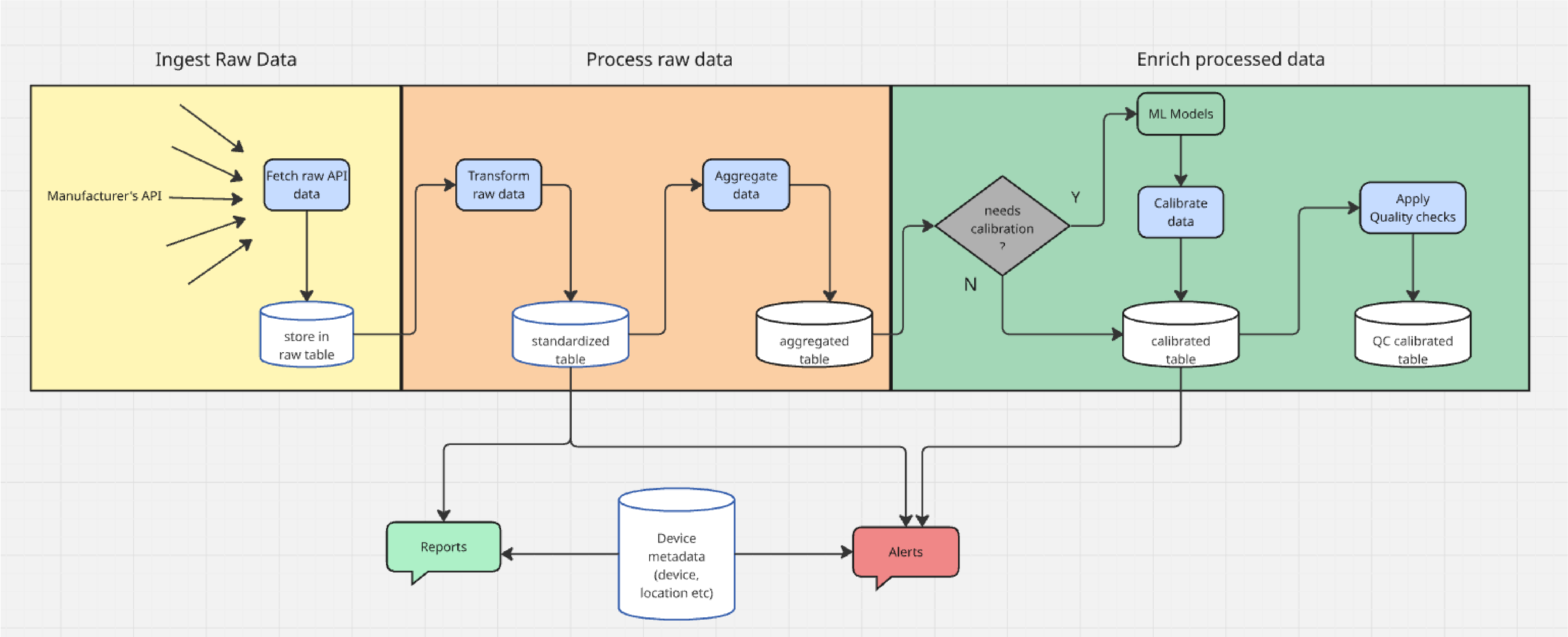
Publication alert

- **Which is the best sensor model to be used for construction dust monitoring?**
- **How to carry out the collocation study for evaluating and calibrating PM sensors?**
- **How to store and process sensor data?**

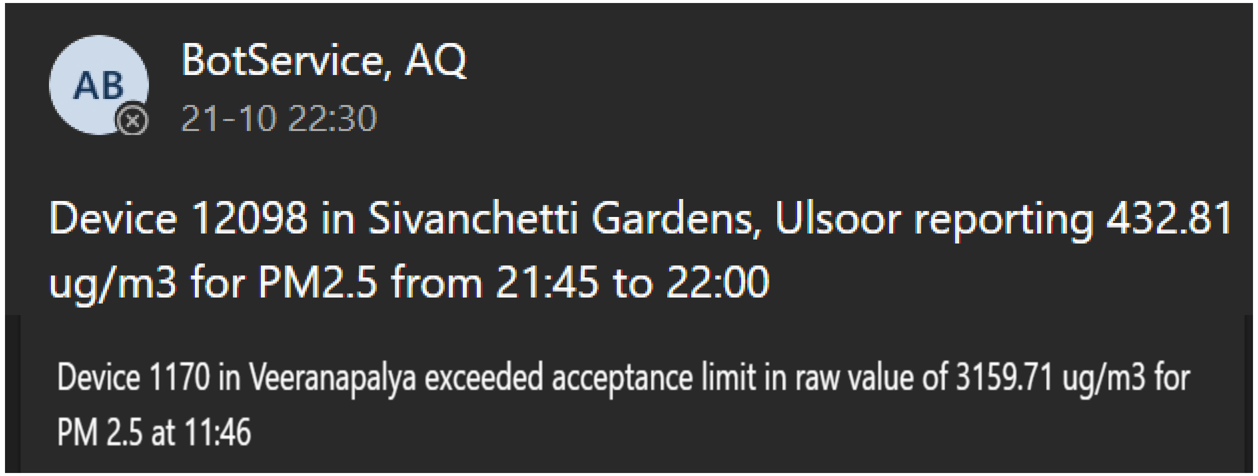
[Guidance for Evaluating the Performance and Use of Air Quality Sensors for Construction Dust (PM10) Monitoring – CSTEP]

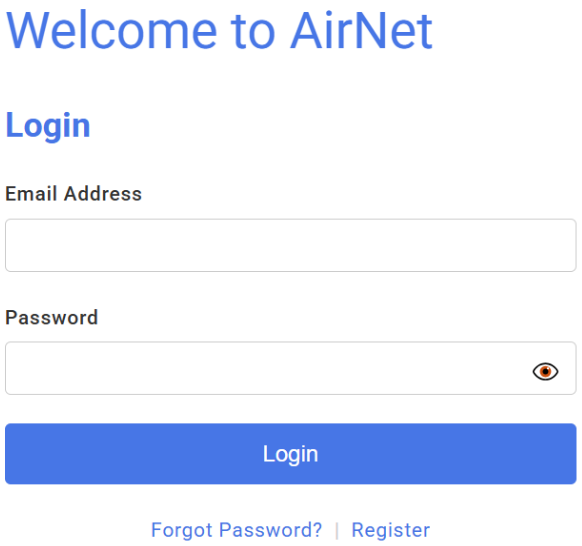
Contributing authors: Dr. Emil Varghese and Dr. R. Subramanian

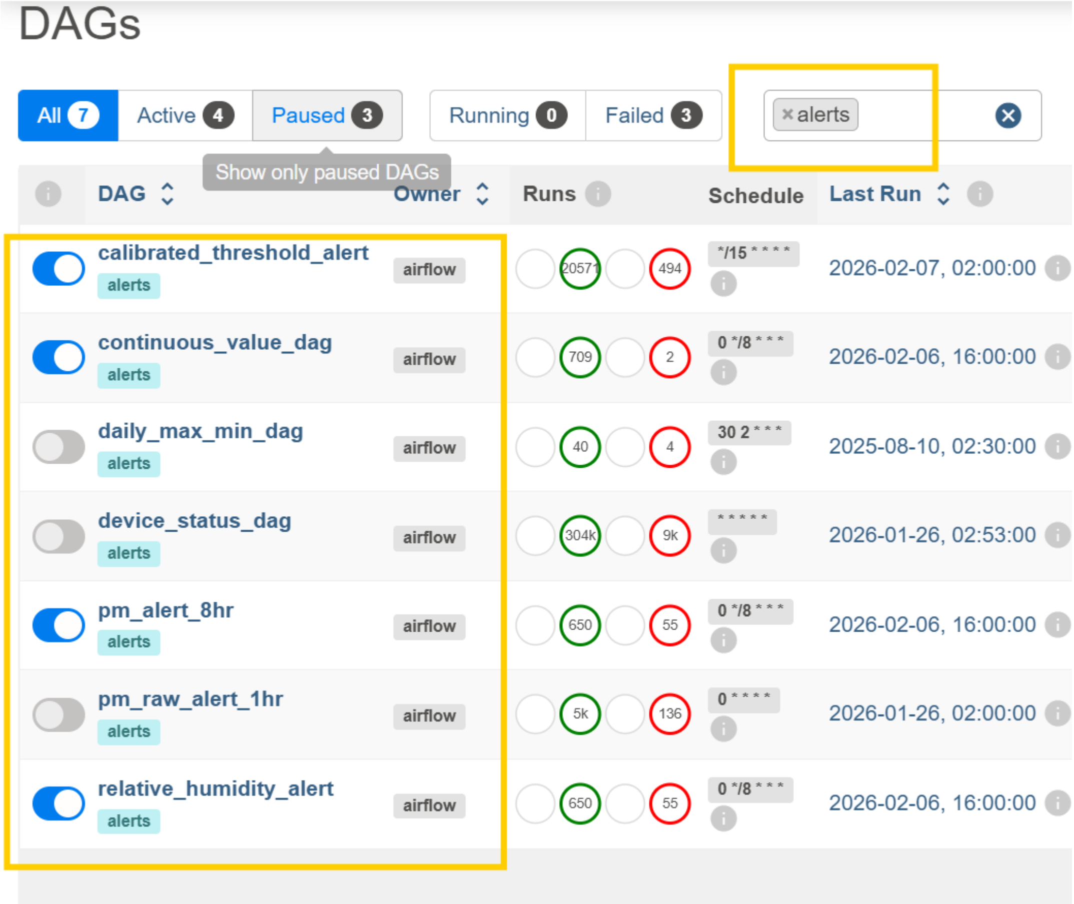
Air Quality – Data Management System (AQ-DMS)



Get AQ data available at neighborhood levels – with alerts carefully flagged



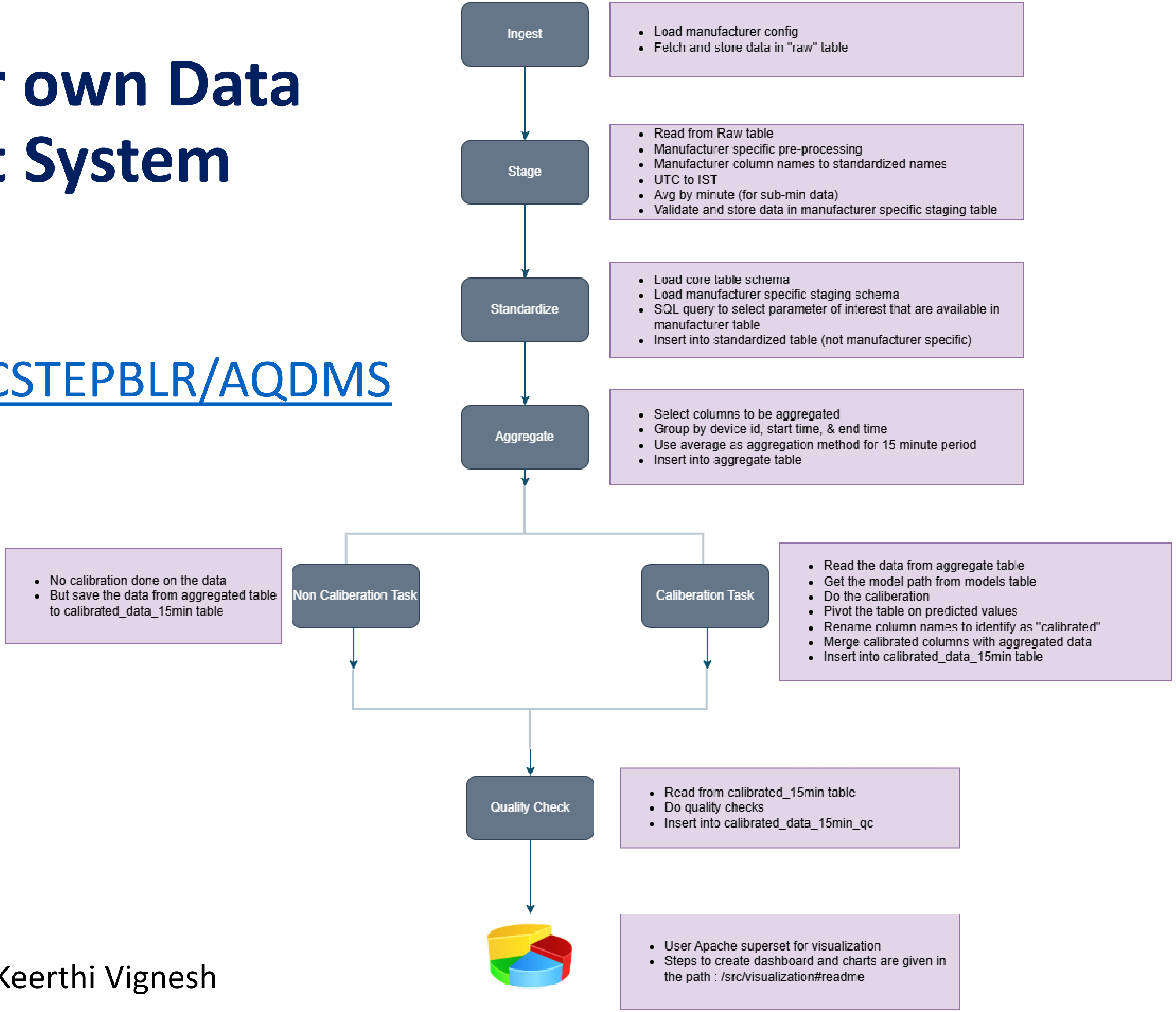




https://airnet.cstep.in/

Develop your own Data Management System

https://github.com/CSTEPBLR/AQDMS



Contributors: Rakshita Kolhar and Keerthi Vignesh

Acknowledgements

Dr. R. Subramanian

Rakshita Kolhar

Dr. Nirav Lekinwala

Srinivas Sridharan

Vinod S.

Swagata Dey

Shobitha Cherian

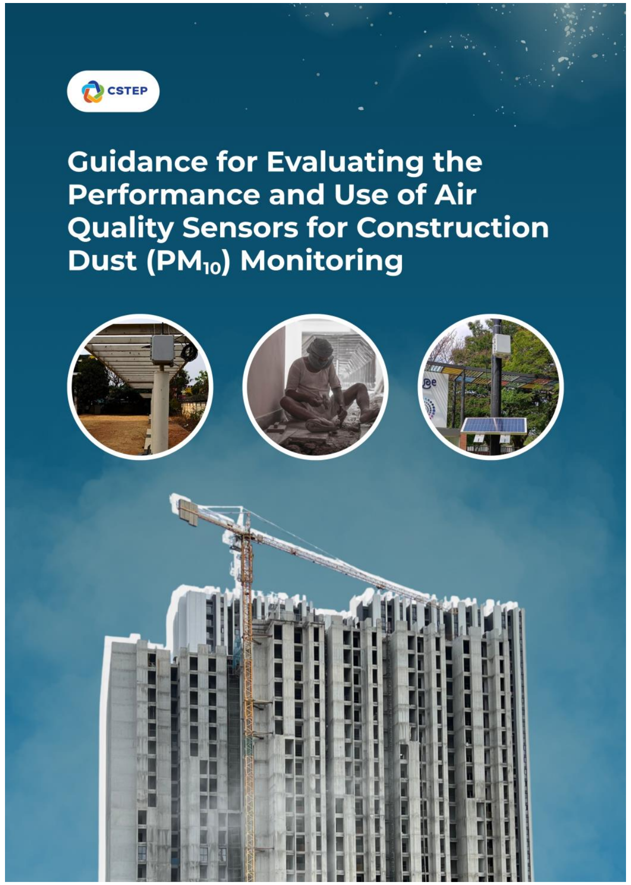
Yashwant Pratap S.Y.

Kavyashree N. Kalkura

Ashwin K.R.

Keerthi Vignesh

Rajesh Shenoy



Center for Study of Science, Technology & Policy (CSTEP), Bengaluru 560094, India

emil.varghese@cstep.in; airquality@cstep.in

References

- Guidance for Evaluating the Performance and Use of Air Quality Sensors for Construction Dust (PM10) Monitoring – CSTEP
- AQ-DMS GitHub repository: https://github.com/CSTEPBLR/AQDMS
- Zimmerman et al. (2018): https://doi.org/10.5194/amt-11-291-2018
- Malings et al. (2019): https://doi.org/10.5194/amt-12-903-2019
- Giordano et al. (2021) : https://doi.org/10.1016/j.jaerosci.2021.105833
- Li et al. (2021): https://doi.org/10.1021/acssensors.1c00729
- Schneider et al. (2019): https://doi.org/10.1021/acs.est.9b03950
- WMO reports on LCS: WMO-No. 1215 (2021), GAW Report No. 293 (2024)
- Enhance Air Sensor Guidebook: EPA/600/R-22/213
- EPA Performance Testing Protocols: EPA/600/R-20/279, EPA/600/R-20/280, EPA/600/R-23/145, EPA/600/R-23/146