CALIBRATION OF A LOW-COST PM10 SENSOR WITH MACHINE LEARNING ALGORITHMS

BDRC-2025-001

Name of the Author(s)

Ernest Werner - AEMET

Gerardo García-Castrillo - AEMET

01 July 2025

TECHNICAL REPORT

Series: Barcelona Dust Regional Center (BDRC) Technical Report

A full list of BDRC Publications can be found on our website under: http://dust.aemet.es/about-us/technical-reports

C/Jordi Girona, 29 | 08034 Barcelona (Spain)

Summary

This report outlines the methodology employed for the calibration of the Low-Cost PM10 sensor, which has been developed by the Barcelona Dust Regional Center as part of CREWS project. The calibration is based on four different methods, all utilizing machine learning.

Contents

1. Introduction

The Climate Risk and Early Warning Systems (CREWS, https://crews-initiative.org) Initiative is a global partnership that aims to significantly increase the availability of, and access to, multi-hazard early warning systems in Least Developed Countries (LDCs) and Small Island Developing States (SIDS). These nations are often the most vulnerable to the impacts of climate change and extreme weather events.

Its core mission is to save lives, protect assets, and secure livelihoods by ensuring these communities have timely, understandable, and actionable warnings about extreme weather and climate events. CREWS achieves this by supporting projects that strengthen all four pillars of effective early warning systems: disaster risk knowledge, hazard monitoring and forecasting, warning dissemination, and preparedness/response capabilities.

Implemented by key partners including the World Meteorological Organization (WMO), the UN Office for Disaster Risk Reduction (UNDRR), and the World Bank (via GFDRR), CREWS is a multi-donor trust fund. The Barcelona Dust Regional Center, for instance, contributes to CREWS by developing and disseminating crucial dust storm forecasts, aiding in early warning for a specific, impactful hazard. Ultimately, CREWS aims to close the capacity gap in climate services and build global resilience against increasing climate risks.5. The Role of the Barcelona Dust Regional Center within CREWS:

The Barcelona Dust Regional Center, hosted by AEMET (Spanish State Meteorological Agency) and BSC (Barcelona Supercomputing Center), plays a vital role within the WMO's Sand and Dust Storm Warning Advisory and Assessment System (SDS-WAS) program. This system is crucial for providing early warnings for sand and dust storms, which are significant meteorological hazards with wide-ranging impacts on health, environment, and various socio-economic sectors.

Supporting Early Warning Systems: Their work directly supports the "Detection, Observation, Monitoring, Analysis, and Forecasting" pillar for dust-related hazards in regions affected by sand and dust storms, like Northern Africa, the Middle East, and Europe. This is precisely where your low-cost PM10 sensor calibration work fits in, as it aims to improve local monitoring and prediction of dust.

In essence, the CREWS Initiative is a critical global effort to enhance resilience against climate and weather-related disasters by ensuring that vulnerable communities receive timely and actionable warnings, with specialized centers like the Barcelona Dust Regional Center contributing expertise in specific hazards like dust storms.

2. Calibration methodology

The purpose of this section is to describe the calibration method and the observation data used.

2.1 Machine learning algorithms

For the calibration we have used four different algorithms:

Multiple Linear Regression (MLR)

Multiple Linear Regression (MLR) is a statistical method used to model the relationship between a dependent variable and two or more independent variables. It aims to find the best-fitting linear equation that describes how changes in the independent variables correspond to changes in the dependent variable. Essentially, it extends simple linear regression by incorporating multiple predictor variables.

MLR is a statistical modeling technique that predicts a continuous dependent variable by fitting a linear relationship to two or more independent variables. It seeks to find the best-fitting hyperplane that minimizes the sum of squared differences between observed and predicted values. It's simple, interpretable, but assumes linearity.

Multiple Linear Regression with Stochastic Gradient Descent (MLR+SGD) (Goodfellow et al, 2016)

Multiple linear regression with stochastic gradient descent (SGD) is a method for finding the optimal parameters of a linear model when dealing with a large dataset. SGD is a variant of gradient descent that updates model parameters based on individual data points (or small batches) rather than the entire dataset, making it more efficient for large datasets.

This method applies Multiple Linear Regression, but instead of using direct matrix calculations to find the optimal coefficients, it employs Stochastic Gradient Descent (SGD) for optimization. SGD iteratively updates the model's parameters (coefficients) by calculating the gradient of the error function on one (or a small batch of) data point(s) at a time, making it particularly efficient for large datasets and less prone to getting stuck in local minima in more complex models.

K-Nearest Neighbors (KNN) (Goodfellow et al, 2016)

KNN is a non-parametric, instance-based learning algorithm used for regression (and classification). To predict the value for a new data point, it identifies the 'k' closest data points in the training set (based on a distance metric like Euclidean distance) and then calculates the average (or weighted average) of their corresponding output values. It's simple and can capture non-linearities but can be computationally intensive during prediction for large datasets.

Deep Neural Network (DNN) (Goodfellow et al, 2016)

A Deep Neural Network is an artificial neural network characterized by multiple hidden layers between the input and output layers. Each layer consists of interconnected nodes (neurons) that apply non-linear activation functions to weighted sums of their inputs, allowing the network to learn complex, hierarchical representations and highly non-linear relationships in data. DNNs are trained using backpropagation and gradient descent variants. They excel at complex pattern recognition but require significant data and computational resources

2.2 Observation data

2.2.1 Low Cost PM10 sensor (LCS)

Our LCS is a NOVA SDS011, we have chosen this sensor base on a previous report, in which there is a comparison between different low-cost sensor and a Beta Attenuation Monitor. LSC was installed at the AEMET Headquarters in Santa Cruz de Tenerife.

2.2.2 Air Quality Station (AQS)

For our calibration we need a reference station, on this report we have used Tome Cano, It is a station from Air Quality Network of Government of the Canary Island. Tome Cano has an altitude of 67 meters and its localization is in Santa Cruz de Tenerife (16.26N, -28.46E).

We have chosen this station for two reason: it’s the nearest station to the LCS and it has a good time series.

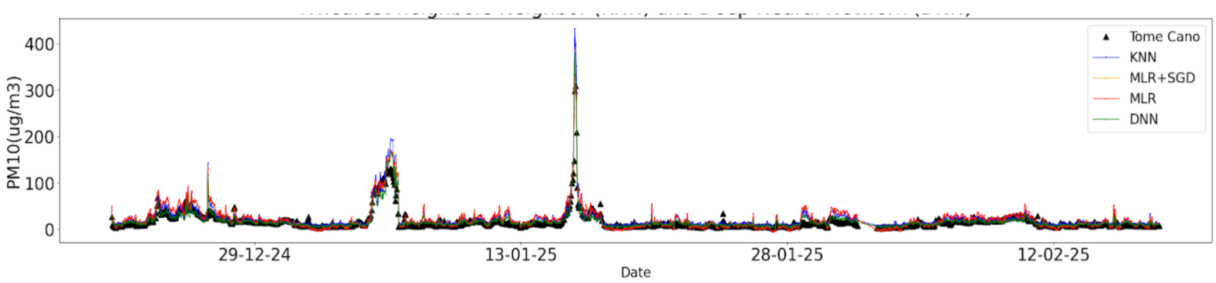
.

3. Results

Dust outbreaks are typically associated with low humidity and high temperatures. Taking advantage of this pattern, we evaluated PM10 concentrations in combination with different environmental parameters: (Figure 1) PM10 alone, (Figure 2) PM10 and humidity, (Figure 3) PM10 and temperature, and (Figure 4) PM10 with both humidity and temperature.

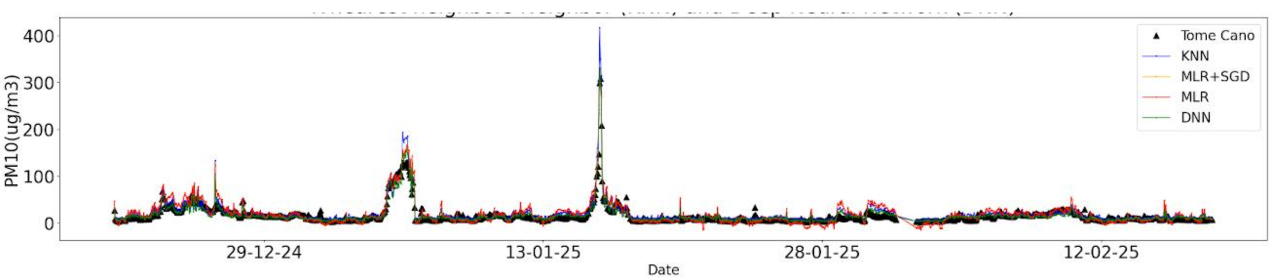
Our objective was to isolate PM10 originating solely from mineral dust, as Air Quality Station (AQS) can also detect soot or other particles resulting from combustion processes. Consequently, the data were filtered based on the following criteria:

- Inclusion of time periods in which at least one significant dust outbreak occurred.
- Data showing a significant discrepancy between LCS and AQS readings were excluded.



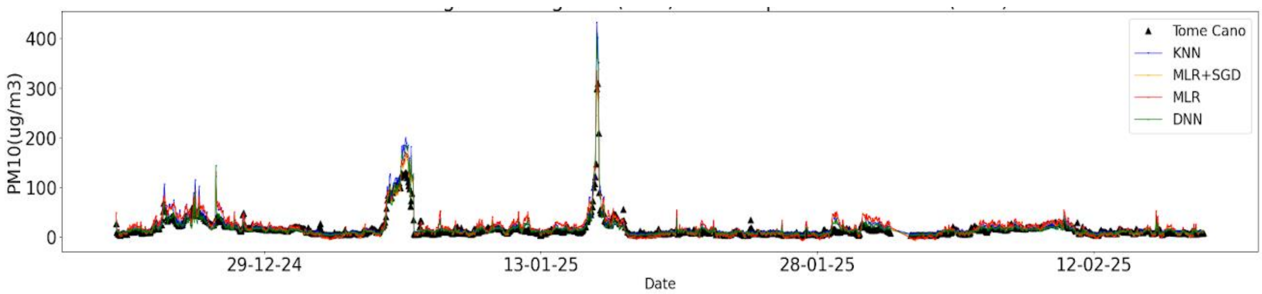
ID	KNN	ID	MLR+SGD	ID	MLR	ID	DNN
Counts	1362	Counts	1362	Counts	1362	Counts	1362
Pearson	0.93	Pearson	0.93	Pearson	0.93	Pearson	0.93
R2 score	0.66	R2 score	0.71	R2 score	0.7	R2 score	0.83
BIAS	5.26	BIAS	4.97	BIAS	4.67	BIAS	0.57
MSE	162.35	MSE	135.45	MSE	142.08	MSE	81.97
RMSE	12.74	RMSE	11.64	RMSE	11.92	RMSE	9.05
MAE	7.07	MAE	7.69	MAE	7.82	MAE	4.75
MAPE	56.5 %	MAPE	57.38 %	MAPE	58.99 %	MAPE	32.84 %
Pm10 = 0.71KNN+1.39		Pm10 = 0.75MLR+SGD+0.63		Pm10 = 0.73MLR+1.26		Pm10 = 0.83DNN+2.46	

Figure 1. Time series comparison of measured PM10 (black triangles) concentrations versus predictions from the KNN (solid blue), MLR+SGD (solid yellow), MLR(solid red) and DNN(solid green).



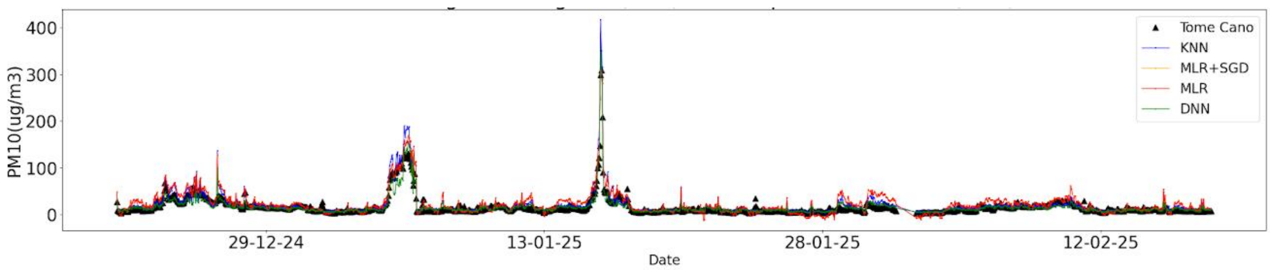
ID	KNN	ID	MLR+SGD	ID	MLR	ID	DNN
Counts	1362	Counts	1362	Counts	1362	Counts	1362
Pearson	0.94	Pearson	0.93	Pearson	0.93	Pearson	0.95
R2 score	0.77	R2 score	0.75	R2 score	0.73	R2 score	0.88
BIAS	3.67	BIAS	2.95	BIAS	2.84	BIAS	-0.29
MSE	108.24	MSE	116.81	MSE	127.46	MSE	54.58
RMSE	10.4	RMSE	10.81	RMSE	11.29	RMSE	7.39
MAE	5.87	MAE	7.37	MAE	7.67	MAE	4.45
MAPE	45.85 %	MAPE	61.38 %	MAPE	64.3 %	MAPE	30.47 %
Pm10 = 0.76KNN+1.41		Pm10 = 0.75MLR+SGD+2.06		Pm10 = -0.49HR+3.64MLR+20.24		Pm10 = 0.9DNN+1.94	

Figure 2. Time series comparison of measured PM10-Humidity (black triangles) concentrations versus predictions from the KNN (solid blue), MLR+SGD (solid yellow), MLR(solid red) and DNN(solid green). In this case, we have added humidity in order to generate the predictions.



ID	KNN	ID	MLR+SGD	ID	MLR	ID	DNN
Counts	1362	Counts	1362	Counts	1362	Counts	1362
Pearson	0.94	Pearson	0.93	Pearson	0.93	Pearson	0.94
R2 score	0.65	R2 score	0.75	R2 score	0.7	R2 score	0.75
BIAS	4.0	BIAS	2.89	BIAS	4.34	BIAS	0.88
MSE	167.47	MSE	116.46	MSE	139.4	MSE	116.78
RMSE	12.94	RMSE	10.79	RMSE	11.81	RMSE	10.81
MAE	6.22	MAE	7.27	MAE	7.76	MAE	4.85
MAPE	40.28 %	MAPE	57.66 %	MAPE	59.08 %	MAPE	28.92 %
Pm10 = 0.67KNN+2.99		Pm10 = 0.76MLR+SGD+2.05		Pm10 = 0.73MLR+1.51		Pm10 = 0.73DNN+4.11	

Figure 3. Time series comparison of measured PM10-Temperature (black triangles) concentrations versus predictions from the KNN (solid blue), MLR+SGD (solid yellow), MLR(solid red) and DNN(solid green). In this case, we have added temperature in order to generate the predictions.



ID	KNN	ID	MLR+SGD	ID	MLR	ID	DNN
Counts	1362	Counts	1362	Counts	1362	Counts	1362
Pearson	0.95	Pearson	0.93	Pearson	0.93	Pearson	0.93
R2 score	0.74	R2 score	0.66	R2 score	0.67	R2 score	0.86
BIAS	3.56	BIAS	6.26	BIAS	6.01	BIAS	-0.91
MSE	124.76	MSE	159.95	MSE	154.84	MSE	66.7
RMSE	11.17	RMSE	12.65	RMSE	12.44	RMSE	8.17
MAE	5.9	MAE	8.64	MAE	8.5	MAE	4.55
MAPE	40.41 %	MAPE	68.0 %	MAPE	67.07 %	MAPE	28.98 %
Pm10 = 0.72KNN+2.4		Pm10 = 0.73MLR+SGD+0.15		Pm10 = -0.57HR+3.64MLR+43.86		Pm10 = 0.91DNN+2.41	

Figure 4. Time series comparison of measured PM10 (black triangles) concentrations versus predictions from the KNN (solid blue), MLR+SGD (solid yellow), MLR(solid red) and DNN(solid green). In this case, we have added temperature and humidity in order to generate the predictions.

Counts	1362	Pearson	R2	BIAS	MSE	RMSE	MAE	MAPE(%)
Normal	KNN	**0.93**	0.62	5.26	162.35	12.74	7.07	56.70
	MLR+SGD	**0.93**	0.71	4.97	135.45	11.64	7.69	57.38
	MLR	**0.93**	0.70	4.67	142.08	11.92	7.82	58.99
	DNN	**0.93**	**0.83**	**0.57**	**81.97**	**9.05**	**4.75**	**32.84**
H	KNN	0.94	0.77	3.67	108.24	10.40	5.87	45.85
	MLR+SGD	0.93	0.75	2.95	116.81	10.81	7.37	61.38
	MLR	0.93	0.73	2.84	127.46	11.29	7.67	64.30
	DNN	**0.95**	**0.88**	**-0.29**	**54.58**	**7.39**	**4.45**	**30.47**
T	KNN	**0.94**	0.65	4.00	167.47	12.94	6.22	40.28
	MLR+SGD	0.93	**0.75**	2.89	**116.46**	**10.79**	7.27	57.66
	MLR	0.93	0.70	4.34	139.40	11.81	7.76	59.08
	DNN	**0.94**	**0.75**	**0.88**	116.78	10.81	**4.85**	**28.92**
H+T	KNN	**0.95**	0.74	3.56	124.76	11.17	5.90	40.41
	MLR+SGD	0.93	0.66	6.26	159.95	12.65	8.64	68.00
	MLR	0.93	0.67	6.01	154.84	12.44	8.50	67.07
	DNN	0.93	**0.86**	**-0.91**	**66.70**	**8.17**	**4.55**	**28.98**

Table 1. Scores for the different methods and parameters. In bold the better scores for each parameter and in gray the best scores.

The Table 1 shows DNN model, particularly its Humidity-PM10 variant, consistently stands out as the top performer across the vast majority of evaluation metrics. It achieves the highest R2 value (indicating the best model fit) and the lowest values for BIAS, MSE, RMSE, and MAE (signifying minimal prediction errors). Furthermore, the DNN (Temperature) variant also demonstrates excellent performance in terms of MAPE (Mean Absolute Percentage Error), making it highly accurate in percentage terms.

In summary, the figures demonstrate that the DNN model serves as a robust and accurate predictor of PM10 concentrations within this dataset, consistently replicating both general trends and extreme pollution events with high fidelity.

The Figure 5 aims to visually assess the DNN model's performance in predicting PM10 concentrations against actual measurements from the 'Tome Cano' station.

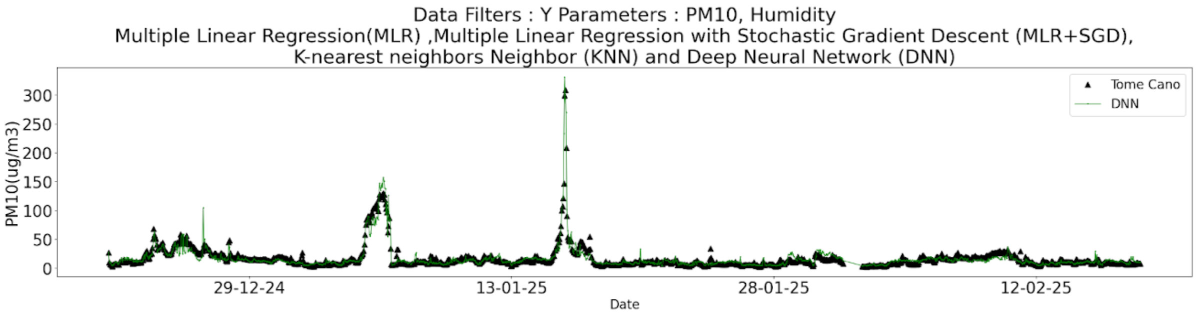


Figure 5. Time series comparison of measured PM10 concentrations versus predictions from the DNN with humidity. The observed data, sourced from the Tome Cano monitoring station, are represented by black triangles. The solid green line indicates the corresponding predictions made by the DNN. In this case, we have added humidity in order to generate the predictions.

The Figure 5 DNN predictions represent the model's strong capability in capturing overall patterns of increase and decrease in PM10 pollution.

The model effectively identifies periods of high mineral dust concentration. Notably, the pronounced peak around January 16, 2025, shows excellent agreement between DNN predictions and actual values, with values exceeding 300 µg/m³.

While the overall fit is good, minor discrepancies exist. The model occasionally underestimates or overestimates peaks.

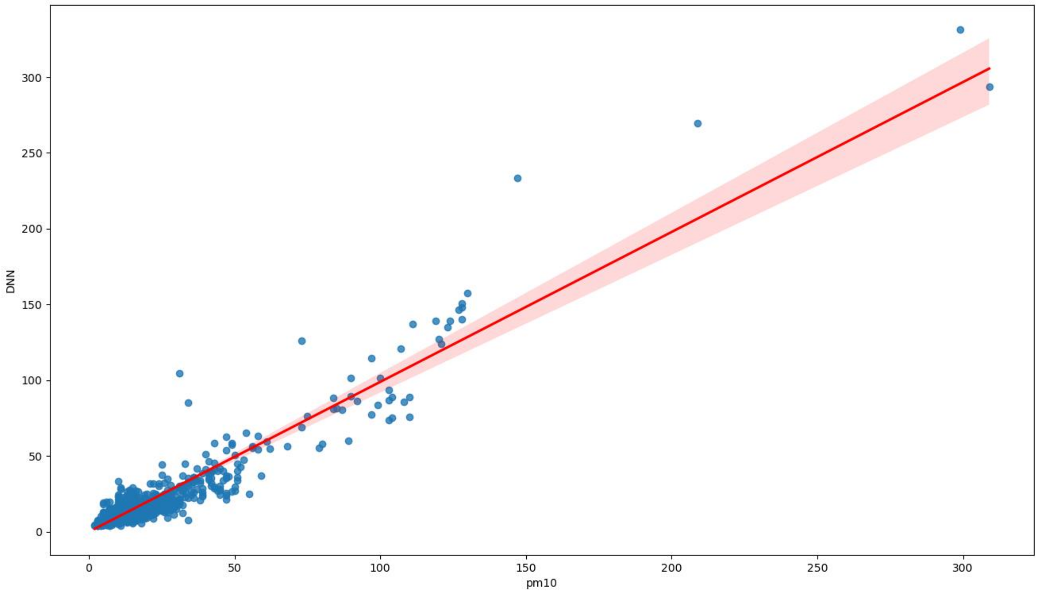


Figure 6. Scatter plot of DNN and PM10 concentration with humidity. The pink shaded area around the regression line is the confidence band. This band indicates the uncertainty in the regression line estimate.

The Figure 6 shows:

The majority of our data have values less than 50 µg/m^3.

DNN has usually less value in comparison with PM10, which is normal because our LCS is higher altitude than the AQS and for this the AQS could detect anthropogenic aerosols. This fact it is only with the minor values, less than 50 µg/m^3.

4. Conclusions

Our analysis of parameter relationships revealed that the Deep Neural Network (DNN) model consistently outperformed other methods across all evaluated scores. Interestingly, the combination of humidity with PM10 generally resulted in the most favorable scores.

The DNN model, specifically its Humidity-PM10 configuration, demonstrated superior performance across the majority of metrics. This variant achieved the highest R^2 values and the lowest (indicating best performance) values for BIAS, MSE, RMSE, and MAE. These results underscore its high accuracy and low error in PM10 concentration predictions.

5. References

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.