REFERENCE: amt-2024-127- "Reviewer 1" Round 2

Title: "*Improving Raw Readings from Low-Cost Ozone Sensors Using Artificial Intelligence for Air Quality Monitoring"*

Authors: *Guillem Montalban-Faet, Eric Meneses-Albala, Santiago Felici-Castell, Juan J. Perez-Solano and Jaume Segura-Garcia*

Departament de Informàtica, ETSE, Universitat de València, Avd. de la Universidad S/N, 46100 Burjassot, (Valencia), Spain

Dear editor and reviewer,

Thank you for giving us the opportunity to address the comments provided by the anonymous reviewers. We have made every effort to respond thoroughly to their feedback. Attached is a response letter with our responses highlighted **in magenta for this Round 2**. The revised manuscript also uses **magenta text** to indicate the changes made, keeping the changes from **Round 1 in blue**.

We would also like to express our gratitude to the anonymous reviewers for their valuable comments and suggestions. We appreciate the time and effort they have invested in improving our work. We firmly believe that this manuscript is now suitable for publication and an excellent contribution to share with the broader research community.

Reviewer's comments (Referee #1 Round 2, March 10 2025) Round 2

First of all, we would like to sincerely thank you for your thoughtful review and comments, which have greatly contributed to improving our work.

In the following sections, we will address all your comments, queries, and suggestions.

Summary: While this draft shows improvement, **more work is needed on the introduction/related work** to set the stage for a strong paper. These sections should clearly set up: 1) what is **already** being **done** in the space; 2) what is **lacking** in the space; 3) what you will do differently to **expand** on what's already been done. There are plenty of other papers already using ML, GB., etc. – what about your model is different?

Likewise, **figures and tables** should be included selectively – many are still **superfluous** and either demonstrate the same information as each other, or information that is already well- established in the field. These figures should be combined or removed as appropriate.

Response 1: Thank you for this comment. We think that the goal and **contribution** of this draft is relatively clear, that is the accuracy improvement of **ground-level ozone** measurements from low-cost sensors but using less expensive air quality monitoring modules, in particular the ZPHS01B multisensor module. The related work and the selected papers used for comparison are using low-cost sensors **ten times more expensive** as it is detailed in the manuscript.

Moreover, since Machine Learning-based algorithms show the best results as discussed in Section 2 in the context of low-cost air quality sensors, in particular for ground-level ozone, **we have focused exclusively on them**, evaluating up to four different models, whereas other studies have only considered one or two. We follow a clear exploratory data analysis, focused on FIA, FS and a detailed HPO process for the different models. Notice that Machine Learning algorithms are the ones that best adapt to the nonlinearities of these sensors, compared to statistical approaches.

In addition, in our models, we include the "**date**" feature (variable), as metadata, as depicted in Section 3.3, which takes into account the effects of aging and detects additional information from road traffic patterns.

Thus, regarding the "**what**" questions:

1) what is already being done in the space; *There are many contributions, and to the best of our knowledge all of them considered in Section 2.*

2) what is lacking in the space; *There is always room for improvement through different aspects: different models and their design, exploratory data analysis, better and different features (variables) and new sensors and platforms to name a few.*

3) what you will do differently to expand on what's already been done: *In our case, we achieve similar or better results with cheaper sensors (10 time less expensive) in an environment with lower ozone concentrations (with a mean value of 55.72 ug/m^3), including all the sensors from ZPHS01B (9 in total) and metadata ("date") in the machine learning process, in a well-defined structured approach for exploratory data analysis. The metadata is used to account for the aging effect and improve the models following road traffic patterns.*

Notice that in addition to our previous manuscript, **we have reviewed and updated the state of the art across various bibliographic databases** from the most important publishers. In particular, we searched for journal publications in IEEE (excluding IEEE Access), Elsevier (ScienceDirect) and Copernicus. Our search focused on calibration methods for low-cost ozone sensors using ML techniques. In practice, there are not that many publications on this topic. Narrowing the search by subject, we found around 50 publications, and after reviewing their contents, we identified only **3 recent references** with a truly similar focus and could be added to update the list of references already included. Briefly, the discarded publications were excluded because they either dealt with tropospheric ozone, integrated additional satellite imaging systems, focused on prediction rather than calibration, used ML for other air quality parameters (without including ground-level ozone), or focused specifically on deep learning (DL).

These new references are:

Cavaliere, A., Brilli, L., Andreini, B. P., Carotenuto, F., Gioli, B., Giordano, T., Stefanelli, M., Vagnoli, C., Zaldei, A., and Gualtieri, G.: Development of low-cost air quality stations for next-generation monitoring networks: calibration and validation of NO_2 and O_3 sensors, Atmospheric Measurement Techniques, 16, 4723–4740, https://doi.org/10.5194/amt-16-4723-2023, 2023.

Wang, G., Yu, C., Guo, K., Guo, H., and Wang, Y.: Research of low-cost air quality monitoring models with different machine learning algorithms, Atmospheric Measurement Techniques, 17, 181–196, https://doi.org/10.5194/amt-17-181-2024, 2024.

Wang, R., Li, Q., Yu, H., Chen, Z., Zhang, Y., Zhang, L., Cui, H., and Zhang, K.: A Category-Based Calibration Approach With Fault Tolerance for Air Monitoring Sensors, IEEE Sensors Journal, 20, 10756–10765, https://doi.org/10.1109/JSEN.2020.2994645, 2020.

These 3 new references included have been explained and included in Section 2 as follows:

The calibration process of these LCS is a challenge as mentioned before, where ML and Deep Learning (DL) models can be used. **In (Wang et al. (2024)), a low-cost multi parameter AQ system based on $PM_{2.5}$, PM_{10}, SO_2, NO_2, CO and O_3, along with Temp and RH is proposed using and evaluating various calibration algorithms. For O_3, the algorithms are ranked from best to worst fit as follows: RF, K-Nearest Neighbors (KNN), Back Propagation (BP), Genetic Algorithm Back Propagation (GA-BP), and Multiple Linear Regression (MLR), with R^2 values (MAE, in $\mu g/m^3$) of 0.98 (2.88), 0.87 (7.33), 0.83 (11.14), 0.83 (10.90), and 0.74 (13.46), respectively. With a mean O_3 concentration of approximately 70 $\mu g/m^3$, as shown in their Figure. 12, the RF model achieves a MRE of 4.11%. In (Cavaliere et al. (2023)), based on O_3 and NO_2 metal oxide sensors, along with Temp and RH, the authors analyzed different calibration options using uni-variate/multi-variate, linear/non-linear and parametric/non-parametric approaches with algorithms such as Linear Regression (LR), Non-Linear Regression (NLR), Support Vector Machines (SVM), RF and GB. They concluded that Multiple Random Forest (MRF) achieved the highest accuracy during Phase I (pre-deployment), with an R^2 of 0.98 and MAE (MRE) of 4.31 (5.74%), considering a mean O_3 during their deployments of 75 $\mu g/m^3$, depicted in their Figure 3 and 7. However, in Phase II (field validation) conducted at a different location, the performance worsened, with the MAE (MRE) 22.22 (29.62%) while MLR 12.96 (17.28%). In this case, MLR provided better results. The authors conclude that MLR may be a more suitable solution for representing physical models beyond the Phase I calibration dataset, demonstrating better transferability across diverse spatial and temporal settings, highlighting that parametric models such as MLR have a defined equation with only a few parameters, making them easier to adjust for potential changes over time. In (Wang et al. (2020)), the authors propose a category-based calibration approach (piecewise) using ML, which builds separate regression models for different pollutant concentration levels. This proposal is tested on CO and O_3 data from two Chinese cities, Fuzhou and Lanzhou, with good and bad AQ, with mean O_3 concentrations of 69.545 $\mu g/m3$ and 49.781 $\mu g/m3$ respectively for 11 months (48 weeks). The achieved metrics for the best results are given by Extreme GB and Light GB machine algorithms (outperforming linear regression and RF) with MAE ($\mu g/m3$) of 10.75 and 10.98 in Lanzhou city respectively, and 13.83 and 14.98 in Fuzhou city, with a MRE greater than 19.88%.**

And in Table 11 (in the new version), we have included also these 3 references (the first 3 row) for comparison as follows:

Study	Location	Platform, Sensor	R^2	MRE [%]	Comment
(Wang et al. (2024))	Zhengzhou (China)	by Hanwei Electronics Corp, O_3 B4 Alphasense	0.93	4.11	52 weeks dataset with RF and HPO
(Cavaliere et al. (2023))	Florence, Montale (Italy)	AirQino LC, NO_2 MiCS-2714, O_3 MiCS-2614	0.98 with MRF	I:MRF 5.74. II:MLR 17.28, MRF 29.62	61 weeks dataset with MRF and MLR, using complete EDA
(Wang et al. (2020))	Lanzhou (China)	Sailhero instrument, -	-	19.88	48 weeks dataset, category-based calibration (piecewise) with Extreme GB and FS
(Zimmerman et al. (2018))	Pittsburg (USA)	RAMP, Alphasense Ox-B431	0.86	15	16 weeks dataset with RF
(Esposito et al. (2016))	Cambridge (UK)	SnaQ, Alphasense B4 Electrochemical	0.69	42	5 weeks dataset using Dynamic NN with a kind of HPO
Our model	Valencia (Spain)	ZPHS01B, Winsen ZE27	0.93	7.21	57 weeks dataset using GB with FIA, FS and HPO

Table 11. Comparison with similar related works

We agree that the Machine Learning techniques used are not innovative, but their use as a tool to achieve the accuracy shown in our results, combined with the methodology, following the recommendations and best practices reported in other scientific works and dealt in the manuscript, make it relevant and interesting for the community (i.e. for researcher and practitioners). These steps and methodology, as well as their explanation and justification are not found in the related work as is highlighted in the new version of the manuscript, as stressed in the previous revision.

We emphasize that some of the steps for data preprocessing, analysis and interpretability are often overlooked, such as Feature Importance Analysis (FIA), Principal Component Analysis (PCA) and Feature Selection (FS). In this line, in the manuscript it is said that the process of optimizing algorithms through the selection of their hyperparameters is also neglected in some environmental research studies. As we mentioned before, all these details are already included in Section 2 “Related work” and checked with these representative papers.

It is worth mentioning that we are combining two different disciplines, air quality and artificial intelligence. And it is difficult to master both disciplines and this is the reason in Section 2 we go into detail with these aspects, checking if the procedures used in the related works are overlooking these steps in these papers.

Thus, **in summary our work provides several contributions of interest,** which we list below:

1. **The multisensor module ZPHS01B is priced at approximately 150 euros and includes 9 different sensors**, making it cheaper than other systems used in the state of the art. This is a distinguishing starting point and of interest to the scientific community. In the related work, systems considered low-cost typically refer to a price of less than 150 euros **per sensor**.

2. **The approach of using a single module with all 9 integrated air quality sensors**, which **enables the evaluation of cross-sensitivity issues** between sensors and their potential added value, is another differentiating element. In the state-of-the-art works reviewed, the systems generally use separate sensors of different types and features, which can be interchanged. We have as many different ozone sensors as there are papers on the related work.

3. **In the studies compared in Table 11 of the new version of the manuscript**, various methods are used for calibration, **including statistical methods (also known as white-boxes) and machine learning (ML)-based approaches (grey-boxes)**, the latter of which tend to yield better results. However, in those comparisons with **ML, only one or two methods are usually applied.** In our case, **we focus exclusively on machine learning and perform a more in-depth evaluation of four different methods**.

4. **Due to the design and characteristics of these low-cost sensors, aging affects their performance over time.** Only in the mentioned article by (Cavaliere et

al.(2023)), it is proposed an adjustment for a linear regression method to account for this, but the adjustment is not applied to other methods, especially not to those based on machine learning. In our machine learning models, **this effect is incorporated through the "date" feature**, which—while technically a **metadata** field—helps reduce error in the models by not relying solely on environmental variables. This feature allows us to capture both the effect of sensor aging and pollution patterns associated with traffic from combustion vehicles. Besides, this feature (date) will allow us to improve the models following the road traffic patterns.

5. **We already evaluated Deep Learning (DL) methods at the request of Reviewer 2 in Round 1 of the revision process** (as can be seen in the discussion forum of the platform), and we also extended the initial dataset from 165 to 239 days. Although DL results are not directly included in the article, we provide a relevant discussion explaining why such methods were discarded in this scenario, as noted in section 3.4. In fact, **we observed that for the datasets obtained during the measurement campaigns, DL models tend to learn and memorize the dataset entirely, leading to overfitting.** We believe this is due to the intrinsic characteristics of the air quality monitoring scenario and the behavior of the low-cost sensors in the ZPHS01B module, **as the datasets generated are limited and constrained for the use of DL techniques.** This is the reason we do not include these results. These results are shown in the response to this Reviewer 2 with R^2=0.9999999999976741, RMSE: 3.514481801502949e-05 and MAE: 2.9925663790820442e-05.

Besides, notice that Machine Learning (ML) models are often more practical, efficient, and interpretable for time-series prediction tasks, especially when datasets are small-to-medium-sized, contain noise, or require explicit domain knowledge. While DL models like LSTMs and TCNs excel in capturing long-term dependencies in very large datasets. Thus, with our dataset, we observed that DL techniques are not able to generalize as the ML approach did. And for this reason, the results using DL techniques are not so robust and reliable, mainly due to overfitting.

6. The results presented share the same characteristics as those presented in **Table 11 of the new version of the manuscript** used for comparison. However, there are two important differentiating elements. On the one hand, as mentioned before, the module used is significantly more affordable (around 10 times cheaper). On the other hand, **the results reported in the referenced works (Wang et al. (2024)) were obtained in environments with much higher ozone concentrations**. It is important to note that these sensors perform worse at **low concentrations** than at high ones due to their sensitivity limitations and the weakness of the signals generated, as well as **interference from other pollutants**. While in our dataset the ozone concentration is lower, with an average ozone concentration of **55.72 µg/m³**—i.e., in the cited studies the values are higher (more than 70 µg/m³). This information is included in the new version of the manuscript as follows:

not all of these works follow and discuss an structured EDA with FIA, FS and HPO. In particular, when compared to the first two works with slightly better results, in (Wang et al. (2024), we appreciate higher O_3 values, mean values higher than 70 $\mu g/m^3$, while in our case we have lower levels (55.72 $\mu g/m^3$)), as well as there is not a complete EDA. It is important to note that these sensors perform worse at low concentrations than at high ones due to their sensitivity limitations and the weakness of the signals generated, as well as interference from other pollutants. Finally, in (Cavaliere et al. (2023)), although the authors use a complete EDA, they only use two sensors (NO_2 and O_3) apart from Temp and RH, and the

7. Finally, although there are a couple of articles that follow a more structured approach, in particular (Cavaliere et al.(2023)), **most do not carry out the recommended steps required to properly apply machine learning algorithms**, such as conducting exploratory data analysis and including **Feature Importance Analysis (FIA), Feature Selection (FS) and Hyperparameter Optimization (HPO)** stages.

All these comments have been incorporated into the wording of the new version, particularly in the abstract, at the end of Section 2 (Related Work), and in the conclusion.

Finally, about the figures and tables included in the manuscript, they are discussed in the following responses.

We have clarified this issue in the new version of the manuscript.

Subscripts are needed throughout for O_3, CO_2, etc.

Abstract: Readers will know what ozone is. This space would be better spent explaining why **you need machine learning enabled calibration**.

Response 2: Thank you for your comment. We have updated the subscripts for the chemical formulation.

In the abstract, regarding the calibration process in general, it arises from a lack of accuracy in low-cost ozone sensors. However, we have improved the wording for clarity as well as a better justification of the machine learning techniques used in this case, but in a brief form for the abstract, as follows:

Abstract. Ground-level ozone (O_3) is a highly oxidizing gas with very reactive properties, harmful at high levels, and generated by complex photochemical reactions when primary pollutants from the combustion of fossil materials react with sunlight. Thus, its concentration indicates the activity of other air pollutants and plays a crucial role in smart cities. **With the growing interest in high-resolution Air Quality (AQ) monitoring, low-cost ozone sensors present an interesting alternative, although they lack accuracy and suffer from cross-sensitivity issues. In this context, artificial intelligence techniques, particularly ensemble Machine Learning (ML) models, can improve the raw readings from these sensors by incorporating additional environmental information to minimize inaccuracies and nonlinearities, as well as by including metadata to account for sensor aging effects and improve the models based on road traffic patterns.** In this paper, **based on the low-cost ZPHS01B multisensor module with nine sensors, we analyze, propose, and compare different techniques using four ML models in a low O_3 concentration scenario (mean value of 55.72 $\mu g/m^3$). We carried out a thorough exploratory data analysis process to extract the main features (variables) and performed hyperparameter optimization for the different models.** As a result, we reduced the estimation error by approximately 94.05%. In particular, using the Gradient Boosting algorithm, we achieved a Mean Absolute Error (MAE) of 4.022 $\mu g/m^3$ and a Mean Relative Error (MRE) of 7.21%, outperforming related work while using a module approximately ten times less expensive. **To carry out this work, we generated two datasets in the city of Valencia (Spain), at two different locations with the same characteristics (close to the ring road but separated by 4.1 km), with 165 and 239 days.**

Line 14: do the authors ever come back to these **guidelines**? If not, this paragraph is not useful. Same with the next paragraph – these standards are not really mentioned again later. I understand that this is trying to establish the "why use low-cost sensors", but it needs to be more clearly related back to what you're actually doing.

Response 3: The reviewer is correct. In Section 1, we introduce, motivate, and contextualize the problem of ground-level ozone based on the Air Quality Guidelines and the objectives set within the mentioned directives. This approach is twofold: on one hand, we focus on air pollutants (particularly ozone) and their impact on health, and on the other hand, we emphasize the importance of higher spatial monitoring resolution for these pollutants.

However, we have specifically improved the wording in the Conclusion section to address this issue and revisit the problem statement covered in this manuscript, as introduced in Section 1. Thus, in the Conclusion, we refer to these standards and guidelines again for closure.

5 Conclusions

This paper focuses on ground-level ozone (O_3), as it serves as an indicator of other pollution levels in urban areas using LCS nodes based on the ZPHS01B module. **These nodes will enable an increase in the spatial sampling of AQ monitoring in cities, following the interest of AQG (Organization et al. (2021)) and in line with the future plans of the related directives, ideally at least one sample per 100 m^2, according to Annex III-B of the European (Directive 2008/50/EC (2008)).**.

Line 27: Why is "**primary**" in quotes but not secondary? Be consistent, but quotes are not necessary. There are also quotes around primary on line 2.

Line 34: What is "**official** equipment"?

Line 51: While not detrimental, this paragraph is unnecessary.

Response 4: Thanks for these corrections. We have removed the "quotes" for primary and secondary. About "official equipment" expression, maybe the term official is not

adequate, and it should be better “regulated”, “certified” or “standardized”. We have explained this and changed this expression. Thus, we refer to regulated equipment, when we refer to “standardized air quality monitoring stations”.

The paragraph in line 51, although it is often found in the research papers, to assist the reader, we have omitted it. If the editor considers incorporating it, it has been just commented % in latex in the source files.

We have revised this expression in the manuscript accordingly.

Related works: see comments about **table 15**, but the information on the specific other **sensors** used for comparison could be restructured, if not removed. The information added here on specific ML **models here is helpful, but it could be improved further by exploring more clearly the strengths and weaknesses of each of these**, and how you will improve upon this and not just repeat what’s already been done.

Response 5: Thanks for this comment. Regarding this table, now **Table 11 of the new version of the manuscript**, we have included some extra details for further information for clarity. Notice that this table only considers some of the modules shown in Table 1, in particular RAMP, AirSensrEUR and ZPHS10B.

Moreover, we highlight that the goal of Table 1 is to compare different commercially available low-cost multisensor modules and alternatives, detailing only their sensors and price range, without considering whether all these modules have been used in related work.

Finally, about the strengths and weaknesses of the related work, it was already considered in the previous review, in blue as follows:

good example of the use of these good practices is shown in (Cavaliere et al. (2023)). In addition, in (Zhu et al. (2023)), it is said that the process of optimizing algorithms through the selection of their hyperparameters (Hyperparameter Optimization (HPO)) is neglected in most of the environmental research studies considered. For instance, in (Johnson et al. (2018)), better results are obtained with GB, but HPO is not performed in the model, which could allow further improvements of the results. In (Malings et al. (2019), (Wang et al. (2020)) and (Zimmerman et al. (2018)), it is taken into account some aspects related to the data analysis focused on the optimization of the problem, but they do not carry out a HPO. In (Esposito et al. (2016)), the authors carry out a kind of simple HPO, based on raw tests of different architectures and modifying hyperparameters, such as the number of hidden layers of the model, tapped delay length and feedback delay line length, concluding that a dynamic approach to these parameters improves the results with respect to a static approach without changing the value of these parameters.

Regarding the selection of parameters, in (Johnson et al. (2018)), the authors does not perform an analysis using techniques such as the aforementioned FIA and FS, but a sensitivity analysis using different meteorological variables (such as Temp and RH), determining that it is useful information for GB. In (Malings et al. (2019)), the quantification of the importance of the model variables is mentioned as a mean to understand which information is useful, concluding that for RF, to add additional information apart from AQ measurements, such as Temp and RH are very helpful. In (Esposito et al. (2016)) and (Wang et al. (2024)), the authors do not include a specific analysis of the relative importance of different variables or features. However, a good example of FS is depicted in (Okafor et al. (2020)), where it is shown that identifying the environmental factors affecting LCS is crucial for improving data quality using data fusion and ML. These factors are then incorporated into the development of the calibration model.

In conclusion, in order to increase the resolution of city-scale AQ monitoring according to the recommendations given by (Directive 2008/50/EC (2008)) as mentioned before, it is necessary to perform a calibration process of these LCS. In this scenario, we focus on O_3 calibration using ensemble ML techniques to minimize inaccuracies and nonlinearities, comparing four different models, considering different environmental variables as well as metadata mainly to account for sensor aging effects. For this purpose, it is necessary to carry out a thorough data treatment with a good practice criteria (Zhu et al. (2023)) including HPO, FIA and/or FS, which are usually overlooked. In a scenario with low O_3 concentration, we achieve interesting results compared with the related work, as shown in Section 4.

Table 3: This table should be removed. There are still no units in this table (temp, RH, PM2.5, CO2, NO2, CO, etc. should all have units attached). The statistics of the measured quantities are not referenced or used anywhere else in the paper, and the reader can't do anything with this information on their own. **Likewise, "stationarity" and "percentage of samples taking Different values" are not analyzed further in the text**. The paragraph beginning on line 181 can be condensed to give the context the table is hoping to provide (**ex. "Sensors X, Y, and Z appeared particularly unreliable and were omitted from our model"**.)

Response 6: Thanks for this comment. We have introduced these units both in the caption and in the text.

However, regarding the content of this table, note that the ZPHS01B model has not been previously used for these issues. For this reason, it is important for us to justify our choice and provide all relevant details and evidence to clarify and characterize its behavior.

The statistical analysis conducted with the datasets may seem redundant if using the same module as other research groups or well-known sensors, but this is not our case. Nonetheless, **we have simplified this table** by removing ' Variance (Var.), Stationarity (Stat.) and 'Seasonality (Seas.)', retaining the more relevant statistics.

In addition, these statistics are used in the results section to calculate additional metrics and parameters, in particular when we estimate the mean relative error.

The new version of this table 3 is as follows:

Table 3. Summary of main statistics of the Dataset: Minimum (Min.), Maximum (Max.), Mean (Mean), Standard Deviation, Median Absolute Deviation (MAD), percentage of samples taking Different values (Diff.) and High correlation (High corr.)

	Temp [ºC]	RH [%]	$PM_{2.5}$ $[\mu g/m^3]$	CO_2 [ppm]	NO_2 $[mg/m^3]$	CO $[mg/m^3]$	CH_2O $[mg/m^3]$	TVOC [levels]	O_3 $[\mu g/m^3]$	O_3ref $[\mu g/m^3]$
Min	5.24	62.29	21.25	693.43	0.78	0	0.005	0	39.57	8.71
Max	42.26	118	83.69	1792.50	18.81	0.75	1.21	2.95	255.76	97.85
Mean	20.60	91.31	49.99	780.33	15.27	0.34	0.021	0.024	114.39	55.72
SD	5.70	18.12	18.14	57.16	5.65	0.28	0.02	0.13	67.11	24.83
MAD	3.92	16.37	13.31	24.53	0.59	0	0.001	0	51.40	16.21
Diff.	99.1%	81.9%	87.9%	97.5%	50.6%	0.2%	81.2%	5.8%	75.0%	30.3%
High corr.	yes	yes	yes	yes	not	yes	not	not	yes	yes

Figure 3: Any ambient pollutant will have a **repeating diurnal pattern** from the boundary layer rising and falling each day, and most sensors will pick up on major sources like traffic.

A DFT is not necessarily needed to show this and confuses the messaging in this section. Since this figure is never referenced again other than to show that a pattern exists, it should be omitted.

Response 7: Of course, a repeating diurnal pattern associated with the day/night cycle is evident once we analyze the DFT, as it reveals ground-level ozone generation through photochemical reactions.

However, when the selected sensors are under test (especially in a low-cost approach like this) and this analysis has not been previously performed, we do not consider this check redundant. We believe that we should not assume certain patterns as obvious without verification.

It is possible that this pattern does not exist or cannot be detected with these sensors, which is precisely why we applied this analysis. In our dataset, particularly for ozone, the pattern is clearly observable, and this method provides a straightforward way to demonstrate it.

Perhaps this analysis has not been included in previous related work because researchers have used well-known low-cost sensors.

That said, we are open to removing this information if the editor deems it unnecessary.

Meanwhile, **we have placed this information in "Appendix A:** Spectral analysis for O3 low-cost readings from ZPHS01B module" of the new version of the manuscript.

Figure 4: While this figure is fine, it's well known in the low-cost sensor space that sensors can capture the general trends of pollutants but need calibration to accurately convey the magnitude. This figure should be **omitted**.

Response 8: As we stated previously, the ZPHS01B module has not been used before in this kind of studies and research. Note that we selected this module for several reasons, as explained in the manuscript. It offers the best price-per-sensor and price-to-quality ratio, embedding 9 different sensors on the same board.

Thus, the data provided by these sensors is valuable for analyzing cross-sensitivity issues, enabling the training of different calibration models and extracting more information than would be possible with single sensors.

For this reason, examining and demonstrating the behavior of the O_3 sensor in this module is particularly relevant. For instance, we observe a positive offset in the raw readings compared to the regulated and standardized O_3 measurements from the AQ station. This trend was also reflected in the error distribution shown in Figure 8, that finally was removed in the new version of the manuscript as it is suggested later in Response 15.

Table 4: If this is all to make a better ozone model, the **FIA** of ozone should be included here to show how much it improves the model. How was **8% importance** selected? It sounds arbitrary. It would also be easier on the reader if the threshold and the table were in the same format (either **both in decimal or both in percent)**.

Response 9: Thank you for your comment. Table 4 presents the normalized output of the FIA using the scikit-learn library for parameters complementary to ozone, for each model used. For clarity, all contributions are expressed per unit (1). From this table, we observe the following:

- On one hand, **Temp, RH, and CO_2** exhibit higher contributions compared to the other parameters. We have highlighted these values in bold.

- On the other hand, **NO_2, $PM_{2.5}$, CH_2O, TVOC, and CO** show lower contributions, falling below the suggested heuristic threshold of 0.08 (8%), as no other criterion applies in this case. Additionally, NO_2, CH_2O, TVOC, and CO were already discussed in the analysis of Table 3, except for $PM_{2.5}$.

We have **refined the wording and improved this table** in the manuscript regarding its analysis as follows:

Table 4. FIA of ozone's complementary parameters **for Random Forest (RF), Gradient Boosting (GB), Adaptive Boost (ADA) and Decision Tree (DT)**, in bold the selected ones, contribution higher than 0.8.

Model	Temp	RH	$PM_{2.5}$	CO_2	NO_2	O_3ref	CO	TVOC	CH_2O
RF	**0.128**	**0.103**	0.069	**0.222**	0.078	**0.269**	0.002	0.003	0.064
GB	**0.107**	**0.105**	0.052	**0.211**	0.057	**0.253**	0.001	0.001	0.068
ADA	**0.119**	**0.097**	0.064	**0.246**	0.067	**0.287**	0.001	0.001	0.066
DT	**0.115**	**0.088**	0.070	**0.232**	0.061	**0.276**	0.001	0.002	0.061

230 Table 4 shows the normalized output of the FIA using the *scikit-learn* library (Pedregosa et al. (2011)), for the parameters complementary to O_3, for each ML models used. **In order to determine the most useful parameters for the models, a threshold is established in 0.08, that is 8% of importance. These parameters are in bold. Notice that the set of parameters with the highest importance, is repeated for all models: Temp, RH, CO_2 and O_3.**

Besides, we must highlight that these are preliminary steps, and it does not mean that we directly will exclude these parameters with lower contribution at this point.

Figure 5 is essentially showing that some sensors are more cross sensitive than others, which is already well established in the field. This figure should be **omitted**.

Response 10: In our opinion, this figure could be considered redundant when dealing with well-known and well-characterized sensors. However, this is not our case. In line with the arguments mentioned in Response 8, the ZPHS01B module has not been used in a similar way before. Thus, the **information provided is valuable for analyzing cross-sensitivity issues. This type of information is part of the exploratory data analysis (EDA) and feature selection (FS).**

Tables 5, 6, 7 and 8 should be combined into a single table with the 4 sub-categories as another column.

Response 11: Since we are dealing with four different models, each with different hyperparameters (both in number and meaning), it is clearer to present this information in separate tables. These tables are different and **cannot be combined** in a clear way.

Table 9 is unnecessary and can be omitted – you and many others have already established that **hyperparameter tuning will make the models fit better**.

Response 12: Thanks. We have **omitted** this table and left only the optimized versions. Simply, we have just introduced a sentence detailing how much improvement the hyperparameter optimization introduces in the different models, as follows:

320 **It is worth mentioning that the improvement achieved by HPO is greater in GB and ADA models than in RF and DT, which are already well-optimized with default values. In particular, for the optimized GB and ADA models, R^2 is improved by 42% and 182%, respectively, while RMSE is reduced by 57% and 66%. However, the execution time required for training is influenced by HPO, increasing to 66.937s and 7.805s for GB and ADA, respectively, as shown in Table 9. We highlight that RF and DT are already well-optimized, and their execution times remain unchanged between**
325 **the default and optimized versions.**

Tables **10 & 11** should also be **combined**.

For tables 9, 10, and 11, and Figure 6, it is not specified in the titles whether it is **training or testing data** – please specify.

In the low-cost sensor field, **it is standard to show both training and testing data - consider adding to tables 9, 10, and 11, and Figure 6.**

Response 13: Thanks for your comment. We have combined both tables in one as follows:

Table 9. Performance metrics for HPO models with 90/10 and 80/20 (training/testing) ratio

Model	GB		RF		ADA		DT	
Ratio	90/10	80/20	90/10	80/20	90/10	80/20	90/10	80/20
R^2	0.938	0.936	0.927	0.924	0.922	0.920	0.878	0.863
RMSE	6.492	6.664	7.093	7.253	7.289	7.416	9.149	9.735
MAE	4.022	4.221	4.185	4.415	3.642	3.833	4.684	5.104
MAPE	0.194	0.206	0.208	0.228	0.160	0.175	0.206	0.226
Time	66.937	61.054	18.316	16.618	7.805	7.078	0.212	0.194

The results shown are **always for testing data** as it is detailed in the manuscript. We do not show the training process. However, in the next response 14 we will discuss this issue again, proving the results from training and validation.

Notice that we split the dataset for training and testing, both sets remain independent and isolated with different training-test ratio percentages: 60%-40%, 70%-30%, 80%-20% and 90%-10%.

And during the training process itself, the dataset is further divided into two parts: one for training and the other for validation. By default, we allocate 80% of the data for training and 20% for validation. In this process, the training and validation datasets are combined across different iterations.

We have improved the wording to clarify this issue in the new version of the manuscript as follows:

4 Results

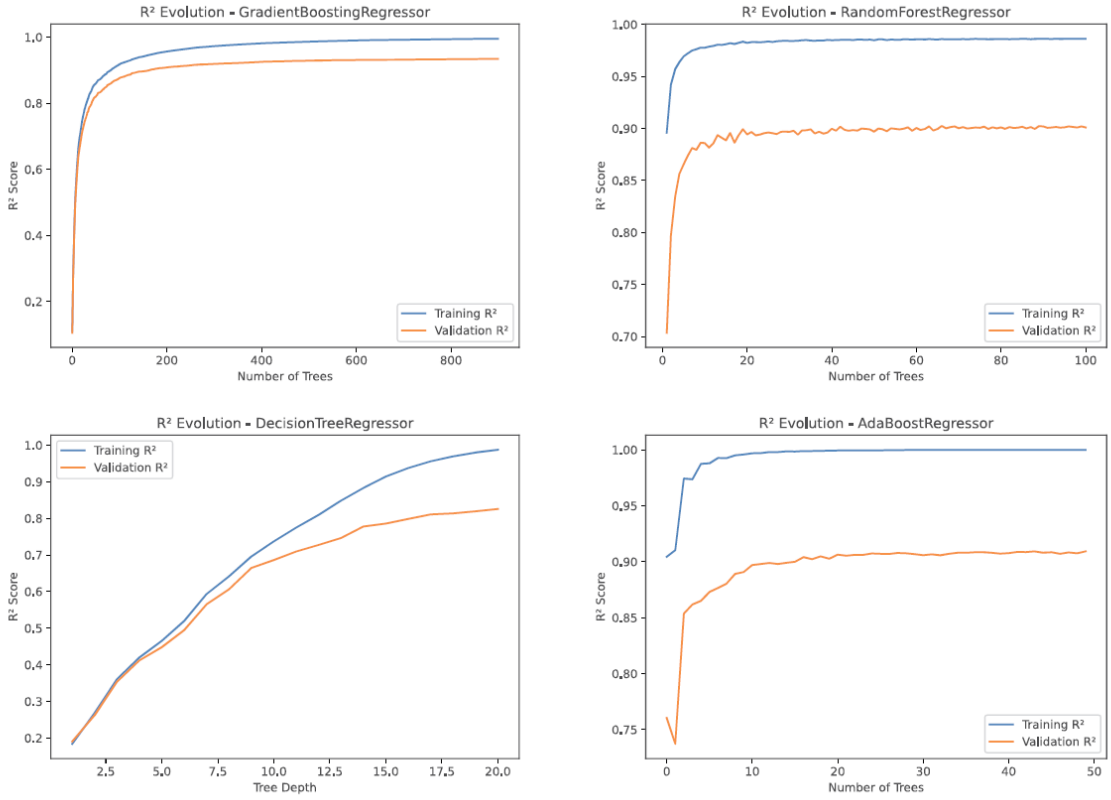
We evaluated the performance metrics of these ML models under different configurations (in terms of R^2, RMSE, MAE in $\mu g/m^3$ and Mean Absolute Percentage Error (MAPE) and execution time in seconds), with the optimized hyperparameters that achieve higher R^2 and lower errors. Also, we used the three different datasets given by different monitoring intervals: 10 and 30 min and 1 h, as depicted in Section 3.2. We tested different training-test ratio percentages from these datasets: 60%-40%, 70%-30%, 80%-20% and 90%-10%, denoted as 60/40, 70/30, 80/20 and 90/10. **Note that when we split the dataset for training and testing, both sets remain independent and isolated. However, during the training process itself, the dataset is further divided into two parts: one for training and the other for validation. By default, we allocate 80% of the data for training and 20% for validation. In this process, the training and validation datasets are combined across different iterations. From all of them, we have achieved the best results in terms of these performance metrics with**

Is the point of Figure 7 just to show that the model isn't **overfitting**? It needs more analysis in the text rather than relying on the reader to interpret.

Response 14: As mentioned above, in Response 13, **we only show the results from testing data.**

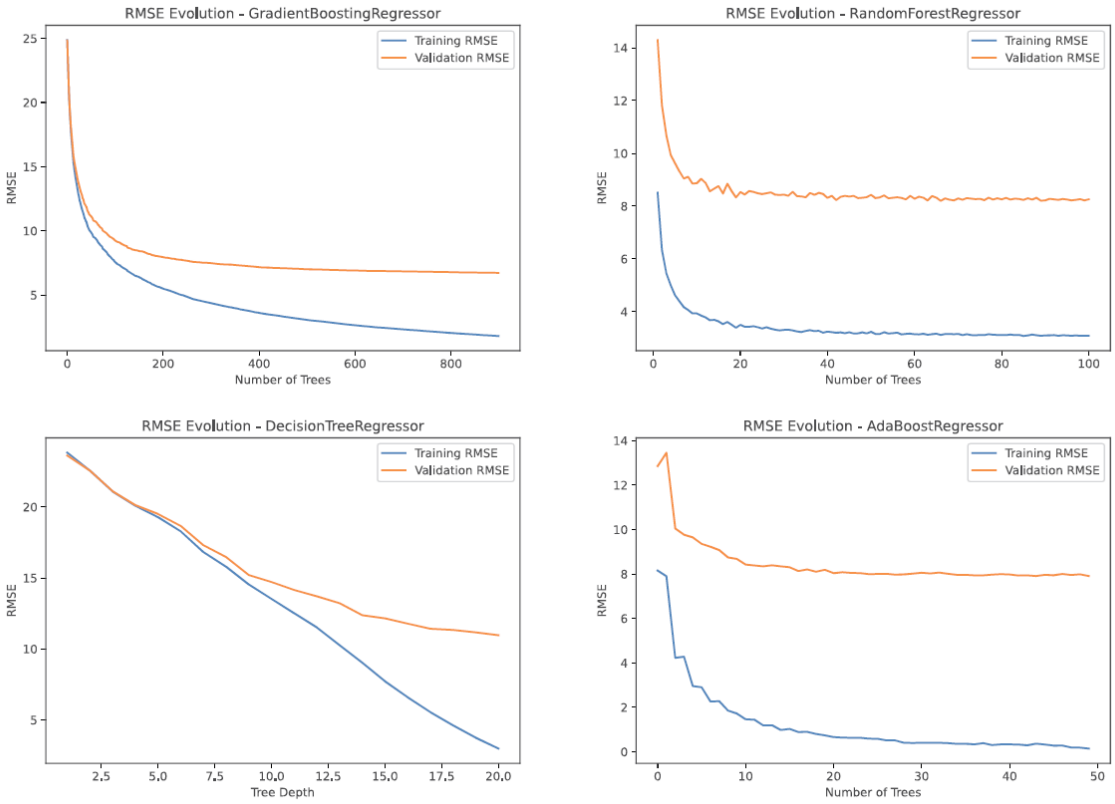
However, in the training process, the used dataset (excluding testing dataset) is further divided into two parts: one for training and the other for validation, by default 80% for training and 20% for validation respectively during the different iterations.

Since the convergence of performance metrics provides information about overfitting for both the training and validation datasets, we have included the following plots, which show the R² and RMSE values across different iterations during the training process for various models. Each model uses a reference hyperparameter for convergence.



We can observe in the above plots during the different iterations the fit of the model in terms of R^2, with a better fit with training than with validation, as expected. In addition, it should be noted that the convergence process with training does not reach a perfect fit in any case, which justifies and supports the conclusion that there is no overfitting in the models.

Moreover, we see that the achieved R^2 score for both training and validation is better than the values shown for testing, which are the ones included in the tables in the manuscript. That is because the testing dataset does not participate in the training process.



About RMSE, the above plots show a similar behavior during different iterations, with a better fit in training compared to validation. As mentioned before, these values are better than those shown in the table from the testing process.

All this information, figures and explanation, has been included in “Appendix B: Results of models’ convergence”.

Again, it is well-established that low-cost sensors need calibration, and that tuning will improve models. **Figure 8 should be omitted**. **If you are insistent on including something like this, an analysis showing the statistical significance in model improvement might be more impactful.**

Is there more analysis or more takeaways to be had from **Table 12**? All the **text** is really saying is that the numbers in the table match the numbers in the figure. Stronger analysis in the text is needed to make the table worth keeping.

Response 15: Thanks. We have omitted Figure 8 and Table 12, which illustrates the error distribution and its analysis, related to standard deviation and confidence intervals.

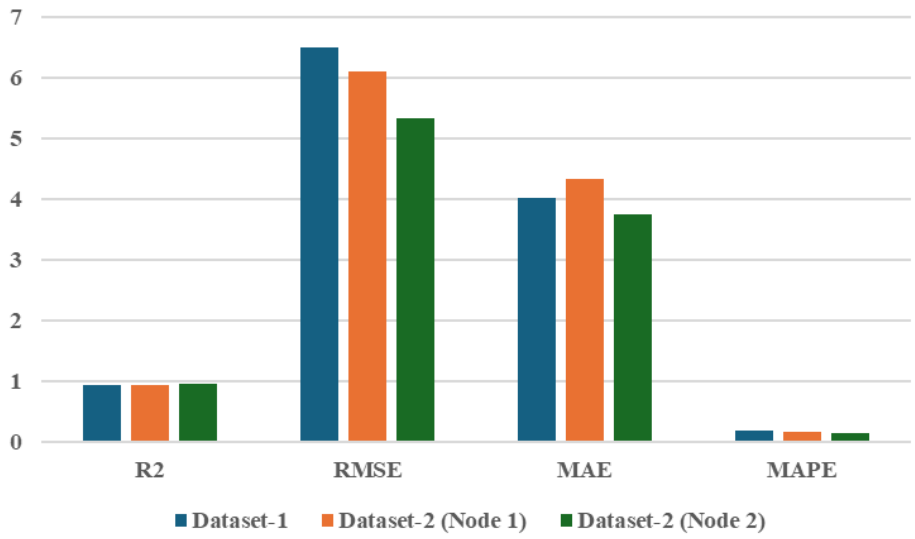
We found it interesting to observe how the adjustment in the calibration process is carried out based on the raw readings, allowing us to identify default deviations and tendencies directly from the embedded sensors in the ZPHS01B module. Since this module has not been used before, as previously mentioned, this information could shed light on important insights into its performance.

Besides, a statistical study based directly on this distribution is more robust and comprehensive. For instance, we could observe how the offset shown in the raw readings in Figure 4 appears as an asymmetry (skewness) in the error distribution.

However, we reconsidered and concluded that the information provided could be omitted, and it was removed.

Is there a better way to visualize the information in **table 13**? It's inclusion is helpful, but a **figure** could be more informative than a table.

Response 16: Thanks. Table 13 summarizes the metrics provided by the Gradient Boosting model for the different datasets used for generalization. In this case, we present the performance metrics for Dataset-1 and Dataset-2 with Node 1 and Node 2, respectively. Additionally, we have included the following bar graph for easier comparison.



All this information has been included in the new version of the manuscript.

Table 14 contains **repetitive** information and should be removed.

Response 17: Since one of the metrics used in the comparisons is the improvement relative to the raw values, we have kept this table (in the new version is number 10) but improved its explanation in the text to clarify these results as follows:

In Table 10, we show the improvement in % using the different ML models for the calibration process from the LCS raw readings of the module, highlighting the better performance of GB model compared to the other models. Notice that with this model, GB, the initial MAE from the raw readings was 67.59 $\mu g/m^3$ reducing it to 4.022 $\mu g/m^3$, that is an improvement of 94.05% as depicted in this table.

Table 15 would be more useful if combined with **table 1** instead of expecting the reader to remember the sensor specs from the very beginning. However, as the authors point out, this is comparing multiple different types of sensors that aren't inherently comparable. I understand that the authors are trying to show the usefulness of their calibration, but I don't think they need to directly compare with others for that message to come across. I recommend **removing tables 1 and 15**, especially because the inclusion of information on these other sensors in the earlier sections muddles the message of what the paper is ultimately trying to convey.

Response 18: As we answered in Response 5, Table 15 (in the new version is 11) is used for comparison and we have included some extra details for clarity. In this table, we compare our models for ozone calibration for low-cost sensors, against the related work with a similar approach, highlighting the location, platform and sensors used, R^2, mean relative error (MRE) with comments about the details of the models used and dataset duration

Notice that this table only considers some of the modules shown in Table 1, in particular RAMP, AirSensrEUR and ZPHS10B. Table 1 is an overview of different commercial sensor modules available, detailing only their sensors and price range, without considering whether all these modules have been used in related work.

This new table (Table 11) was already shown in Response 1, and its explanations have included in the new version of the manuscript as follows:

Finally, in Table 11, we compare our models for O_3 calibration for LCS, against the related work with a similar approach, highlighting the location, platform (and sensors used), R^2, MRE along with additional comments about the detail of the models used and dataset duration. First, we must stress that the starting point of the selected papers is slightly different compared to ours, since these studies have used more reliable and expensive LCS, approximately ten times more expensive that the ZPHS01B module. Moreover, since ML-based algorithms show the best results as discussed in Section 2, we have focused exclusively on them, evaluating up to four different models, whereas other studies have only considered one or two. Our model, in particular GB with 4 features (including "date" as metadata), as shown in Section 3.3, achieves a MRE of 7.21% (given by MAE 4.022 $\mu g/m^3$ with 90/10 dataset (Table 9) and the mean O_3 value of 55.72 $\mu g/m^3$ (Table 3)). Besides, not all of these works follow and discuss an structured EDA with FIA, FS and HPO. In particular, when compared to the first two works with slightly better results, in (Wang et al. (2024), we appreciate higher O_3 values, mean values higher than 70 $\mu g/m^3$, while in our case we have lower levels (55.72 $\mu g/m^3$)), as well as there is not a complete EDA. It is important to note that these sensors perform worse at low concentrations than at high ones due to their sensitivity limitations and the weakness of the signals generated, as well as interference from other pollutants. Finally, in (Cavaliere et al. (2023)), although the authors use a complete EDA, they only use two sensors (NO_2 and O_3) apart from Temp and RH, and the aging effect is considered a posteriori, while this information is included in our case by date in our models, which also detects other patterns derived from road traffic.

Line 321: This paragraph isn't indented, but all the others in this section are.

Response 19: It is the default AMT template.

Line **324-325**: Which **model** are these statistics from? The abstract suggests **GB**, but this should be clearly stated in the conclusions as well.

Line 350: Missing a **period** at the end of the sentence.

Response 20: These details were already included in the previous manuscript. After comparing the different models, we identified Gradient Boosting (GB) as the best model and have highlighted its performance metrics in both the abstract and conclusions.

Finally, we put this period.

Finally, thank you for your thoughtful review and comments which will enable us to improve this work. We appreciate the time and effort invested in your review.