Response to comments from Referee #1

Citation: https://doi.org/10.5194/egusphere-2025-2677-RC1

We would like to thank you for taking the time to review our manuscript and provide valuable feedback. Our responses and proposed revisions, which we believe enhance the quality of the paper, are presented below. The comments from Referee #1 are provided in black, our responses appear in brown, and the revised or newly added text in the manuscript is shown in *italics*.

First of all, I would like congratulate the authors for the work carried out and presented in this paper. After having read the full document, I'm not sure that the conclusion or the study really answer the question asked in the title. In fact, the author ask the question of the need of re-calibration of low-cost senors but they do not really answer it in the document as the present an interesting use of sensor for ambient air monitoring ("pairwise calibration strategy") based on a monthly exchange of LCS between a collocation site and a measurement site. This strategy, somehow interesting when looking at the sensors performances is much more time consuming than a classic network installation as, at the end, 2 LCS are always running adding the necessity of installation/removal every month. However, the interesting comparison of calibration results using several training length against both US-EPA and European standards brings a lot of valuable information.

In recent years, multiple recognized organizations such as the EPA and CEN have released state-of-the-art test protocols for air sensors. These are important and much-needed tools that help to communicate the possible end-use applications of low-cost sensors to the public after their evaluation. In this work, these test protocols provide guidance for evaluating and contextualizing the actual impact of different training lengths (extended training (ET)) compared to a shorter training period (single training (ST)) on sensor performance. However, the main research question is if and how recalibration must be designed to maximize performance of the sensors.

The conclusions section (Sect. 4) of this work offers the following statements related to the question asked in the title (Recalibration of low-cost air pollution sensors: Is it worth it?) of this study:

1. Our findings suggest that for quantitative studies, during periods characterized by elevated ground level ozone concentrations (ozone season), recalibration is advisable after each month of O_3 LCS operation. In particular, the machine learning techniques RF and XGB benefited from the increased amount of summer training data resulting from monthly recalibrations.

2. If extended training via monthly recalibration is feasible, RF and XGB calibration models appear to be the more sensible choice, as their quantitative performance

aligns particularly well with EPA guidelines for non-regulatory supplemental and informational monitoring devices targeting O_3.

3. A MLR calibration model using ET was the only calibration model that met all EPA-recommended performance metric goals for assessing the quantitative strength of $PM_{2.5}$ LCS data.

4. The REU values suggest that extended training of the employed calibration models enables the generation of a continuous LCS time series from two identical sensor model units, more consistently meeting a targeted DQO (e.g. indicative measurements). This approach also contributes to reduced measurement uncertainty, which becomes visually noticeable as a pollutant concentration increases. Again, extending the calibration model training period and therefore expanding the calibration space is especially advised for machine learning methods to reduce the LCS measurement uncertainty.

5. We conclude that achieving the highest possible quantitative validity for low-cost air sensors requires regular in-season recalibration using high-quality reference data. The response of the sensor units to changing environmental conditions at the station site, along with improved performance resulting from regular recalibration that aligns sensor output more closely with EPA and CEN recommendations, highlights how important regular sensor maintenance is to enhance their applicability.

We understand the reviewer's point that the current title may not fully reflect the content, which could be expected given its provocative nature. If the title seems too strong, we propose the following possible revisions:

Recalibration of low-cost air pollution sensors: Linking practices to state-of-the-art test protocols

Recalibration of low-cost air pollution sensors: Connecting calibration practices with modern test protocols

Recalibration of low-cost air pollution sensors for advanced performance

Furthermore, we agree that a pairwise calibration strategy is more time-consuming than a classic network installation, particularly when sensors are installed and removed monthly in a large-scale network. However, considering our observed sensor performances, we see value and the possibility in applying a pairwise calibration strategy in small networks, especially when LCS measurement systems are deployed at locations with high densities of vulnerable populations, such as retirement homes, schools, kindergartens, or outdoor workplaces. Implementing multiple smaller-scale

LCS networks by various groups with access to adequate infrastructure for sensor calibration (e.g. research institutions, state organizations), focused on at-risk population hotspots, could help LCS realize their potential and, in fact, gain recognition as long-term supplemental monitoring systems, integrated into official networks to serve the most vulnerable people of society.

I also made some minor comment along the document listed below:

- Line 153: length of this stabilization phase ?

We clarified the stabilization phase in the manuscript as follows:

Only after their stabilization phase the LCS output is eligible for measurements of their respective target pollutant (Gäbel et al., 2022). The stabilization phase observed in the LCS outputs was shorter than one day. The first 24 hours of all LCS data were thus removed and not considered for this study.

- Line 155: coma could be removed.

Done.

- Line 157: The 3 of O3 should be in subscript.

Done.

- Line 165: Are the daily means for LCS based on the hourly values or on the raw values ? The end of this paragraph suggest that the daily means has been calculated using hourly values. Did you check the impact on the data ?

We have clarified this in the manuscript as follows in lines 164-168:

Raw LCS measurements and reference measurements given by the AEMS were aggregated to hourly means for LCS calibration. Calibrated PM2.5 measurements were aggregated to daily means. Daily means were required for the performance evaluation of the low-cost particulate matter sensor SAG-SPS30 based on the technical specification developed by CEN (CEN/TS 17660-2:2024, 2024) and the test protocol developed by EPA (Duvall et al., 2021a).

- Line 183: This PM sensor sentence seems to me to be not in the right paragraph as the PM data has been discussed on the previous one.

We moved line 183 to the previous paragraph to line 168:

Calibrated PM2.5 measurements were aggregated to daily means. Daily means were required for the performance evaluation of the low-cost particulate matter sensor SAG-

SPS30 based on the technical specification developed by CEN (CEN/TS 17660-2:2024, 2024) and the test protocol developed by EPA (Duvall et al., 2021a). The SAG-SPS30 provides outputs in mass concentrations by default.

- Line184-189: This explanation could maybe be moved a after the first paragraph of 2.4 where the use of T and RH in the calibration models is explained. It was somehow confusing to me to read first that the data from the BME280 were not used to then see that they are finally used. Only on a second read I pay attention to the fact that the BME280 data were not used for the gas sensors.

The purpose of the paragraph is to emphasize that mass concentrations are required for sensor evaluation according to CEN/TS 17660-1:2021 and to specify which meteorological data we considered in order to calculate mass concentrations as accurately as possible. Therefore, we would prefer to keep these lines in the data treatment section, as the contents of the full paragraph are too closely interwoven.

To prevent confusion, we adjusted the lines 184-189:

We exclusively used low-cost meteorological data from the Bosch BME280 sensors as input for the calibration models (Sect. 2.4). To calculate mass concentrations from the output of the calibration models we did not rely on BME280 meteorological data, but used the weather station data, because the former are highly biased due to solar radiation. The bias stems from solar heating of the AELCM units, which could not be mitigated by the integrated fan, as it causes an exchange of air between the inside and outside, failing to reduce the heating effect. It is planned to upgrade the AELCM units with radiation shields in the future to reduce the effect of solar radiation on the low-cost meteorological measurements.

- Table 1: the first row is not the easiest to read, in particular for O3 and NO2 as there is not a clear separation between the T (end of O3) and VNO2 (beginning of NO2).

We improved the readability of the table:

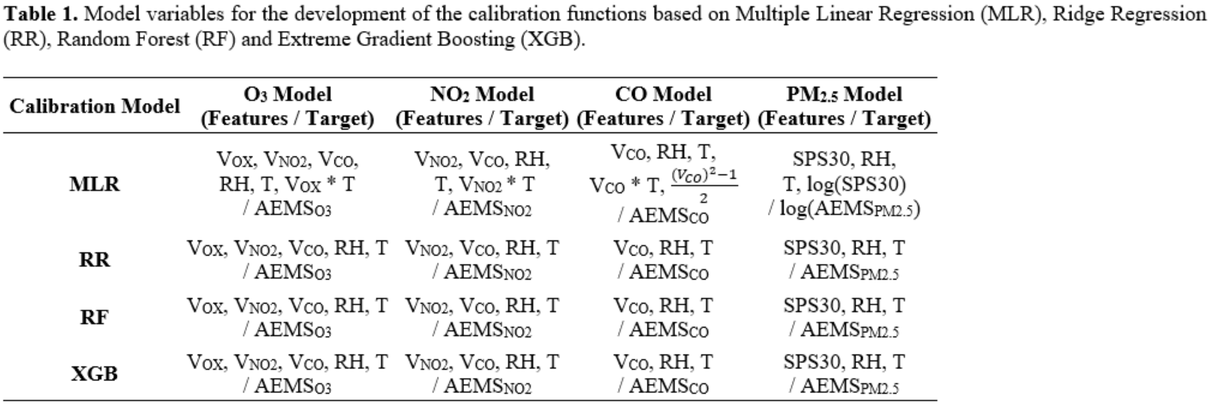

Table 1. Model variables for the development of the calibration functions based on Multiple Linear Regression (MLR), Ridge Regression (RR), Random Forest (RF) and Extreme Gradient Boosting (XGB).

Calibration Model	O_3 Model (Features / Target)	NO_2 Model (Features / Target)	CO Model (Features / Target)	$PM_{2.5}$ Model (Features / Target)
MLR	V_{OX}, V_{NO2}, V_{CO}, RH, T, V_{OX} * T / $AEMS_{O3}$	V_{NO2}, V_{CO}, RH, T, V_{NO2} * T / $AEMS_{NO2}$	V_{CO}, RH, T, V_{CO} * T, $\frac{(V_{CO})^2-1}{2}$ / $AEMS_{CO}$	SPS30, RH, T, log(SPS30) / log($AEMS_{PM2.5}$)
RR	V_{OX}, V_{NO2}, V_{CO}, RH, T / $AEMS_{O3}$	V_{NO2}, V_{CO}, RH, T / $AEMS_{NO2}$	V_{CO}, RH, T / $AEMS_{CO}$	SPS30, RH, T / $AEMS_{PM2.5}$
RF	V_{OX}, V_{NO2}, V_{CO}, RH, T / $AEMS_{O3}$	V_{NO2}, V_{CO}, RH, T / $AEMS_{NO2}$	V_{CO}, RH, T / $AEMS_{CO}$	SPS30, RH, T / $AEMS_{PM2.5}$
XGB	V_{OX}, V_{NO2}, V_{CO}, RH, T / $AEMS_{O3}$	V_{NO2}, V_{CO}, RH, T / $AEMS_{NO2}$	V_{CO}, RH, T / $AEMS_{CO}$	SPS30, RH, T / $AEMS_{PM2.5}$

- Line 218: what do you mean by merging the data by hour ? is it the mean calculation ?

We aligned the hourly reference station data with the hourly raw LCS data by matching timestamps. We think it is redundant to mention this, since time alignment is the standard procedure when comparing a reference method with a candidate method. Therefore, we removed "*and merging the data by hour*" in line 218.

- Line 395: you should mention in the previous paragraph 2.7 Performance metrics and target values that the measurement thus the evaluation has been carried out only for a urban background site whereas the CEN document ask for different testing site, for example a rural site for O3.

I've added this information in Section 2.7 at line 369:

It should also be noted that the LCS evaluation was performed only at a single urban background site (AEMS), whereas the technical specifications by CEN call for evaluations at different sites, for instance, testing NO_2 sensors at traffic and background sites.

- Figure 8, 9, 10 and 11: I would advice the authors to write the title of the different graphs on a clearer way, at a first look, it is not easy to see the difference between each plot.

We adjusted the title position and title size of each of the figures mentioned to enhance readability. The adjustments can be seen further down below:

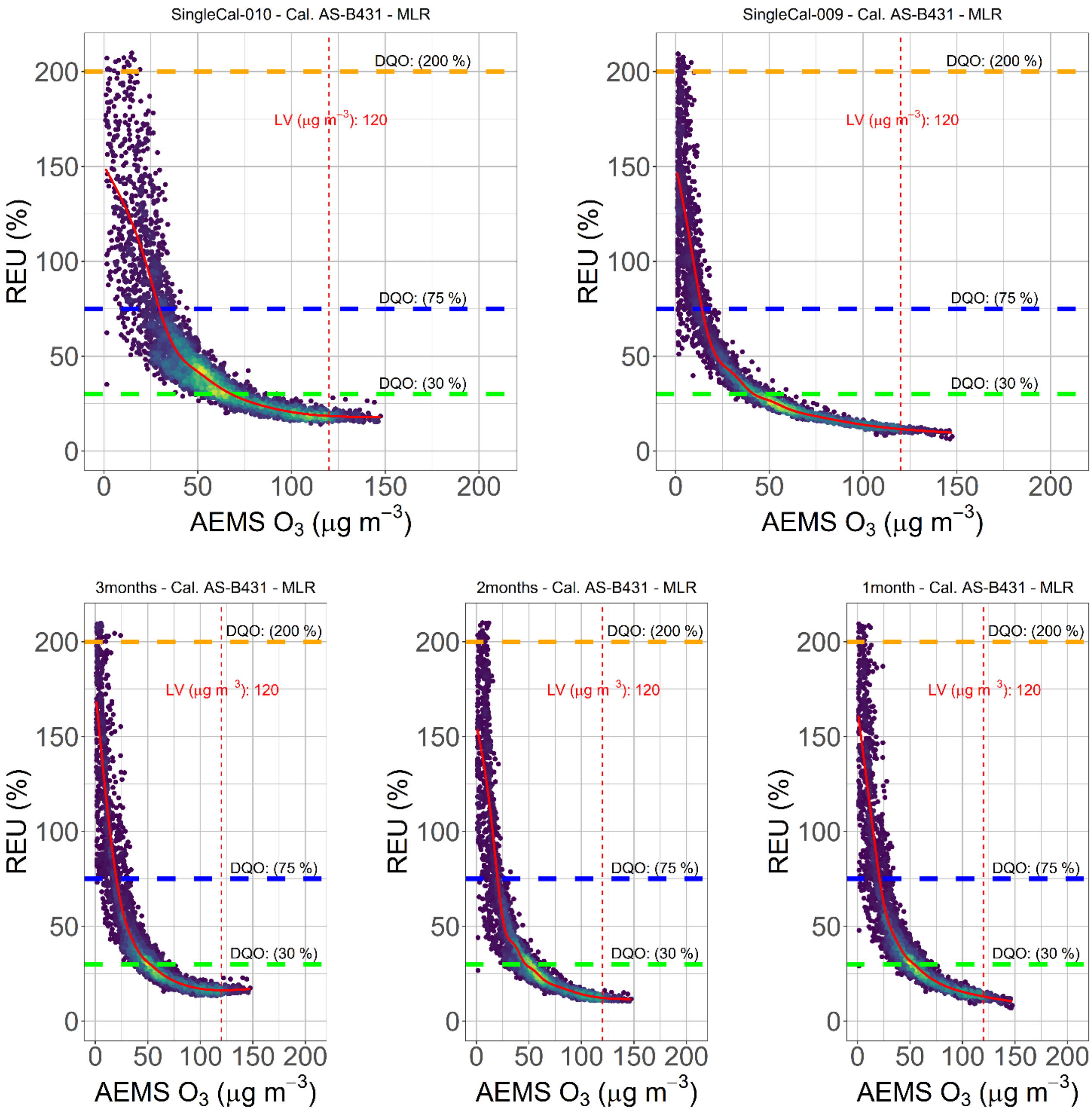

Figure 8. Calculated REU values for MLR calibrated O_3 LCS hourly data belonging to the test periods (TP1–TP7, 10 June 2022–11 January 2023) of AELCM009 and AELCM010. The calibration variants are single training (ST) (top row, left: AELCM010, right: AELCM009) and extended training (ET) (bottom row). The extended training is characterized by ET variants of 1, 2 and 3 months for each AELCM box. Horizontal dashed lines describe the data quality objectives (O_3 Class 1 DQO = 30 %, Class 2 DQO = 75 % and Class 3 DQO = 200 %). The vertical dashed line describes the limit value for O_3 (LV = 120 µg m^{-3}). The fitted smooth curve (red) is based on a generalized additive model (GAM). Data density is shown through colour, where darker colours express lower data density and brighter colours express higher data density.

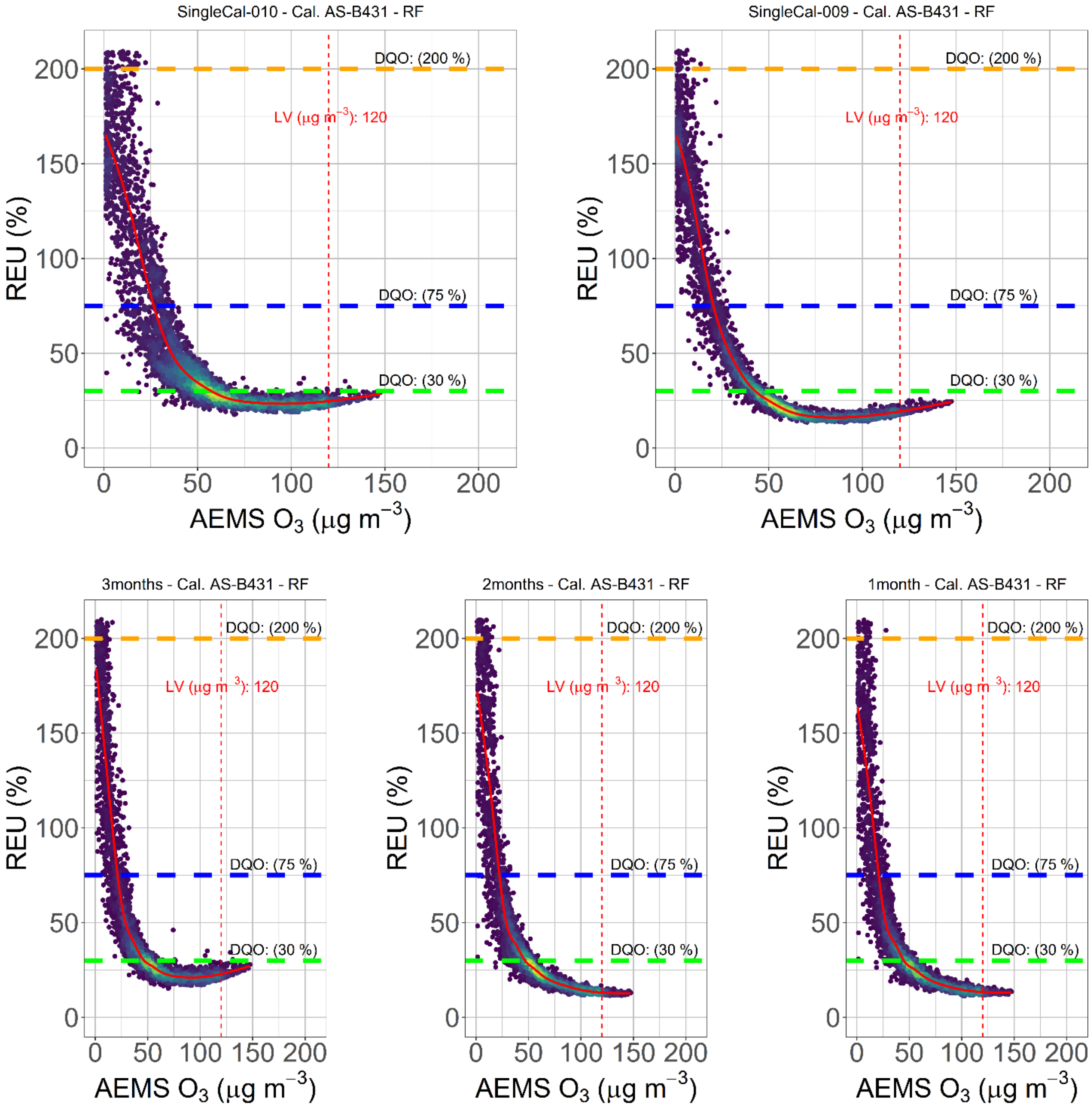

Figure 9. Calculated REU values for RF calibrated O_3 LCS hourly data belonging to the test periods (TP1–TP7, 10 June 2022–11 January 2023) of AELCM009 and AELCM010. The calibration variants are single training (ST) (top row, left: AELCM010, right: AELCM009) and extended training (ET) (bottom row). The extended training is characterized by ET variants of 1, 2 and 3 months for each AELCM box. Horizontal dashed lines describe the data quality objectives (O_3 Class 1 DQO = 30 %, Class 2 DQO = 75 % and Class 3 DQO = 200 %). The vertical dashed line describes the limit value for O_3 (LV = 120 µg m^{-3}). The fitted smooth curve (red) is based on a generalized additive model (GAM). Data density is shown through colour, where darker colours express lower data density and brighter colours express higher data density.

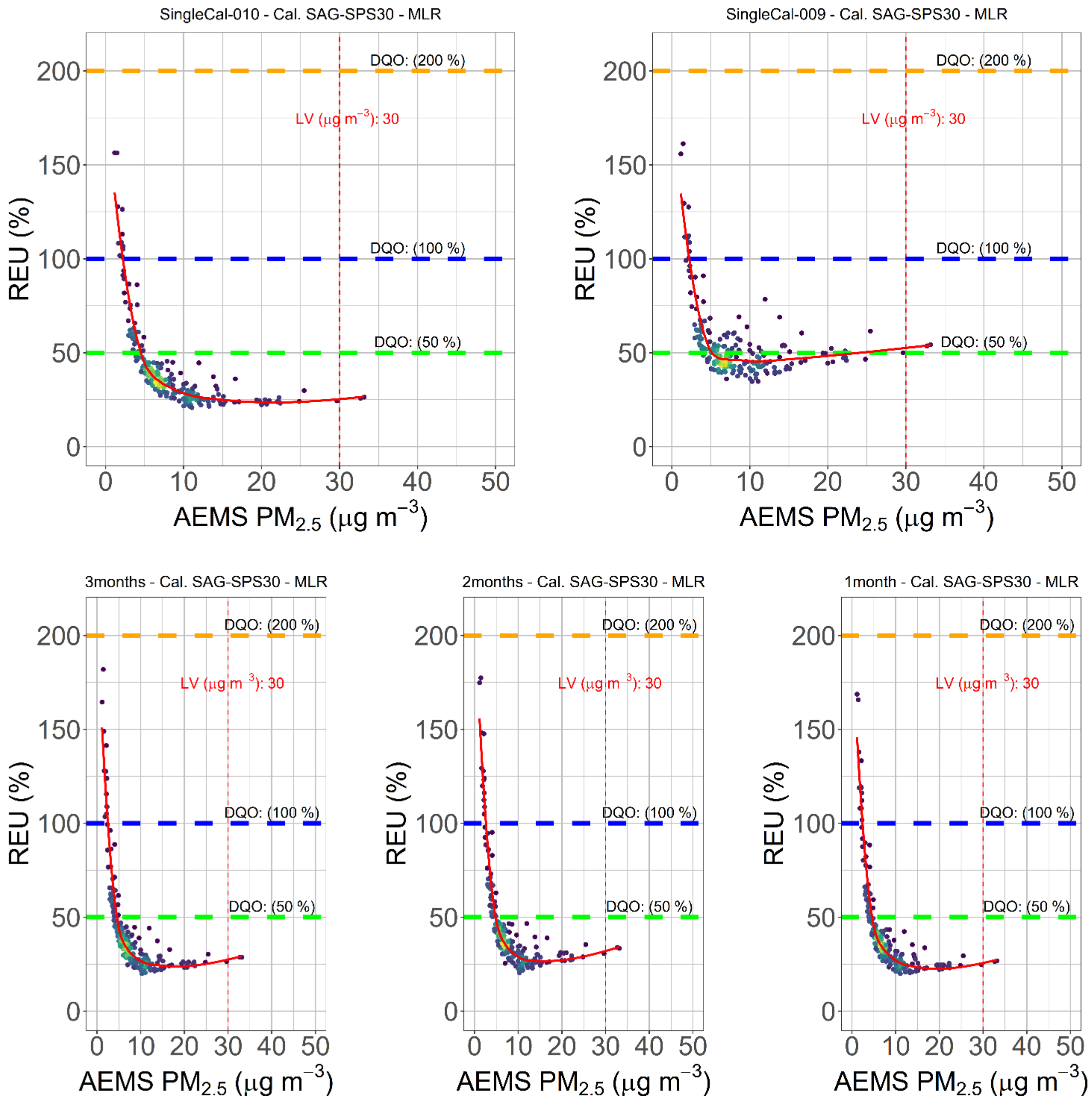

Figure 10. Calculated REU values for MLR calibrated $PM_{2.5}$ LCS daily data belonging to the test periods (TP1–TP7, 11 June 2022–6 January 2023) of AELCM009 and AELCM010. The calibration variants are single training (ST) (top row, left: AELCM010, right: AELCM009) and extended training (ET) (bottom row). The extended training is characterized by ET variants of 1, 2 and 3 months for each AELCM box. Horizontal dashed lines describe the data quality objectives ($PM_{2.5}$ Class 1 DQO = 50 %, Class 2 DQO = 100 % and Class 3 DQO = 200 %). The vertical dashed line describes the limit value for $PM_{2.5}$ (LV = 30 μg m^{-3}). The fitted smooth curve (red) is based on locally estimated scatterplot smoothing (LOESS). Data density is shown through colour, where darker colours express lower data density and brighter colours express higher data density.

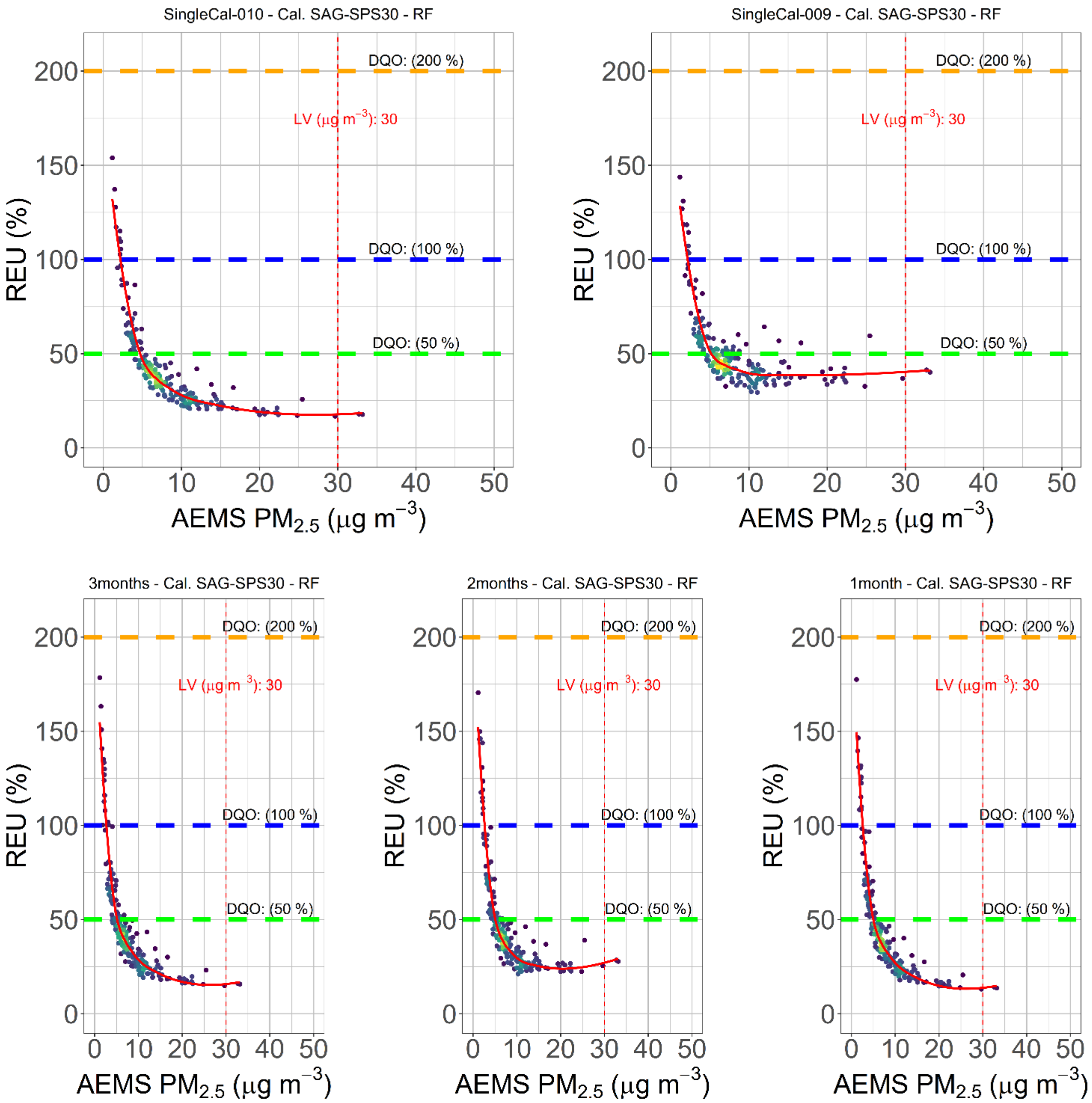

Figure 11. Calculated REU values for RF calibrated $PM_{2.5}$ LCS daily data belonging to the test periods (TP1–TP7, 11 June 2022–6 January 2023) of AELCM009 and AELCM010. The calibration variants are single training (ST) (top row, left: AELCM010, right: AELCM009) and extended training (ET) (bottom row). The extended training is characterized by ET variants of 1, 2 and 3 months for each AELCM box. Horizontal dashed lines describe the data quality objectives ($PM_{2.5}$ Class 1 DQO = 50 %, Class 2 DQO = 100 % and Class 3 DQO = 200 %). The vertical dashed line describes the limit value for $PM_{2.5}$ (LV = 30 µg m^{-3}). The fitted smooth curve (red) is based on locally estimated scatterplot smoothing (LOESS). Data density is shown through colour, where darker colours express lower data density and brighter colours express higher data density.