A comparative Study of Calibration Methods for Low-Cost Ozone Sensors in IoT Platforms

Pau Ferrer-Cid, Jose M. Barcelo-Ordinas, Jorge Garcia-Vidal, Anna Ripoll, Mar Viana

***Abstract*—This paper shows the result of the calibration process of an IoT platform for the measurement of tropospheric ozone (O_3). This platform, formed by sixty nodes, deployed in Italy, Spain and Austria, consisted of one hundred and forty metal-oxide O_3 sensors, twenty-five electro-chemical O_3 sensors, twenty-five electro-chemical NO_2 sensors and sixty temperature and relative humidity sensors. As ozone is a seasonal pollutant, which appears in summer in Europe, the biggest challenge is to calibrate the sensors in a short period of time. In the paper, we compare four calibration methods in the presence of a large data set for model training and we also study the impact of a limited training data set on the long-range predictions. We show that the difficulty in calibrating these sensor technologies in a real deployment is mainly due to the bias produced by the different environmental conditions found in the prediction with respect to those found in the data training phase.**

***Index Terms*—IoT platform, Sensor Calibration, Low-cost sensors, Uncontrolled environments, Quality of Information (QoI)**

I. Introduction

FOUR point two million deaths are produced every year as a result of exposure to ambient (outdoor) air pollution according to WHO[1] (World Health Organization). Moreover, around 91% of the world's population live in places where air quality levels exceed WHO limits. National and regional governmental organizations measure pollutants using highly accurate instruments. However, these equipments are costly to deploy and maintain, being its number low with respect to large density areas. Low-cost air pollution sensors mounted on nodes forming an Internet of things (IoT) platform can help to estimate and understand the pollution in areas with low number of accurate instruments.

One of the most discussed topics [1], [2] with low-cost sensor networks is the accuracy of the data they provide. In recent years, there has been greater interest in learning how low-cost sensors behave in terms of quality of information (QoI) metrics such as the root mean square error (RMSE), mean bias error (MBE), or the short-term or large-term capacity prediction of the sensors. Many of the low-cost sensors in IoT platforms are not calibrated by the manufacturers or if they are calibrated by them, the calibration has been done in laboratory chambers and not in the environmental conditions of the place where the nodes are deployed [3], [4]. In this case, the sensors of the IoT platform is calibrated during network deployment in an uncontrolled environment without laboratory instruments [5], [6].

For this reason, much research has focused on the interaction of environmental conditions such as temperature and relative humidity [3], [4], [7], [8], [9] or on the interactions of other pollutants [10], [11] with respect to one pollutant sensor. In addition, there is recently a greater interest in comparing and studying [11], [12], [13], [14], [15] how signal processing techniques behave for calibrating different air pollution low-cost sensors in IoT platforms. Many of these investigations focus on comparing what is the error obtained using several linear and non-linear machine learning algorithms. The authors choose some commercial sensors, take data during a large period of time and compare how the sensor behave in terms of QoI metrics. Most of the time the goal is to evaluate which commercial sensor performs best or whether a commercial sensor performs well.

H2020 CAPTOR[2] project (2016-2018) works on the assumption that the combination of citizen science, collaborative networks and environmental grassroots social activism helps to raise awareness and find solutions to air pollution problems. During the project, three testbeds forming an IoT platform, aimed at increasing public awareness of tropospheric ozone (O_3), have been deployed in Austria, Italy and Spain [16]. Sixty wireless nodes have been deployed with 140 metal-oxide O_3 sensors, 25 electro-chemical O_3 sensors, 25 electro-chemical NO_2 sensors and 60 temperature and relative humidity sensors.

One of the main challenges in a real IoT sensor deployment with air pollution sensors is that the amount of time for calibrating the sensors is of few weeks. This means that one of the most difficult challenges to solve is *how to predict the pollutant concentrations in the long-term given that the calibration has been carried out in a fixed and not very long period of time*. As a consequence, the IoT nodes have to be calibrated in an uncontrolled environment and knowing that the environmental conditions will change in time.

In this paper, we describe the process of calibrating metal-oxide and electro-chemical low-cost O_3 sensors in a real IoT

Pau Ferrer-Cid (pauferrercid12@gmail.com), Jose M. Barcelo-Ordinas (joseb@ac.upc.edu) and Jorge Garcia-Vidal (jorge@ac.upc.edu) are with the Universitat Politecnica de Catalunya, Barcelona, Spain.

Anna Ripoll (anna.ripoll@idaea.csic.es) and Mar Viana (mar.viana@idaea.csic.es) are with the Institute for Environmental Assessment and Water Research, Spanish National Research Council (IDAEA-CSIC), Barcelona, Spain

This work is supported by the National Spanish funding TIN2016-78473-C3-1-R, regional project 2017SGR-990 and by European H2020 CAPTOR project. The authors also thanks the collaboration of the staff at the Department of the Environment of the Generalitat de Catalunya for providing support for the deployment of the sensing nodes at reference stations and access to the reference data.

[1]https://www.who.int/airpollution/ambient/en/

[2]https://www.captor-project.eu/en/

network deployed during the H2020 CAPTOR project [16]. For that purpose we assume that (i) O_3 is seasonal, and therefore, has large peaks during summer in Europe, (ii) the calibration process is performed just before the summer and the objective is to learn how accurate are the predictions of O_3 concentration during the summer. For that reason, a set of linear (multiple linear regression) and non-linear (K-nearest neighbors, support-vector regression and random forest) algorithms are compared, both in the short and long-term predictions.

The outline of the paper is as follows: section II enumerates the related work. Section III explains the testbeds and data sets employed for the analysis. Section IV illustrates the calibration algorithms used for comparing sensor behavior. Section V describes the results showing how the sensors behave in the short and long-term and the impact of the environmental conditions. Finally, section VI concludes the paper.

II. Related work

There has recently been a large number of studies devoted to calibration in many fields related to low-cost sensor in IoT platforms including weather, air quality monitoring, target discovery, synchronization or localization [6]. Signal processing techniques have been applied to calibrate low-cost sensors in IoT. In general, temperature and relative humidity low-cost sensors follow linear patterns, and linear regression has been the main technique used for calibrating these sensors. Several authors [10], [17], [18] have shown that in order to calibrate air pollution sensors such as CO, NO_2, CH_4, O_3, CeO_2 or C_3H_8 it is needed an array of sensors. The idea of sensor array calibration consists in measuring all the cross-sensitivities to compensate for all interfering pollutants and environmental conditions [5], [6]. For example, to calibrate a NO_2 sensor, NO_2, O_3, temperature, and relative humidity are measured.

There are several calibration approaches to calibrate a sensor node [6]. The most typical is the approach in which uncalibrated sensor nodes are collocated (placed) a few meters away from the reference node [11], [12]. Other possible calibration approaches assume a distributed network of nodes. An example is to calibrate nodes using a multi-hop calibrated network, in which a node is calibrated using an already calibrated node [17].

All these works use data processing algorithms for calibrating the low-cost sensors. Whenever the sensor response has a linear behavior with respect to the reference data, multiple linear regression (MLR) [11], [12], [14], [19] is used for calibrating the sensors. Nevertheless, when the response is non-linear, models such as K-nearest neighbors (KNN) [19], [20], Gaussian processes [20] and more recently support-vector regression (SVR) [14], [15], [21], random forest (RF) [13], [14], [20] and artificial neural networks (ANN) [11], [15], [20] have been used to calibrate low-cost sensors. Most of these works are focused on studying and analyzing the quality of different commercial low-cost sensors and the performance of electro-chemical sensors. In general, the authors deploy a sensor collocated with reference instruments, take data for a large amount of time, e.g., several months, and compare one or several calibration models or compare sensors from different manufacturers with large data sets. This methodology is correct for assessing how good is a calibration model or how good is a sensor technology if a large amount of samples are available for calibrating the sensor. But, few of them deploy sensors with the target of a real IoT network deployment in which the objective is to calibrate the sensor in few weeks, thus with a short amount of samples, and assess how these sensors will behave in the long-term and how it will drift.

The drift in electro-chemical low-cost sensors has also been paid attention. This drift is a degradation mainly due to poisoning and aging of the sensor material. For example, Martinelli et al. [22] propose a modified version of an artificial immune system (AIS) algorithm that having some form of memory, is less affected by drift. Other authors [7], [12], [13], [17] propose recalibration as a way of fighting drifts in low-cost sensor networks. For example, Mijling et al. [7] and Barcelo-Ordinas et al. [12] propose a pre-post calibration approach, in which the sensors are calibrated in the pre-campaign followed by a second calibration period after the measurement campaign is finished with the aim of assessing and compensating the individual sensor drift in the IoT sensor nodes. Saukh et al. [17] mount nodes in a network of buses and re-calibrate the air pollution sensors each time that the buses opportunistically cross a reference monitoring station. This type of calibration is called opportunistic or periodic calibration depending on whether the recalibration is opportunistically or periodically scheduled. Wei et al. [8] also study the drift and the impact of environmental parameters such as temperature and relative humidity. In general, short and long-term predictions are quite sensitive to temperature and relative humidity [8], and these parameters have to be included in the calibration model.

In this work, we compare linear (MLR) and non-linear (KNN, RF and SVR) models in O_3 sensors with metal-oxide and electro-chemical technologies. We compared the models for calibration in the presence of a large training data set, following the literature, but added (i) the analysis of the size of the training data set, and (ii) what happens when the training data set is small and a long-term prediction is made.

III. Data set and testbed

During H2020 CAPTOR project, three network testbeds in Spain, Italy and Austria were deployed during two summers in 2017 and 2018 [16]. Since the main objective of the project was to raise awareness on O_3, and this pollutant is seasonal in Europe, from mid May to mid September, the nodes were calibrated during part of May and June, deployed in volunteer houses in large areas from July to mid September and recovered for post-calibration from mid September to October. However, several nodes were permanently deployed in a reference station during the entire measurement campaign. The objective was to carry out calibration studies in the reference stations where the nodes of the volunteers had been deployed. Two kind of sensor technologies were deployed: SGX Sensortech MICS 2614 metal-oxide O_3 sensors in nodes called *Captors* and Alphasense O3B4 electro-chemical O_3 sensor in nodes called *Raptors*.

For calibrating metal-oxide O_3 sensors, it is needed to measure O_3, temperature and relative humidity. Captor nodes

TABLE I: Sensor Deployment information.

Node Name	Sensor Labels	Sensor Type	Calibration Place	Period	# of Samples
Captor C17013	s1,s2,s3,s4	MICS 2614	Manlleu (Spain)	08/05/2017-04/10/2017	6745
Captor C17016	s1,s2,s3,s4	MICS 2614	Vic (Spain)	26/05/2017-05/10/2017	6149
Captor C17017	s1,s2,s3,s4	MICS 2614	Tona (Spain)	08/05/2017-05/10/2017	6944
Raptor R69-17	s1	OX-B431	MonteCucco (Italy)	06/07/2017-11/10/2017	1797
Raptor R308-17	s1	OX-B431	Weiz Bahnhof (Austria)	07/06/2017-27/09/2017	1439
Raptor R69-18	s1	OX-B431	MonteCucco (Italy)	20/06/2018-26/09/2018	2295
Raptor R202-18	s1	OX-B431	Colli Euganei (Italy)	18/06/2018-30/09/2018	2254
Raptor R212-18	s1	OX-B431	Osio Sotto (Italy)	26/06/2018-25/09/2018	2148

have been built by Universitat Politecnica de Catalunya (UPC) in Spain. Each Captor node uses Arduino technology with a sensor shield board that attaches four SGX Sensortech MICS 2614 metal-oxide O_3 sensors, a temperature (T) sensor and a relative humidity (RH) sensor and it is powered by an external power supply. Metal-oxide Sensortech sensors measure O_3 using a voltage divider circuit that has a load resistor and a variable resistor. Whenever the O_3 concentration changes, the variable resistor changes. The resistor value representing the O_3 sample is obtained by measuring the voltage V_L in the load resistor after quantizing the signal with an A/D converter and converting this voltage to the raw measurement:

$$s_{O_3} = R_L(1 - \frac{V_{cc}}{V_L}). \quad (1)$$

Where s_{O_3} is the raw O_3 measurement in kiloohm, R_L is the load resistor, V_{cc} is the input voltage and V_L is the voltage measured by the A/D converter. Reference monitoring stations show pollutants every half-hour or hour as the concentrations of these pollutants change slowly over time. Thus, a Captor node sends one measurement to a database repository every half hour. In order to have a representative value, each measurement is the average of 100 consecutive samples in which the 10 highest and the 10 lowest are removed to avoid outliers. Each measurement is a tuple with *RawData*$_{Captor}$={Timestamp, $s1_{O_3}$, $s2_{O_3}$, $s3_{O_3}$, $s4_{O_3}$, $s5_T$, $s6_{RH}$} where si_{O_3} (i=1,2,3,4) is a O_3 sensor raw measure, $s5_T$ is a temperature sensor measure and $s6_{RH}$ is a relative humidity sensor measure. This tuple is sent via 3G or Wifi to an IoT platform repository using a REST Web service.

For calibrating electro-chemical O_3 sensors, it is needed to measure NO_2, O_3, temperature (T) and relative humidity (RH). Raptor nodes have been built by Universite Clermont Auvergne (UCA) in France. Each Raptor node uses Raspberry Pi technology with one Alphasense OX-B431 electro-chemical O_3 sensor, one Alphasense NO2-B43F electro-chemical NO_2 sensor, a temperature sensor and a relative humidity sensor. The Raptor platform is composed by two boxes: an outdoor box is powered by a 9V 4000mAh battery for a lifetime of 3 months, and connected using a IEEE802.15.4 (ZigBee) wireless access medium to a indoor box that acts as local server, powered by an external power supply and connected to Internet using Wifi or 3G. The measure raw data is obtained by averaging data taken every minute during half hour.

Alphasense OX-B431 and NO2-B43F electro-chemical sensors use the Alphasense support circuits Individual Sensor Board (ISB) that outputs two signals for each sensor. The signals are called the working electrode (WE) and the auxiliary electrode (AE) used to compensate for zero current and both give values in the range of milivolts. The NO_2 sensor measures only NO_2 and the difference between the two sensors, after passing through an A/D converter, gives the O_3 concentration, so the raw O_3 measurement in milivolts is:

$$\begin{aligned} s_{O_3} &= (WE_{O_3} - AE_{O_3}) - (WE_{NO_2} - AE_{NO_2}), \\ s_{NO_2} &= WE_{NO_2} - AE_{NO_2}. \end{aligned} \quad (2)$$

Each measurement is a tuple with *RawData*$_{Raptor}$= {Timestamp, $s1_{O_3}$, $s2_{NO_2}$, $s3_T$, $s4_{RH}$}, where $s1_{O_3}$ is a O_3 sensor raw measure, $s2_{NO_2}$ is a NO_2 sensor raw measure, $s3_T$ is a temperature sensor measure and $s4_{RH}$ is a relative humidity sensor measure. This tuple is sent via 3G or Wifi to an IoT platform repository using a REST Web service.

Table I shows the nodes used in this study. We only show nodes that were during a large amount of time collocated in reference stations. The table shows the place, interval of time, number of sensors and technology of the O_3 sensors employed.

IV. Calibration Algorithms

Usually the MICS 2614 metal-oxide O_3 sensor response is linear with respect to O_3 concentration, temperature and relative humidity. Alphasense electro-chemical O_3 sensor is linear with respect to O_3 and NO_2 concentrations, temperature and relative humidity. Thus, the most used method for calibrating metal-oxide and electro-chemical O_3 sensor is a multivariate linear regression (MLR).

However, when a sensor is calibrated, sometimes nonlinearities appear due to the nature of low-cost sensing techniques or sometimes due to impurity and aging of the sensor material. To overcome these problems, nonlinear calibration methods such as K-nearest neighbors (KNN), random forest (RF) and support vector regression (SVR) can be introduced to fit the nonlinear responses of the sensor. These methods differ in the quality of information obtained and in the complexity in the execution of the method when calibrating the sensor.

For the nonlinear methods some hyperparameters are needed. Hyperparameters are configuration variables whose value are set before the training phase is executed. Hyperparameters are found using a grid search and they have a large impact in the training and testing time execution. In order to find the best set of hyperparameters per each algorithm a 10-fold cross-validation strategy is used.

A. Multivariate linear regression (MLR)

Let us consider an array of M sensors. A MLR model in multi-array calibration sensor assumes M predictors, one for each sensor of the array (O_3, T and RH), taking the form of:

$$\hat{y}(x_i) \sim f(\beta, x_i) = \beta_0 + \beta^T x_i + \epsilon_i, \qquad i = 1, .., N. \quad (3)$$

Where $x_i \in \mathbb{R}^M$ is a vector with the sensor data, β_0 is the offset, $\beta \in \mathbb{R}^M$ are the gains and ϵ_i is random noise following a Gaussian distribution with zero mean and variance σ^2.

B. K-nearest neighbors (KNN)

The K-nearest neighbors method falls in the category of memory-based methods where the training data is the model itself where the data space forms a cube of dimension M. Then, in order to obtain a new prediction for a point x we find the k closest points in the cloud and average their values:

$$\hat{y}(x) = \frac{1}{k} \sum_{x_i \in N(x)} y(x_i) \tag{4}$$

Where $N(x)$ is the set of point belonging to the neighborhood of x. Defining a neighborhood of closest points implies a distance metric to find them. In our case the distance metric is the *Minkowski* distance. Two hyperparameters are present in this model: the number of neighbors k in the KNN model and the Minkowski distance power p.

C. Random forest (RF)

Nowadays, random forests are becoming widely used in the environmental sciences field [13], [14], [20]. Random forest is an ensemble learner, it mainly constructs a forest of uncorrelated decision trees (week learners). The main benefit of using this ensemble methods is reducing the variance of the response obtaining a better model than just using one simple decision tree. The random forest algorithm proceeds as follows: it grows T trees using T bootstrap samples of the training data, at each node of the decision trees $F \leq M$ features are randomly sampled and taken into account for the split, the depth D of the decision trees can be limited to avoid over-fitting. Finally, the output of the learner is the trees' outputs average.

$$\hat{y}(x) = \frac{1}{T} \sum_{i}^{T} tree_i(x) \tag{5}$$

During the building procedure three hyperparameters can be found to be selected via cross-validation: the number of trees T, the number of features F and the maximum tree depth D.

D. Support vector regression (SVR)

Support vector regression has also been proposed for calibrating low-cost pollution sensors [14], [15], [21]. SVR [23] is a kernel method that is the analogous of support vector machines (SVM), but using continuous values instead of classifying as SVM. It makes use of the "kernel trick" where the data is implicitly mapped to a higher dimension in order to find a better regression curve but doing all computations in input space via a kernel function $k(x, x')$. The points that are far away from the correct regression plane will be the ones important for the correct model building. This is achieved via the ϵ-insensitive error loss, where only the points with error greater than ϵ are considered. The resulting SVR function is the following:

$$\hat{y}(x) = \sum_{i=1}^{N} (\hat{\alpha_i^*} - \hat{\alpha_i}) K(x, x_i) + b \tag{6}$$

The values for the parameters $\hat{\alpha_i^*}, \hat{\alpha_i}$ are found by solving a quadratic programming problem. The objective function to solve is obtained with the dual formulation of the problem, minimizing a loss function [23]. We have chosen to work with the radial basis function (*RBF*) kernel. The *RBF* kernel is proven to have an implicit map of infinite dimension. Finally, the hyperparameters optimized via cross-validation are the variance of the *RBF* kernel, the ϵ in the loss function and a penalization term C.

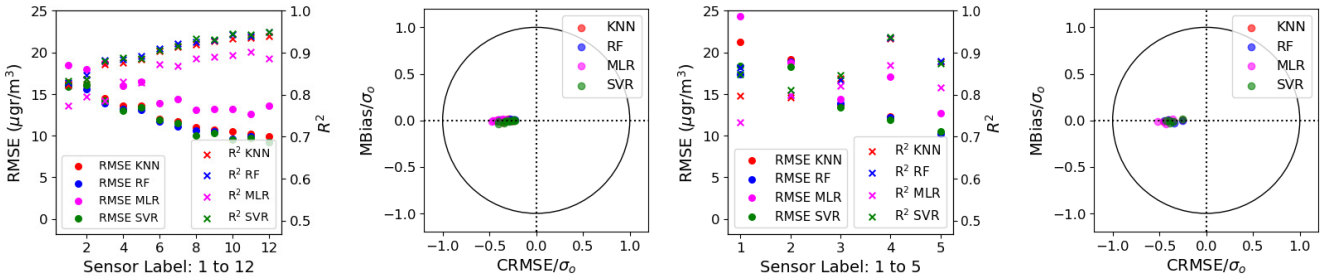
E. Assessment metrics

In order to calculate the calibration coefficients, the data set is split in two parts: a training set of size N_{tr} for calculating calibration parameters and a test set of size N_{ts} for assessing the calibration models. For comparing the different models we use the *mean bias error (MBE)*, the *root mean-squared error (RMSE)*, the *coefficient of determination (R^2)* and the *target diagram*.

The MBE and the RMSE consider the magnitude of the error in the prediction of a model. A value of R^2 close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. The target diagram [24], [25] visualises different aspects of model performance in one single plot, specifically, the MBE, the standard deviation, the RMSE, the centred root mean square error (CRMSE) and the correlation coefficient R. Let us define the σ_y and σ_x as the standard deviation of the reference data and the measured data respectively. The target diagram [25] is a circle of radius one. The x-axis represents the CRMSE normalized by σ_y. The y-axis represents the mean bias also normalized by σ_y. It can be proven that for a value within the circle unit, the RMSE normalized by σ_y is the magnitude between the origin and the value and it is referred as the *target indicator*. By definition, the CRMSE always is positive, however, target indicator points can be split into those that have $\sigma_y < \sigma_x$ (positive axis) and those that have $\sigma_y > \sigma_x$ (negative axis). Moreover, those ones that are out of the circle unit have a *model efficiency score (MEF)* negative. The MEF is defined as:

$$MEF = 1 - (\frac{\text{RMSE}}{\sigma_y})^2 \tag{7}$$

A MEF value near one [24] means a close match between reference data and model predictions. A value of zero indicates that the model predicts individual measurements no better than the average of the reference data. Values less than zero mean that the reference data average would be a better predictor than the model results. Thus, in the target diagram, for negative MEF a point is outside the circle unit while for positive MEF a point is inside the unit circle, being a point in the origin of the target diagram a perfect match between reference data and model predictions.



(a) Captor nodes (RMSE&R^2). (b) Captor nodes (target diagram). (c) Raptor nodes (RMSE&R^2). (d) Raptor nodes (target diagram).

Fig. 1: Test RMSE, R^2 and target diagram for Captor nodes C17013, C17016 and C17017 (12 metal-oxide sensors) and Raptor nodes R69-17, R308-17, R69-18, R202-18 and R212-18 (5 electro-chemical sensors).

V. Results

Our objective is to evaluate the calibration of low cost sensors in a real measurement IoT campaign with three testbeds in Spain, Italy and Austria. To do this, we first evaluate the performance of the sensors based on their technology: metal-oxide or electro-chemical, and the calibration model: MLR, KNN, RF, and SVR[3]. Then, we evaluate the capacity of the different models to evaluate O_3 concentrations in the long-term.

A. Linear versus non-Linear calibration methods

In this section, we compare the performance of a linear method (MLR) against non-linear methods (KNN, RF and SVR) in a large data set. We also compare how these methods behave in two technologies such as metal-oxide SGX Sensortech MICS 2614 O_3 sensors and electro-chemical Alphasense OX-B431 O_3 sensors. The data set is first shuffled and secondly split in 75% for the training set and 25% for the testing set. Figures 1.(a) and 1.(c) show the test RMSE and R^2 obtained for Captor nodes C17013, C17016 and C17017 (12 metal-oxide sensors) and Raptor nodes R69-17, R308-17, R69-18, R202-18 and R212-18 (5 electro-chemical sensors). The RMSE are sorted in decreasing order. Figures 1.(c) and 1.(d) show the target diagram for Captor and Raptor nodes. It can be observed several aspects:

i) identical sensors behave with large variability given the same calibration method. For example, for MICS 2614, the RMSE range for MLR is between 12 and 20 μgr/m^3 with a R^2 that ranges between 0.91 and 0.76. The same behavior can be observed with the non-linear methods and with the electro-chemical sensors,

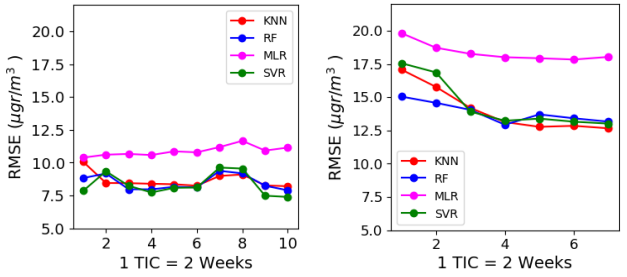
ii) low RMSE values are obtained with high R^2 values, indicating that a large proportion of the variability in the response has been explained by the model. Even in those cases in which RMSE increase, the R^2 values are higher than 0.7,

iii) non-linear models behave better than linear models in both technologies. The behavior between non-linear models is similar, with the SVR being better than RF and finally KNN. However, there are no major differences between the three non-linear models,

iv) the target diagram shows that captors and raptors do not have biases when there is a large number of samples in the training set. In this case, each point of the target diagram represents a sensor. Different colors indicate different calibration models. It can be observed, that the RMSE is mostly due to the variance, being greater the variance and therefore the RMSE, when using an MLR than when using a non-linear model (SVR, RF or KNN),

v) the target diagram also shows that there are practically no biases when the data set is very large as there are samples that represent a wide variety of information, i.e., environmental conditions.

[3]The raw data and the calibrated data can be found at doi:10.5281/zenodo.3233516. The code to obtain the calibrated data can be found at http://sans.ac.upc.edu/?q=node/231.

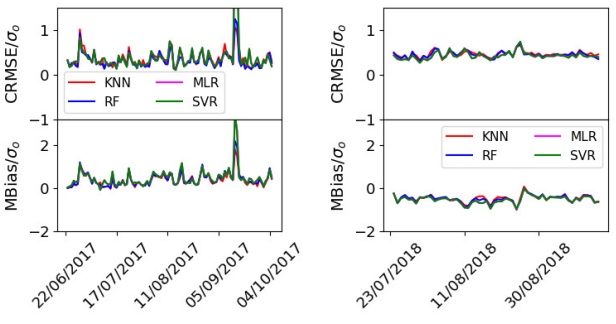


(a) Captor C17016, sensor s4 (b) Raptor R69-18

Fig. 2: RMSE vs training size in Captor C17016 and Raptor R69-18.

B. Training Set Size

In a real IoT deployment, like the one done in the H2020 CAPTOR project [16], the time to calibrate the sensors is limited to a few weeks. Once the sensors are calibrated, they should be able to predict the O_3 values for as long as possible. We are interested in learning what is the impact of the training set size in the RMSE. Figure 2 illustrates the RMSE as a function of the training set size. We consider one-week sample size training sets. The size of the training set is increased at one-week intervals. The size of the test data set remains fixed for all training sizes. As the size of the Captor nodes dataset is larger, we increase the size of the training set up to ten



(a) Captor C17016, sensor s4. (b) Raptor R212-18.

Fig. 3: Normalized bias and normalized CRMSE in Captor C17016 and Raptor R212-18.

weeks, with seven weeks at the end for the size of the test set. For Raptor nodes, on the other hand, as their dataset size is smaller, the size of the training set is increased to seven weeks, and the size of the test set is also fixed to seven weeks. It can be seen that a size between three and four weeks is enough to calibrate both Captor and Raptor nodes. This behavior has been observed in all metal-oxide and electro-chemical sensors. In some cases such as the MLR calibration, two weeks is enough for the RMSE to converge. However, in the non-linear cases the time interval for converging has ranged from five to seven weeks. In general, non-linear models need more samples for the RMSE to converge.

C. Long Term Prediction

In this section, we study how the different models behave when the training set is limited because the nodes have to be deployed in a real network and these nodes have to give data as accurately as possible over a long period of time. We, thus, set the training set to four weeks and observe the quality of the calibration day by day, that is, the test set has a size of one day but slides over a period of two months. Moreover, we consider the case in which the node is brought for recalibration. In this case, we consider two more instances: (i) the node is relocated to the reference station for one week, and the data is added to the training set (we call this case *augmented*), (ii) the node is relocated to the reference station for four weeks (we call this case *re-calibration*).

Figure 3 depicts the variation of the bias and CRMSE over time for the Captor 17017 node. Figures 4 and 5 show the target diagram for Captor and Raptor nodes C17016 and R212-18. Figures 4.(a),(d),(g),(j) and 5.(a),(d),(g),(j) is the general case, without re-calibration. Figures 4.(b),(e),(h),(k) and 5.(b),(e),(h),(k) plot the augmented training set. Finally, Figures 4.(c),(f),(i),(l) and 5.(c),(f),(i),(l) draw the re-calibration instance. For each instance, we consider the four calibration models, MLR, KNN, RF and SVR. Each point of the target diagram now represents a day. Several observations can be made made of these figures:

i) the fact that some points are in the right plane or in the left plane is due to the calibration model sometimes overestimates the variance and others underestimates it,
ii) having a large sample size in the training set implied that there were almost no biases, Figure 1.(b) and (c), however, when long-term predictions are made, biases appear, Figures 4.(a),(d),(g),(j) and 5.(a),(d),(g),(j). This bias is very variable, and depends on environmental conditions. Metal-oxide technologies have more bias than electro-chemical, but in both cases appear,
iii) the four calibration methods present bias. In the case of MLR, moreover, there is greater variance than in non-linear methods, which explains a higher RMSE,
iv) red dots have similar environmental conditions to blue dots (spring and autumn). Since the training set was taken in spring, these dots have fewer biases than those taken in summer,
v) increase (augmented instance) the size of the training dataset, Figures 4.(b),(e),(h),(k) and 5.(b),(e),(h),(k), slightly decrease the bias, having increased in a few weeks this dataset, the improvement is not much noticeable,
vi) recalibration, on the other hand, Figures 4.(c),(f),(i),(l) and 5.(c),(f),(i),(l), does improve bias and variance. In this situation, non-linear models behave better than linear models for O_3 metal-oxide technology. However, linear and non-linear models behave similar for O_3 electro-chemical technology.

Summarizing, recalibration improves bias at the cost of extracting the node from the deployment to put it back in the reference station.

VI. Conclusions

In this paper we have studied how to calibrate O_3 sensors with metal-oxide and electro-chemical technologies in a real deployment in Italy, Spain and Austria. Sixty wireless nodes were deployed with 140 metal-oxide O_3 sensors, 25 electro-chemical O_3 sensors, 25 electro-chemical NO_2 sensors and 60 temperature and relative humidity sensors. Four calibration methods have been compared (MLR, KNN, RF, SVR). In the case of having a large data set, several months, non-linear methods, and above all the SVR gives the best results in terms of RMSE. Also RMSE is mostly due to variance, with very little bias. This is because there are samples in all environmental conditions.

In general, all methods take about three to four weeks to calibrate O_3. However, when you have a few weeks of data, a normal situation in a real deployment, the long-term prediction presents bias. This is because the environmental conditions in which the training set was taken are different from those they present when predicting O_3 concentrations. Increasing the training size or re-calibrating improves the bias, but has the cost of having to extract the nodes from the deployment to relocate them to the reference stations. We think that other solutions, left as a future research, as the fusion of data, can improve the long-term predictions.

References

[1] E. G. Snyder, T. H. Watkins, P. A. Solomon, E. D. Thoma, R. W. Williams, G. S. Hagler, D. Shelow, D. A. Hindin, V. J. Kilaru, and P. W. Preuss, "The changing paradigm of air pollution monitoring," *Environmental science & technology*, vol. 47, no. 20, p. 11369, 2013.

[2] D. E. Williams, G. S. Henshaw, M. Bart, G. Laing, J. Wagner, S. Naisbitt, and J. A. Salmond, "Validation of low-cost ozone measurement instruments suitable for use in an air-quality monitoring network," *Measurement Science and Technology*, vol. 24, no. 6, p. 065803, 2013.

[3] K. Yamamoto, T. Togami, N. Yamaguchi, and S. Ninomiya, "Machine learning-based calibration of low-cost air temperature sensors using environmental data," *Sensors*, vol. 17, no. 6, p. 1290, 2017.

[4] N. Castell, F. R. Dauge, P. Schneider, M. Vogt, U. Lerner, B. Fishbain, D. Broday, and A. Bartonova, "Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates?" *Environment international*, vol. 99, pp. 293–302, 2017.

[5] B. Maag, Z. Zhou, and L. Thiele, "A survey on sensor calibration in air pollution monitoring deployments," *IEEE Internet of Things Journal*, vol. 5, no. 6, pp. 4857–4870, Dec 2018.

[6] J. M. Barcelo-Ordinas, M. Doudou, J. Garcia-Vidal, and N. Badache, "Self-calibration methods for uncontrolled environments in sensor networks: A reference survey," *Ad Hoc Networks*, vol. 88, p. 142, 2019.

[7] B. Mijling, Q. Jiang, D. de Jonge, and S. Bocconi, "Field calibration of electrochemical no$_2$ sensors in a citizen science context," *Atmospheric Measurement Techniques*, vol. 11, no. 3, pp. 1297–1312, 2018.

[8] P. Wei, Z. Ning, S. Ye, L. Sun, F. Yang, K. C. Wong, D. Westerdahl, and P. K. Louie, "Impact analysis of temperature and humidity conditions on electrochemical sensor response in ambient air quality monitoring," *Sensors*, vol. 18, no. 2, p. 59, 2018.

[9] J. M. Barcelo-Ordinas, P. Ferrer-Cid, J. Garcia-Vidal, A. Ripoll, and M. Viana, "Distributed multi-scale calibration of low-cost ozone sensors in wireless sensor networks," *Sensors*, vol. 19, no. 11, 2019.

[10] M. Mueller, J. Meyer, and C. Hueglin, "Design of an ozone and nitrogen dioxide sensor unit and its long-term operation within a sensor network in the city of zurich," *Atmospheric Measurement Techniques*, vol. 10, no. 10, pp. 3783–3799, 2017.

[11] L. Spinelle, M. Gerboles, M. G. Villani, M. Aleixandre, and F. Bonavitacola, "Field calibration of a cluster of low-cost commercially available sensors for air quality monitoring. part b: NO, CO and CO2," *Sensors and Actuators B: Chemical*, vol. 238, pp. 706–715, 2017.

[12] J. M. Barcelo-Ordinas, J. Garcia-Vidal, M. Doudou, S. Rodrigo-Muñoz, and A. Cerezo-Llavero, "Calibrating low-cost air quality sensors using multiple arrays of sensors," in *Wireless Communications and Networking Conference (WCNC)*. IEEE, 2018, pp. 1–6.

[13] N. Zimmerman, A. A. Presto, S. P. Kumar, J. Gu, A. Hauryliuk, E. S. Robinson, A. l. L. Robinson, and R. Subramanian, "A machine learning calibration model using random forests to improve sensor performance for lower-cost air quality monitoring." *Atmospheric Measurement Techniques*, vol. 11, no. 1, 2018.

[14] A. Bigi, M. Mueller, S. K. Grange, G. Ghermandi, and C. Hueglin, "Performance of no, no$_2$ low cost sensors and three calibration approaches within a real world application," *Atmospheric Measurement Techniques*, vol. 11, no. 6, pp. 3717–3735, 2018.

[15] S. De Vito, E. Esposito, M. Salvato, O. Popoola, F. Formisano, R. Jones, and G. Di Francia, "Calibrating chemical multisensory devices for real world applications: An in-depth comparison of quantitative machine learning approaches," *Sensors and Actuators B: Chemical*, vol. 255, pp. 1191–1210, 2018.

[16] A. Ripoll, M. Viana, M. Padrosa, X. Querol, A. Minutolo, K. M. Hou, J. M. Barcelo-Ordinas, and J. García-Vidal, "Testing the performance of sensors for ozone pollution monitoring in a citizen science approach," *Science of the Total Environment*, vol. 651, pp. 1166–1179, 2019.

[17] O. Saukh, D. Hasenfratz, and L. Thiele, "Reducing multi-hop calibration errors in large-scale mobile sensor networks," in *Proceedings of the 14th International Conference on Information Processing in Sensor Networks*, ser. IPSN '15. New York, NY, USA: ACM, 2015, pp. 274–285.

[18] Y. Liu, K. Zhou, and Y. Lei, "Using bayesian inference framework towards identifying gas species and concentration from high temperature resistive sensor array data," *Journal of Sensors*, vol. 2015, 2015.

[19] D. Hagan, G. Isaacman-VanWertz, J. Franklin, L. Wallace, B. Kocar, C. Heald, and J. Kroll, "Calibration and assessment of electrochemical air quality sensors by colocation with regulatory-grade instruments," *Atmospheric Measurement Techniques*, vol. 11, no. 1, pp. 315–328, 2018.

[20] C. Malings, R. Tanzer, A. Hauryliuk, S. P. N. Kumar, N. Zimmerman, L. B. Kara, A. A. Presto, and R. Subramanian, "Development of a general calibration model and long-term performance evaluation of low-cost sensors for air pollutant gas monitoring," *Atmospheric Measurement Techniques*, vol. 12, no. 2, pp. 903–920, 2019.

[21] R. Rossini, E. Ferrera, D. Conzon, and C. Pastrone, "Wsns self-calibration approach for smart city applications leveraging incremental machine learning techniques," in *2016 8th IFIP International Conference on New Technologies, Mobility and Security (NTMS)*, Nov 2016.

[22] E. Martinelli, G. Magna, S. De Vito, R. Di Fuccio, G. Di Francia, A. Vergara, and C. Di Natale, "An adaptive classification model based on the artificial immune system for chemical sensor drift mitigation," *Sensors and Actuators B: Chemical*, vol. 177, pp. 1017–1026, 2013.

[23] H. Drucker, C. J. Burges, L. Kaufman, A. J. Smola, and V. Vapnik, "Support vector regression machines," in *Advances in neural information processing systems*, 1997, pp. 155–161.

[24] C. A. Stow, J. Jolliff, D. J. McGillicuddy Jr, S. C. Doney, J. I. Allen, M. A. Friedrichs, K. A. Rose, and P. Wallhead, "Skill assessment for coupled biological/physical models of marine systems," *Journal of Marine Systems*, vol. 76, no. 1-2, pp. 4–15, 2009.

[25] A. Pederzoli, P. Thunis, E. Georgieva, R. Borge, D. Carruthers, and D. Pernigotti, "Performance criteria for the benchmarking of air quality model regulatory applications: the targetapproach," *International Journal of Environment and Pollution*, vol. 50, no. 1-4, pp. 175–189, 2012.

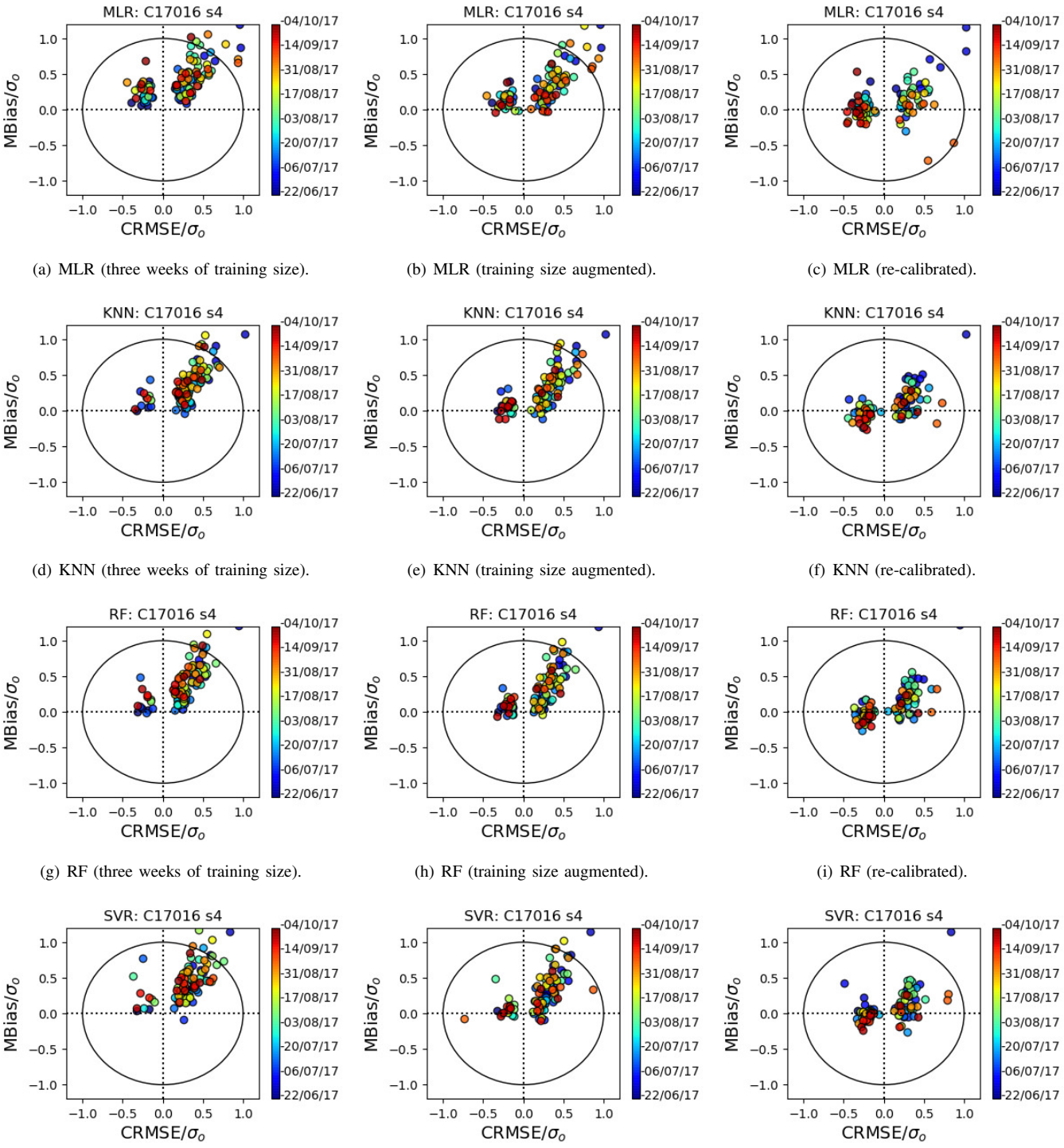
Pau Ferrer-Cid is a research assistant at the Statistical Analysis of Networks and Systems (SANS) research group, Universitat Politecnica de Catalunya (UPC). He holds a B.Sc in Computer Science and a M.Sc in Data Science by the UPC. His main research interests are the applications of novel data analysis methods to sensor data coming from IoT platforms and the analysis of other kinds of data from fields like biology and computer vision.

Jose M. Barcelo-Ordinas is an Associate Professor at Universitat Politecnica de Catalunya (UPC) from 1999. He holds a PhD and B.Sc+M.Sc in Telecommunication Engineering and a B.Sc+M.Sc in Mathematics. He has participated in many European projects such as EXPLOIT, BAF, EXPERT, NETPERF, MOEBIUS, WIDENS, EuroNGI, EuroNFI, EuroNF NoE and H2020 CAPTOR. His currently research areas are wireless sensor networks, mobility patterns, and the statistical analysis of sensor data.

Jorge Garcia-Vidal is since 2003, full professor at the Computer Architecture Department of UPC, and since 2012 responsible of the Smart Cities Initiative at Barcelona Supercomputing Center-Centro Nacional de Supercomputacion (BSC-CNS). Currently he is coordinator of the projects H2020 CAPTOR and responsible of BSC participation in the H2020 project ASGARD. His main current research interest is in problems related with the capture, processing and statistical analysis of sensor data.

Anna Ripoll is a contracted researcher at the Institute of Environmental Assessment and Water Research (IDAEA-CSIC) in Barcelona, Spain. She holds a PhD degree in environmental sciences, and has participated in several European projects including H2020 CAPTOR and ACTRIS. One of her main current research interests is the analysis of air quality sensor time series and of calibration methods and tools for sensor technologies.

Mar Viana is a staff researcher at the Institute of Environmental Assessment and Water Research (IDAEA-CSIC) in Barcelona, Spain, since 2011. Her research focuses on atmospheric aerosols, with main interests in outdoor and indoor air pollution, source apportionment and links with health. She is coordinator of ERANET project CERASAFE and IP of H2020 project CAPTOR, and participates in several other European and national projects. Her current research focuses on assessing the performance of sensor technologies for air quality monitoring.



(a) MLR (three weeks of training size).

(b) MLR (training size augmented).

(c) MLR (re-calibrated).

(d) KNN (three weeks of training size).

(e) KNN (training size augmented).

(f) KNN (re-calibrated).

(g) RF (three weeks of training size).

(h) RF (training size augmented).

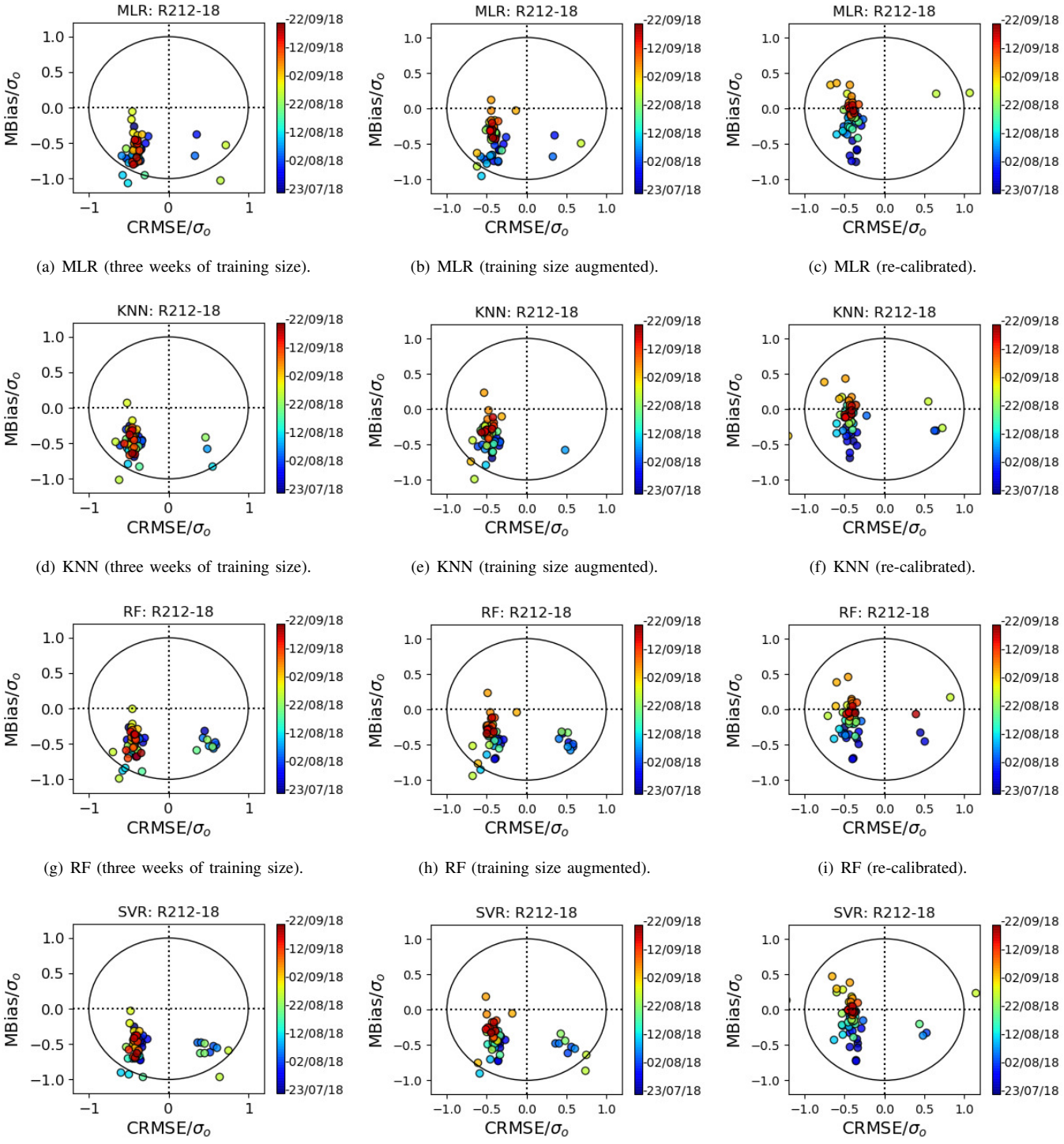
(i) RF (re-calibrated).

(j) SVR (three weeks of training size).

(k) SVR (training size augmented).

(l) SVR (re-calibrated).

Fig. 4: Long term prediction: target diagram for Captor node C17016 (sensor s4).



(a) MLR (three weeks of training size).

(b) MLR (training size augmented).

(c) MLR (re-calibrated).

(d) KNN (three weeks of training size).

(e) KNN (training size augmented).

(f) KNN (re-calibrated).

(g) RF (three weeks of training size).

(h) RF (training size augmented).

(i) RF (re-calibrated).

(j) SVR (three weeks of training size).

(k) SVR (training size augmented).

(l) SVR (re-calibrated).

Fig. 5: Long term prediction: target diagram for Raptor R212-18.