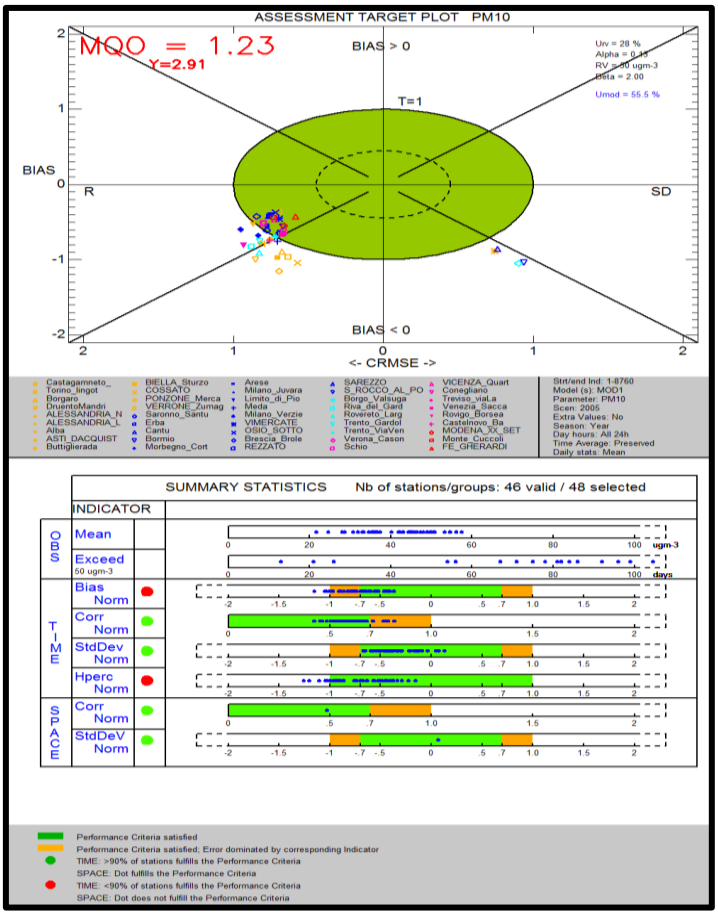
Guidance Document on Modelling Quality Objectives and Benchmarking

Stijn Janssen, Cristina Guerreiro, Peter Viaene, Emilia Georgieva, Philippe Thunis

with contributions from: Kees Cuvelier, Elke Trimpeneers, Joost Wesseling, Alexandra Montero, Ana Miranda, Jenny Stocker, Helge Rørdam Olesen, Gabriela Sousa Santos, Keith Vincent, Claudio Carnevale, Michele Stortini, Giovanni Bonafè, Enrico Minguzzi, Laure Malherbe, Frederik Meleux, Amy Stidworthy, Bino Maiheu and Marco Deserti

Version 2.1 – February 2017



FAIRMODE
Forum for air quality modelling in Europe

Table of contents

List of Figures:

List of Tables:

1. EXECUTIVE SUMMARY

The development of the procedure for air quality model benchmarking in the context of the Air Quality Directive 2008/50/EC(AQD) has been an on-going activity in the context of the FAIRMODE[1] community. Central part of the studies was the definition of proper modelling quality indicators and criteria to be fulfilled in order to allow sufficient level of quality for a given model application under the AQD. The focus at the beginning was on applications related to air quality assessment, and gradually it has been expanded to other applications, such as forecasting and planning.

The main purpose of this Guidance Document is to explain and summarize the current concepts of the modelling quality objective methodology, elaborated in various papers and documents in the FAIRMODE community, addressing mainly model applications for air quality assessment. Other goals of the Document are linked to presentation and explanation of templates for harmonized reporting of modelling results. Giving an overview of still open issues in the implementation of the presented methodology, the Document aims at triggering further research and discussions.

A core set of statistical indicators is defined using pairs of measurement-modelled data. The core set is the basis for the definition of a modelling quality indicator (MQI) and additional modelling performance indicators (MPI), which also take into account the measurement uncertainty.

The MQI describes the discrepancy between measurements and modelling results (linked to RMSE), normalized by measurement uncertainty and a scaling factor. The modelling quality objective (MQO) requires MQI to be less than or equal to 1. With an arbitrary selection of the scaling factor of 2, the fulfilment of the MQO means the allowed deviation between modelled and measured concentrations is twice the measurement uncertainty. Expressions for the MQI calculation based on time series and yearly data are introduced. MPI refer to aspects of correlation, bias and standard deviation, also MPI related to spatial variation are defined. The modelling performance criteria (MPC) are defined for the MPI; they are necessary, but not sufficient criteria to determine whether the MQO is fulfilled. The MQO is required to be fulfilled at 90% of the stations, a criteria which is implicitly taken into account in the derivation of the MQI. The associated modelling uncertainty is formulated, showing that in case of MQO fulfilment the modelling uncertainty must not exceed 1.75 times the measurement one (with the scaling factor fixed to 2).

The measurement uncertainty is expressed in terms of concentration and its associated uncertainty. The methodology for estimating the measurement uncertainty is overviewed and the parameters for its calculation for PM, NO_2 and O_3 are provided. An expression for the associated modelling uncertainty is also given.

A reporting template is presented and explained for hourly and yearly average data. In both cases there is a diagram and a table with summary statistics. In a separate section open issues are

[1] The Forum for Air quality Modeling (FAIRMODE) is an initiative to bring together air quality modelers and users in order to promote and support the harmonized use of models by EU Member States, with emphasis on model application under the European Air Quality Directives. FAIRMODE is currently being chaired by JRC.

discussed and an overview of related publications and tools is provided. Finally, a chapter on modelling quality objectives for forecast models is introduced. This methodology is still under discussion and need some further fine tuning and testing.

2. VERSION HISTORY

Version	Release date	Modifications
1.0	6/02/2015	First version
1.1	1/04/2015	- NILU application added to Examples of good practice. - Small textual corrections.
2.0	26/05/2016	- Update of Section 8, definition of Modelling Quality Objective - Update of Section 9 on performance reporting - Open Issue list updated according to ongoing discussions within WG1 - Update of CERC, NILU and IRCEL contribution in the Best Practice section
2.1	13/02/2017	- Improvement of the overall readability of the text. - Introduction of a chapter on “Definitions” & “Related Tools” - Update of Section on Modelling uncertainty - Introduction of a chapter on MQO for forecast models - Removal of Section “Examples of Best Practice”. The user feedback will be reused in a publication (*Montero et al, 2017*)

3. INTRODUCTION

3.1. Why Modelling Quality Objectives?

In general, the quality of models is understood in terms of their 'fitness for purpose'. The modelling experience indicates that there are no 'good' or 'bad' models. Evidence is rather based on the question of whether a model is suitable for the intended application and specified objectives. As such, the quality of a model is always relative and is measured against the quality objectives for any particular model application. Statistical performance indicators which provide insight on model performance are generally used to assess the performance of a model against measurements for a given application. They do not however tell whether model results have reached a sufficient level of quality for a given application. This is the reason for which modelling quality objectives (MQO), defined as the minimum level of quality to be achieved by a model for policy use, need to be set.

Modelling quality objectives are described in Annex I of the Air Quality Directive 2008/50/EC (AQD) along with the monitoring quality objectives. They are expressed as a relative uncertainty (%) which is then further defined in the AQ Directive. But as mentioned in the FAIRMODE technical guidance document [http://fairmode.jrc.ec.europa.eu/downloads.html] the wording of the AQD text needs further clarification in order to become operational. It is important to note that these modelling quality objectives apply only to assessment of the current air quality when reporting exceedances, and do not refer to other model applications, such as planning or forecasting. However, there is clearly an expectation when using models for these other applications that they have been verified and validated in an appropriate, albeit unspecified, way.

3.2. What are the purposes of this Document?

The main objectives of this Guidance Document are to:

- Explain MQO concepts and methodology developed within FAIRMODE
- Provide recommendations and guidance for accessing model performance related to a given air quality model application in the frame of the AQD, based on the experience and elaborations in the FAIRMODE community. In a first step PM_{10}, $PM_{2.5}$, NO_2 and O_3 are prioritized but ultimately the methodology should also cover other pollutants such as heavy metals and polycyclic aromatic hydrocarbons. The focus of this document is mainly on the use of air quality models for the assessment of air quality, however hints for forecast applications are also provided.
- Promote consistency in model evaluation for policy applications related to the AQD
- Promote harmonized reporting of modelling performance in the EU Member States
- Promote further development of open issues

3.3. Who is the target audience of this Document?

This Guidance Document is intended primary to environmental experts using air quality models in the context of the EU AQD. Some of these experts apply tools (software), developed around the concepts in this document (as DELTA, ATMOSYS or MyAir, see Section 12) and thus, the current text provides additional support to the respective Users' Guides. Developers of the mentioned tools might also benefit from the notes in this document.

A wider target audience consists of air quality modellers, who are interested in methods and criteria for evaluating model performance and follow recent developments in various modelling communities.

3.4. What are the main components of this Document?

This Document is built upon the following major components:

- The concept and methodology for modelling quality objectives and modelling performance criteria developed within the FAIRMODE community (Chapter 7)
- The techniques for reporting of model performance in harmonized way (Chapter 8)
- Opens issues to the above components, which merit consideration and further development within the FAIRMODE community (Chapter 9)
- Summary of additional resources (publications, tools), (Chapters 10, 12 and 13)

This Guidance Document is periodically revised to ensure that new FAIRMODE developments or expanded regulatory requirements are incorporated, as well as to account for User's feedback.

4. BENCHMARKING: A WORD OF CAUTION

Based on the UNESCO[2] definition, adapted to the context of air quality modelling, benchmarking can be defined as follows:

- a standardized method for collecting and reporting model outputs in a way that enables relevant comparisons, with a view to establishing good practice, diagnosing problems in performance, and identifying areas of strength;
- a self-improvement system allowing model validation and model inter-comparison regarding some aspects of performance, with a view to finding ways to improve current performance;
- a diagnostic mechanism for the evaluation of model results that can aid the judgment of models quality and promote good practices.

When we talk about benchmarking, it is normally implicitly assumed that the best model is one which produces results the closest to measured values. In many cases, this is a reasonable assumption. However, it is important to recognize that this is not always the case, so one should proceed with caution when interpreting benchmarking results. Here are three examples in which blind faith in benchmarking statistics would be misplaced:

- Emission inventories are seldom perfect. If not all emission sources are included in the inventory used by the model then a perfect model should not match the observations, but have a bias. In that case seemingly good results would be the result of compensating errors.
- If the geographical pattern of concentrations is very patchy – such as in urban hot spots – monitoring stations are only representative of a very limited area. It can be a major challenge – and possibly an unreasonable challenge – for a model to be asked to reproduce such monitoring results.
- Measurement data are not error free and a model should not always be in close agreement with monitored values.

In general, in the EU member states there are different situations which pose different challenges to modelling including among others the availability of input data, emission patterns and the complexity of atmospheric flows due to topography.

The implication of all the above remarks is that if one wishes to avoid drawing unwarranted conclusions from benchmarking results, then it is not sufficient to inspect benchmarking results. Background information should be acquired on the underlying data to consider the challenges they represent.

[2] Vlãsceanu, L., Grünberg, L., and Pârlea, D., 2004, /Quality Assurance and Accreditation: A Glossary of Basic Terms and Definitions /(Bucharest, UNESCO-CEPES) Papers on Higher Education, ISBN 92-9069-178-6. http://www.cepes.ro/publications/Default.htm

Good benchmarking results are therefore not a guarantee that everything is perfect. Poor benchmarking results should be followed by a closer analysis of their causes. This should include examination of the underlying data and some exploratory data analysis.

Benchmarking in the context of FAIRMODE strategy is intended as the compilation of different approaches and the subsequent development and testing of a standardized evaluation/inter-comparison methodology for collecting and reporting model inputs/outputs in a way that enables relevant comparisons. The aim is to identify good practices and propose ways to diagnose problems in performance.

5. DEFINITIONS

Modelling Quality Indicator (MQI)

In the context of this Guidance document, a Modelling Quality Indicator is a statistical indicator calculated on the basis of measurements and modelling results. It is used to determine whether the Modelling Quality Objectives are fulfilled. It describes the discrepancy between measurements and modelling results, normalized by the measurement uncertainty and a scaling factor. It is calculated according to the equation (**Error! Reference source not found.**) or (18).

Modelling Quality Objective (MQO)

Criteria for the value of the Modelling Quality Indicator (MQI). The MQO is said to be fulfilled if MQI is less than or equal to unity.

Modelling Performance Indicator (MPI)

They are statistical indicators calculated on the basis of measurements and modelling results. In the context of the present Guidance document several Modelling Performance Indicators are defined. Each of the Modelling Performance indicators describes a certain aspect of the discrepancy between measurement and modelling results. Thus, there are Modelling Performance Indicators referring to the three aspects of correlation, bias and normalized mean square deviation. Furthermore, there are Model Performance Indicators related to spatial variation. Finally, also the Modelling Quality Indicator might be regarded as a MPI. However, it has a special status and is assigned its own name because it determines whether the MQO is fulfilled. See section 7.5 for definitions of the MPI's.

Modelling Performance Criteria

Criteria that Model Performance Indicators are expected to fulfil. Such criteria are defined for certain MPI's. They are necessary, but not sufficient criteria to determine whether the Modelling Quality Objectives are fulfilled.

Measurement uncertainty

Uncertainty related to the measurement of ambient concentrations. FAIRMODE relied on the expertise of the AQUILA network to define those quantities.

Modelling uncertainty

Associated to measurement uncertainty, calculated with equation (22)

Model evaluation

The sum of processes that need to be followed in order to determine and quantify a model's performance capabilities, weaknesses and advantages in relation to the range of applications for which it has been designed.

Note: The present Guidance document does not prescribe a procedure for model evaluation.

[SOURCE: EEA Technical Reference Guide No. 10, 2011]

Model validation

Comparison of model predictions with experimental observations, using a range of model quality indicators.

[SOURCE: EEA Technical Reference Guide No. 10, 2011]

6. MAIN ASSUMPTIONS

The focus of this Guidance Document is on presenting the modelling quality objective (MQO) and associated modelling performance criteria (MPC) for different statistical indicators related to a given air quality model application for air quality assessment in the frame of the AQD. These statistical indicators are produced by comparing air quality model results and measurements at monitoring sites. This has the following consequences:

1. Data availability

A minimum of data availability is required for statistics to be produced at a given station. Presently the requested percentage of available data over the selected period is 75%. Statistics for a single station are only produced when data availability of paired modelled and observed data is for at least 75% of the time period considered. When time averaging operations are performed the same availability criteria of 75% applies. For example, daily averages will be performed only if data for 18 hours are available. Similarly, an 8 hour average value for calculating the O_3 daily maximum 8-hour means is only calculated for the 8 hour periods in which 6 hourly values are available. In open issues §9.4 the choice of the data availability criterion is further elaborated.

2. Species and time frame considered

The modelling quality objective (MQO) and modelling performance criteria (MPC) are in this document defined only for pollutants and temporal scales that are relevant to the AQD. Currently only O_3, NO_2, PM_{10} and $PM_{2.5}$ data covering an entire calendar year are considered.

3. Fulfilment criteria

According to the Data Quality Objectives in Annex I of the AQD the uncertainty for modelling is defined as the maximum deviation of the measured and calculated concentration levels for 90 % of individual monitoring points over the period considered near the limit value (or target value in the case of ozone) and this without taking into account the timing of the events. While the MQO and MPC proposed in this document do consider the timing of the events, we also need to select a minimum value for the number of stations in which the model performance criterion has to be fulfilled and propose to also set this number to 90 %. This means that the model performance criteria must be fulfilled for at least 90% of the available stations. This is further detailed in Section 7.3.3.

4. Measurement uncertainty [3]

A novelty in the concept for defining MQO and MPC is the introduction of measurement uncertainty in the respective statistical parameters. The measurement uncertainty is expressed as dependent on the concentration. Methods for estimating the parameters for the key species treated are further explained in Section 7.2.

[3] In previous versions of the Guidance Document, this term was often interchanged with "observation uncertainty". Further on, we will use only the term "measurement uncertainty".

Main assumptions

- Pollutants covered: O_3, NO_2, PM_{10} and $PM_{2.5}$
- Paired data series of model results and observations at fixed locations
- Minimum data availability 75% for the selected time interval
- Fulfilment criteria at 90% of individual locations
- Measurement uncertainty included in the modelling quality indicator (MQI) and in the modelling performance indicators (MPI)

7. MODELLING INDICATORS AND CRITERIA

7.1. Statistical indicators

Models applied for regulatory air quality assessment are commonly evaluated on the basis of comparisons against observations. This element of the model evaluation process is also known as operational model evaluation or statistical performance analysis, since statistical indicators and graphical analysis are used to determine the capability of an air quality model to reproduce measured concentrations. It is generally recommended to apply multiple statistical indicators regardless of the model application since each one has its advantages and disadvantages.

To cover all aspects of the model performance in terms of amplitude, phase and bias the following ***core set*** of statistical indicators has been proposed within FAIRMODE for the statistical analysis of model performance with M_i and O_i respectively the modelled and observed values where i is a number (rank) between 1 and N and N the total number of modelled or observed values:

Table 1: Core set of statistical indicators

Indicator	Formula	
Root Mean Square Error (***RMSE***)	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}$ with $\bar{O} = \frac{\sum_{i=1}^{N} O_i}{N}$ the average observed value and $\bar{M} = \frac{\sum_{i=1}^{N} M_i}{N}$ the average modelled value.	(1)
Correlation coefficient ***(R)***	$R = \frac{\sum_{i=1}^{N}(M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(M_i - \bar{M})^2}\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}}$	(2)
Normalised Mean Bias ***(NMB)***	$NMB = \frac{BIAS}{\bar{O}}$ where $BIAS = \bar{M} - \bar{O}$	(3)
Normalised Mean Standard Deviation ***(NMSD)***	$NMSD = \frac{(\sigma_M - \sigma_O)}{\sigma_O}$ with $\sigma_O = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - \bar{O})^2}$ the standard deviation of the observed values and $\sigma_M = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2}$ the standard deviation of the modelled values.	(4)

Although statistical performance indicators provide insight on model performance in general they do not tell whether model results have reached a sufficient level of quality for a given application, e.g. for policy support. In the literature different recommended values can be found for some of the statistical indicators for assessment of modelling performance.

In the FAIRMODE community a main statistical indicator has been introduced based on measurement and modelling results - the modelling quality indicator (MQI). Then a criteria for this MQI is defined, the modelling quality objective (MQO), representing the minimum level of quality to be achieved by a model for policy use. A specific feature of the MQI is its link to the measurement uncertainty. In the following, first the measurement uncertainty will be discussed, as a second step the formulation for MQI and MQO will be presented.

Core set of statistical indicators

- Includes: Root Mean Square Error, Correlation, Normalised Mean Bias, Normalized Mean Standard Deviation
- Serves as a basis for the definition of the main model quality indicator MQI (linked to RMSE) and additional Model Performance Indicators (MPI), linked to the remaining core statistical indicators.

7.2. Measurement uncertainty

7.2.1. A simple expression for the measurement uncertainty U(O_i) for time series

We derive here a simplified and general expression for the measurement uncertainty U(O_i). U(O_i) represents the expanded measurement uncertainty which can be expressed in terms of the combined uncertainty, $u_c(O_i)$ by multiplying with a coverage factor k:

$$U(O_i) = k u_c(O_i). \quad (5)$$

Each value of k gives a particular confidence level so that the true value is within the confidence interval bounded by $O_i \pm k u_c(O_i)$. Coverage factors of k = 2.0 and k = 2.6 correspond to confidence levels of around respectively 95 and 99%, so that the unknown true value lies within the estimated confidence intervals.

In Thunis *et al.*, 2013 a general expression for the combined measurement uncertainty is derived by considering that $u_c(O_i)$ of a measurement O_i, can be decomposed into a component that is proportional, $u_p(O_i)$ to the concentration level and a non-proportional contribution, $u_{np}(O_i)$:

$$u_c{}^2(O_i) = u_p{}^2(O_i) + u_{np}{}^2(O_i) \quad (6)$$

The non-proportional contribution $u_{np}(O_i)$ is by definition independent of the concentration and can therefore be estimated at a concentration level of choice that is taken to be the reference value (RV). If $u_r{}^{RV}$ represents the relative standard measurement uncertainty around the reference value

(RV) for a reference time averaging, e.g. the daily/hourly Limit Values of the AQD, then $u_{np}(O_i)$ can be defined as a fraction α (ranging between 0 and 1) of the uncertainty at the reference value:

$$u_{np}{}^2(O_i) = \alpha^2 (u_r{}^{RV} RV)^2 \quad (7)$$

Similarly the proportional component $u_p(O_i)$ can be estimated from:

$$u_p{}^2(O_i) = (1-\alpha^2)\,(u_r{}^{RV} O_i)^2 \quad (8)$$

As representative values of the measurement uncertainty, we will select in our formulation the 95th percentile highest values among all uncertainty values calculated, $U_{95}(O_i)$. To derive expressions for the uncertainty estimation for PM_{10} and $PM_{2.5}$ the results of a JRC instrument inter-comparison (Lagler et al. 2011) have been used, whereas a set of EU AIRBASE stations available for a series of meteorological years has been used for NO_2 , and analytical relationships have been used for O_3. We will use the subscript "95" to recall this choice, i.e.:

$$ku_c^{RV} = ku_{95,c}^{RV} = U_{95}^{RV} \quad (9)$$

Combining (6) – (9) U_{95} can be expressed as:

$$U_{95}(O_i) = U_{95,r}^{RV}\sqrt{(1-\alpha^2)O_i^2 + \alpha^2 . RV^2} \quad (10)$$

From Equation (10) it is possible to derive an expression for RMS_U as:

$$RMS_U = \sqrt{\frac{\sum_{i=1}^{N}\left(U_{95}(O_i)\right)^2}{N}} = U_{95,r}^{RV}\sqrt{(1-\alpha^2)(\bar{O}^2 + \sigma_o^2) + \alpha^2 . RV^2} \quad (11)$$

in which $\bar{O}$ and σ_0 are the mean and the standard deviation of the measurement time series, respectively.

7.2.2. Measurement uncertainty for yearly averaged data

Pernigotti et al (2013) derive the following expression for the expanded 95th percentile measurement uncertainty of the mean concentration:

$$\mathrm{U_{95}}(\bar{\mathrm{O}}) = \mathrm{U_{95,r}^{RV}}\sqrt{\frac{(1-\alpha^2)}{\mathrm{N_p^*}}(\bar{\mathrm{O}}^2 + \sigma_\mathrm{o}^2) + \frac{\alpha^2 . \mathrm{RV}^2}{\mathrm{N_{np}}}} \cong \mathrm{U_{95,r}^{RV}}\sqrt{\frac{(1-\alpha^2)}{\mathrm{N_p}}\bar{\mathrm{O}}^2 + \frac{\alpha^2 . \mathrm{RV}^2}{\mathrm{N_{np}}}} \quad (12)$$

where N_p and N_{np} are two coefficients that are only used for annual averages and that account for the compensation of errors (and therefore a smaller uncertainty) due to random noise and other factors like periodic re-calibration of the instruments. Details on the derivation of (12) and in particular the parameters N_p and N_{np} are provided in Pernigotti et al. (2013).

7.2.3. Parameters for calculating the measurement uncertainty

The values for the parameters in equations (11) and (12) are given in Table 2. All values are as reported in Pernigotti et al. (2013) and Thunis et al. (2012) with the exception of the N_p and N_{np} parameters for PM_{10} that have been updated to better account for the yearly average measurement uncertainty range with current values set to reflect uncertainties associated to the β-ray measurement technique. Because of insufficient data for $PM_{2.5}$, values of N_p and N_{np} similar to those for PM_{10} have been set. The value of U_r^{RV} has also been updated for O_3 where the coverage factor (k) has been updated to 2 (not 1.4 as in Thunis et al. 2012).

Note also that the value of α for $PM_{2.5}$ referred to in the Pernigotti et al. (2014) working note has been arbitrarily modified from 0.13 to 0.30 to avoid larger uncertainties for PM_{10} than $PM_{2.5}$ in the lowest range of concentrations.

Table 2: List of the parameters used to calculate the measurement uncertainty

	$U_{95,r}^{RV}$	RV	α	N_p	N_{np}
NO_2	0.24	200 µg/m3	0.20	5.2	5.5
O_3	0.18	120 µg/m3	0.79	11	3
PM_{10}	0.28	50 µg/m3	0.13	30	0.25
$PM_{2.5}$	0.36	25 µg/m3	0.30	30	0.25

7.2.4. Practical derivation of the measurement uncertainty parameters

We provide here some hints for the estimation of U_{95}, the relative uncertainty around a reference value and α, the non-proportional fraction around the reference value. We re-write equation (10) as:

$$U_{95}^2 = \alpha^2 (U_{95}^{RV})^2 + \left(U_{r,95}^{RV}\right)^2 (1-\alpha^2) O_i^2 \quad (13)$$

This is a linear relationship with slope, $m = (1-\alpha^2)\left(U_{r,95}^{RV}\right)^2$ and intercept, $q = \alpha^2 (U_{95}^{RV})^2$ which can be used to derive values for $U_{r,95}^{RV}$ and α by fitting measured squared uncertainties U_{95}^2 to squared observed values $(O_i)^2$.

An alternative procedure for calculating U_{95}^2 and α can be derived by rewriting (10) as:

$$U_{95}^2 = (U_{95}^L)^2 + \frac{\left(U_{r,95}^{RV}\right)^2 - \left(U_{r,95}^L\right)^2}{RV^2 - L^2} (O_i)^2 \quad (14)$$

where L is a low range concentration value (*i.e.* close to zero) and U_{95}^L its associated expanded uncertainty. Comparing the two formulations (13) and (14) we obtain:

$$\alpha = \frac{U_{95}^{L}}{U_{95}^{RV}}$$

$$\left(U_{95}^{L}\right)^2 = \left(U_{95}^{RV}\right)^2 - \left(\frac{U_{95}^{RV}}{RV}\right)^2 (1-\alpha^2)(RV^2 - L^2)$$

The two above relations (13) and (14) allow switching from one formulation to the other. The first formulation (13) requires defining values for both α and U_{95}^{RV} around an arbitrarily fixed reference value (RV) and requires values of U_{95}^2 over a range of measured concentrations, while the second formulation (14) requires defining uncertainties around only two arbitrarily fixed concentrations (RV and L).

7.3. Modelling quality indicator (MQI) and modelling quality objective (MQO)

7.3.1. MQI and MQO for time series data

The Modelling Quality Indicator (MQI) is a statistical indicator calculated on the basis of measurements and modelling results. It is defined as the ratio between the model-measured bias at a fixed time (i) and a quantity proportional to the measurement uncertainty as:

$$\mathrm{MQI} = \frac{|\mathrm{O_i} - \mathrm{M_i}|}{\beta \mathrm{U_{95}(O_i)}} \tag{15}$$

with β as a coefficient of proportionality.

The MQO is the criteria for the MQI. The MQO is fulfilled when the MQI is less or equal to 1, i.e.:

$$\mathrm{MQO: MQI} \leq 1 \tag{16}$$

In Figure 1, the MQO is fulfilled for example on days 3 to 10, whereas it is not fulfilled on days 1, 2 and 11. We will also use the condition $|\mathrm{O_i} - \mathrm{M_i}| \leq \mathrm{U_{95}(O_i)}$ in the MQO related diagrams (see Section 8) to indicate when model-measurement differences are within the measurement uncertainty (e.g. days 5 and 12 in Figure 1).

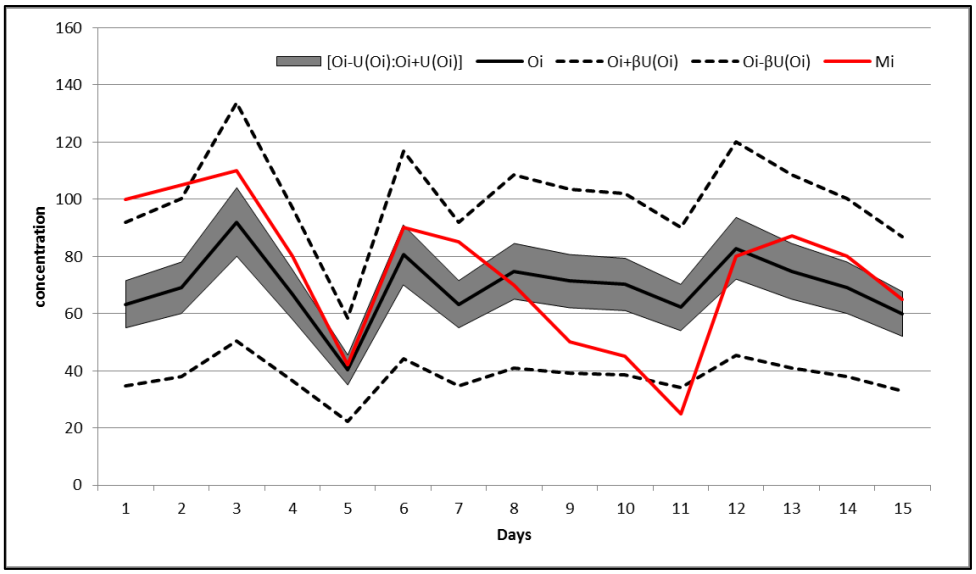


Figure 1: MQI and MQO explained on PM_{10} time series: measured (bold black) and modelled (bold red) concentrations are represented for a single station. The grey shaded area is indicative of the measurement uncertainty whereas the dashed black lines represent the MQI limits (proportional to the measurement uncertainty). Modelled data fulfilling the MQO must be within the dashed lines.

Equation (15) can then be used to generalize the MQI to a time series:

$$MQI = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}}{\beta\sqrt{\frac{1}{N}\sum_{i=1}^{N} U_{95}(\mathrm{O}_i)^2}} = \frac{\mathrm{RMSE}}{\beta RMS_U} \quad \text{and MQO: MQI} \leq 1 \tag{17}$$

With this MQO formulation, the RMSE between measured Oi and modelled Mi values (numerator) is compared to a value representative of the maximum allowed uncertainty (denominator). The value of β determines the stringency of the MQO. From now on in this Guidance Document (and in the DELTA tool) β is arbitrarily set equal to 2, allowing thus deviation between modelled and measured concentrations as twice the measurement uncertainty.

7.3.2. MQI and MQO for yearly average model results

For air quality models that provide yearly averaged pollutant concentrations, the MQI is defined as the mean bias between modelled and measured annual averaged concentrations normalized by the expanded measurement uncertainty of the mean concentration:

$$MQI = \frac{|\bar{O} - \bar{M}|}{\beta U_{95}(\bar{\mathrm{O}})} \qquad and\ MQO: MQI \leq 1 \tag{18}$$

The MQO is fulfilled when the MQI is less than or equal to 1, as in the case for time series data.

7.3.3. The 90% principle

For all statistical indicators used in DELTA for benchmarking purposes the approach currently used in the AQD has been followed. This means that the MQO must be fulfilled for at least 90% of the

available stations. The practical implementation of this approach consists in calculating the MQI associated to each station, rank them in ascending order and inferring the 90th percentile value according to the following linear interpolation (for nstat station):

$$\mathrm{MQI}_{90\mathrm{th}} = \mathrm{MQI}(\mathrm{stat}_{90}) + [\mathrm{MQI}(\mathrm{stat}_{90}+1) - \mathrm{MQI}(\mathrm{stat}_{90})] * \mathrm{dist} \quad (19)$$

where stat90 = integer(nstat*0.9) and dist= $[\mathrm{nstat} * 0.9 - \mathrm{integer}(\mathrm{nstat} * 0.9)]$. If only one station is used in the benchmarking, $\mathrm{MQI}_{90\mathrm{th}} = \mathrm{MQI}(\mathrm{station}) * 0.9$.

The MQO is then expressed as:

$$\mathrm{MQO}: \quad \mathrm{MQI}_{90\mathrm{th}} \leq 1 \quad (20)$$

7.4. Calculation of the associated model uncertainty

To give some information about the model uncertainty, we will use the normalized deviation indicator E_n (ref: ISO 13528). It scales the model (M)-measurement (O) difference with the measurement and model uncertainties [$\mathrm{U}_{95}(\mathrm{O_i})$ and $\mathrm{U}(\mathrm{M_i})$] associated to this difference:

$$\mathrm{E_n} = \frac{|\mathrm{O_i} - \mathrm{M_i}|}{\sqrt{\mathrm{U}_{95}(\mathrm{O_i})^2 + \mathrm{U}(\mathrm{M_i})^2}} \quad (21)$$

E_n equals to unity implies that the model and measurement uncertainties are compatible with the model-measurement bias. We use this relation, i.e. E_n=1, in DELTA to estimate the minimum model uncertainty compatible with the resulting model-measurement bias as follows:

$$\mathrm{E_n} = 1 \Rightarrow \mathrm{U}(\mathrm{M_i}) = \mathrm{U}_{95}(\mathrm{O_i}) \sqrt{\left(\frac{\mathrm{O_i} - \mathrm{M_i}}{\mathrm{U}_{95}(\mathrm{O_i})}\right)^2 - 1} \quad (22)$$

Relation (22) does not apply to cases in which $|O_i - M_i| < U_{95}(O_i)$, i.e. when the bias is inferior to the measurement uncertainty, cases in which no meaningful improvement of the model can be made. It is interesting to note that the fulfilment of the MQO proposed in (16) and (18) implies therefore that the model uncertainty must not exceed 1.75 times the measurement one [this value if obtained by substituting the bias term in (22) by its maximum allowed value in the MQO, i.e. $\beta \mathrm{U}_{95}$(Oi) with β=2].

We can generalise equation (22) for a time series and for time averaged values as:

$$RMS_{U_M} = RMS_U \sqrt{\left(\frac{\mathrm{RMSE}}{RMS_U}\right)^2 - 1} \quad (23)$$

and

$$U(\bar{M}) = U_{95}(\bar{O}) \sqrt{\left(\frac{\mathrm{Bias}}{U_{95}(\bar{O})}\right)^2 - 1} \quad (24)$$

In DELTA the value of the ratio (RMS_{U_M}/RMS_U) or ($U(\bar{M})/U_{95}(\bar{O})$) is used to scale the measurement uncertainty around the reference value ($U_{95,r}^{RV}$) and to provide information about the minimum model uncertainty reached around the reference value.
The 90% principle (see Section 7.3.3) is also applied to the corresponding model uncertainty. The minimum model uncertainty is the value of the uncertainty associated to the 90th percentile station. This information is provided in some benchmarking diagrams (see Section 8).

Measurement uncertainty for a given pollutant:

- Depends on the concentration
- is estimated as 95th percentile highest value, based on: JRC instrument inter-comparison results (for PM); data from EU AIRBASE stations for series of meteorological years (for NO_2) ; and on analytical relationships (for O_3)
- is expressed by (11) for time series
- is given by (12) for yearly averaged values
- uses parameters in its calculation as defined in Table 2

7.5. Comparison to values in the AQD

The Table below lists the values currently used in FAIRMODE as compared to those available in the AQD. The data quality objective (DQO) and the maximum bias at limit value (defined as model quality objective in the AQD) can be compared with the reference measurement uncertainty around the limit value LV, $U_{O,r}^{LV}$ and the maximum bias used in FAIRMODE. Obviously the FAIRMODE max bias is concentration dependent and applies to the whole range of concentration (equal to $2U_O$) but is only reported here around the LV. The last column shows the modelling uncertainty. Note that the values are obtained with fixed β =2.

Table 3: Comparison to AQD values

			2008 AQ Directive			FAIRMODE	
	Frequency	*Limit value* ug/m^3	*DQO at LV*	*Max bias at LV*	$U_{O,r}^{LV}$	*Max bias at LV*	$U_{M,r}^{LV}$
NO2	Hour	200	15%	50%	24.0%	48%	42%
	Year	40	-	30%	14.5%	29%	25%
O3	8h	120	15%	50%	18%	36%	18%
PM10	day	50	25%	-	28%	56%	49%
	year	40	-	50%	6.4%	13%	11%
PM25	Day	"25 ug/m3"	25%	-	36%	72%	63%
	year		-	50%	10%	20%	17%

7.6. Modelling performance indicators (MPI) and criteria (MPC)

Modelling performance indicators are statistical indicators that describe certain aspect of the discrepancy between measurement and modelling results. The MQI can be treated as a kind of MPI related to one of the core statistical parameters defined in 7.1, namely the RMSE. We define here MPI related to correlation, bias and standard deviation (i.e. the remaining core statistical parameters). Furthermore, we define also MPI related to the spatial variability. The criteria that MPI are expected to fulfil are defined as modelling performance criteria (MPC).

7.6.1. Temporal MPI and MPC

A characteristic of the proposed MQI is that errors in BIAS, σ_M and R are condensed into a single number. These three different statistics are however related as follows:

$$\mathrm{MQI}^2 = \frac{\mathrm{RMSE}^2}{(\beta\mathrm{RMS_U})^2} = \frac{\mathrm{BIAS}^2}{(\beta\mathrm{RMS_U})^2} + \frac{(\sigma_M - \sigma_O)^2}{(\beta\mathrm{RMS_U})^2} + \frac{2\sigma_O\sigma_M(1-\mathrm{R})}{(\beta\mathrm{RMS_U})^2} \qquad (25)$$

By considering ideal cases where two out of three indicators perform perfectly, separate MPI and respective MPC can be derived from (25) for each of these three statistics. For example, assuming R=1 and $\sigma_M = \sigma_O$ in equation (25) leads to an expression for the bias model performance indicator (MPI) and bias model performance criterion (MPC) as:

$$\mathrm{MPI} = \frac{\mathrm{BIAS}}{(\beta\mathrm{RMS_U})} \quad \text{and} \quad \mathrm{MPC}: \frac{\mathrm{BIAS}}{\beta\mathrm{RMS_U}} \leq 1$$

.

This approach can be generalised to the other two temporal MPI (see Table 4).

Table 4: Model performance indicators and criteria for temporal statistics

MPI	MPC	
BIAS $(R = 1, \sigma_O = \sigma_M)$	$\lvert BIAS \rvert \leq \beta\ RMS_U$	(26)
R $(BIAS = 0, \sigma_O = \sigma_M)$	$\mathrm{R} \geq 1 - 0.5\beta^2 \frac{{RMS_U}^2}{\sigma_O \sigma_M}$	(27)
Std. dev. $(BIAS = 0, R = 1)$	$\lvert \sigma_M - \sigma_O \rvert \leq \beta RMS_U$	(28)

One of the main advantages of this approach for deriving separate MPI is that it provides a selection of statistical indicators with a consistent set of performance criteria based on one single input: the measurement uncertainty U(O_i). The MQI is based on the RMSE indicator and provides a general overview of the model performance while the associated MPI for correlation, standard deviation and bias can be used to highlight which of the model performance aspects need to be improved. It is important to note that **the MPC for bias, correlation, and standard deviation represent necessary but not sufficient conditions to ensure fulfilment of the MQO.**

7.6.2. Spatial MPI and MPC

Spatial statistics are calculated in the benchmarking performance report (see Chapter 8). For hourly frequency, the model results are first averaged yearly at each station. A correlation and a standard deviation indicator are then calculated for this set of averaged values. Formulas (27) and (28) are still used but $\mathrm{RMS_U}$ is substituted by $\mathrm{RMS_{\overline{U}}}$ where $\mathrm{RMS_{\overline{U}}} = \sqrt{\frac{1}{\mathrm{N}} \sum \mathrm{U}(\overline{\mathrm{O}})^2}$. The same approach holds for yearly frequency output.

Table 5: Model performance indicators and criteria for spatial statistics

MPI	MPC	
Correlation	$\mathrm{R} \geq 1 - 0.5\beta^2 \frac{{RMS_{\overline{U}}}^2}{\sigma_O \sigma_M}$	(29)
Std. dev.	$\lvert \sigma_M - \sigma_O \rvert \leq \beta RMS_{\overline{U}}$	(30)

Modelling Quality Indicator (MQI) and Model Quality Objective (MQO):

- MQI is the main modelling performance indicator
- MQI is defined as the ratio of the RMSE between measured and modelled values and a value proportional to the measurement uncertainty RMS_U
- The proportionality coefficient β is arbitrarily set equal to 2, allowing thus deviation between modelled and measured concentrations as large as twice the measurement uncertainty.
- MQI for time series is given by (**Error! Reference source not found.**), for yearly averaged data by (18). Values for MQI based on time series or annual data may differ
- MQO is the criteria for MQI: MQO is fulfilled when MQI is less than or equal to 1
- MQO does not depend on pollutant , scale and data frequency

Modelling Performance Indicators (MPI) and Model Performance Criteria (MPC):

- MPI are performance indicators additional to the main MQI, highlighting which aspect of the modelling result needs to be improved
- MPI are related to temporal correlation, bias, and standard deviation, spatial correlation and bias. They all depend on measurement uncertainty
- MPC are the criteria for MPI, defined in Table 4
- MPC represent necessary but not sufficient conditions to ensure fulfilment of the MQO

8. REPORTING MODEL PERFORMANCE

Benchmarking reports are currently defined for the hourly NO_2, the 8h daily maximum O_3 and daily PM_{10} and $PM_{2.5}$ concentrations. The reports for the evaluation of hourly and yearly average model results are different. Below we present details for these two types of reports.

8.1. Hourly data

The report consists of a Target diagram followed by a summary table.

8.1.1. Target Diagram

The MQI as described by Eq (**Error! Reference source not found.**) is used as main indicator. In the uncertainty normalised Target diagram, the MQI represents the distance between the origin and a given station point, for this reason in previous documents the MQI was called also target indicator. The performance criterion for the MQI, defined as MQO, is set to unity regardless of spatial scale and pollutant and it is expected to be fulfilled by at least 90% of the available stations. A MQI value representative of the 90th percentile is calculated according to (19).

In the Target diagram, Figure 2, the X and Y axis correspond to the BIAS and $CRMSE$ which are normalized by the measurement uncertainty, RMS_U. The $CRMSE$ is defined as:

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(M_i - \bar{M}) - (O_i - \bar{O})]^2} \tag{31}$$

and is related to RMSE and BIAS as follows:

$$RMSE^2 = BIAS^2 + CRMSE^2 \tag{32}$$

and to the standard deviation, σ and correlation, R :

$$CRMSE^2 = \sigma_o^2 + \sigma_m^2 - 2\sigma_o\sigma_m R \tag{33}$$

For each point representing one station on the diagram the ordinate is then $BIAS/\beta RMS_U$, the abscissa is $CRMSE/\beta RMS_U$ and the radius is proportional to $RMSE$. The green area on the Target plot identifies the area of fulfilment of the MQO, i.e. MPI less than or equal to 1.

Because $CRMSE$ is always positive only the right hand side of the diagram would be needed in the Target plot. The negative X axis section can then be used to provide additional information. This information is obtained through relation (33) which is used to further investigate the $CRMSE$ related error and see whether it is dominated by R or by σ. The ratio of two $CRMSE$, one obtained assuming a perfect correlation ($R = 1$, numerator), the other assuming a perfect standard deviation ($\sigma_M = \sigma_O$, denominator) is calculated and serves as basis to decide on which side of the Target diagram the point will be located:

$$\frac{CRMSE(R=1)}{CRMSE\ (\sigma_M=\sigma_{O)}} = \frac{|\sigma_M - \sigma_O|}{\sigma_O\sqrt{2(1-R)}} \left\{ \begin{matrix} > 1: \ \sigma\ dominates\ R: right \\ < 1: \ R\ dominates\ \sigma: left \end{matrix} \right\} \quad (34)$$

For ratios larger than 1 the σ error dominates and the station is represented on the right, whereas the reverse applies for values smaller than 1.

The MQI associated to the 90th percentile worst station is calculated (see previous section) and indicated in the upper left corner. It is meant to be used as the main indicator in the benchmarking procedure and should be less or equal to one. Below this main indicator, also the MQI when using yearly average data is provided. Some consistency problems are discussed in the Open issue Section 9.3. The measurement uncertainty parameters (α, β, U_r^{RV} and RV) used to produce the diagram are listed on the top right-hand side. In blue color, the resulting model uncertainty is calculated according to equation (22) and is provided as output information. If relevant, the value of the MQI obtained, if all data were to be yearly averaged, is also provided.

In addition to the information mentioned above the proposed Target diagram also provides the following information:

- A distinction between stations according to whether their error is dominated by bias (either negative or positive), by correlation or standard deviation. The sectors where each of these dominates are delineated on the Target diagram by the diagonals in Figure 2.
- Identification of performances for single stations or group of stations by the use of different symbols and colours.

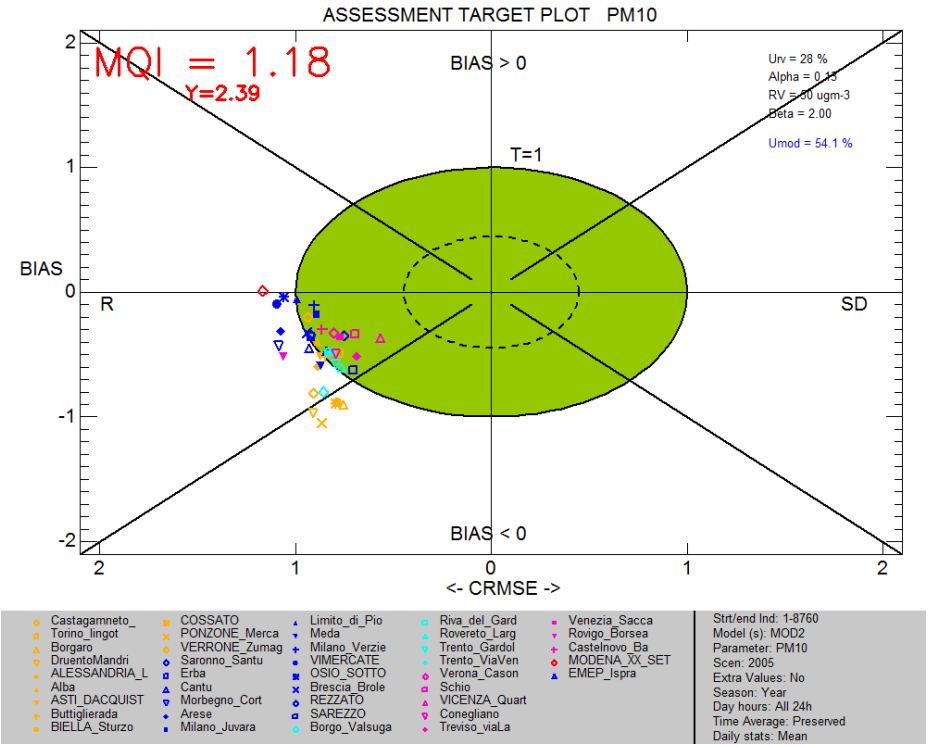


Figure 2: Example of Target diagram to visualize the main aspects of model performance. Each symbol represents a single station.

8.1.2. Summary Report

The summary statistics table, Figure 3, provides additional information on model performances. It is meant as an **additional** and **complementary** source of information to the MQO (Target diagram) to identify model strengths and weaknesses. The summary report is structured as follows:

- ROWS 1-2 provide the measured observed yearly means calculated from the hourly values and the number of exceedances for the selected stations. In benchmarking mode, the threshold values for calculating the exceedances are set automatically to 50, 200 and 120 $\mu g/m^3$ for the daily PM_{10}, the hourly NO_2 and the 8h daily O_3 maximum, respectively. For other variables ($PM_{2.5}$, WS…) no exceedances are shown.
- ROWS 3-6 provide an overview of the temporal statistics for bias (row 3), correlation (row 4) and standard deviation (row 5) as well as information on the ability of the model to capture the highest range of concentration values (row 6). Each point represents a specific station. Values for these four parameters are estimated using equations (26) to (30). The points for stations for which the model performance criterion is fulfilled lie within the green and the orange shaded areas. If a point falls within the orange shaded area the error associated with the particular statistical indicator is dominant. Note again that fulfilment of the bias, correlation, standard deviation and high percentile related indicators does not guarantee that the overall MQO based on the MQI (or RMSE, visible in the Target diagram) is fulfilled.
- ROWS 7-8 provide an overview of spatial statistics for correlation and standard deviation. Average values over the selected time period are first calculated for each station and these values are then used to compute the averaged spatial correlation and standard deviation. As a result only one point representing the spatial correlation of all selected stations is plotted. Colour shading follows the same rules as for rows 3-5.

Note that for indicators in rows 3 to 8, values beyond the proposed scale will be represented by the station symbol being plotted in the middle of the dashed zone on the right/left side of the proposed scale. For all indicators, the second column with the coloured circle provides information on the number of stations fulfilling the performance criteria: the circle is coloured green if more than 90% of the stations fulfil the criterion and red if the number of stations is lower than 90%.

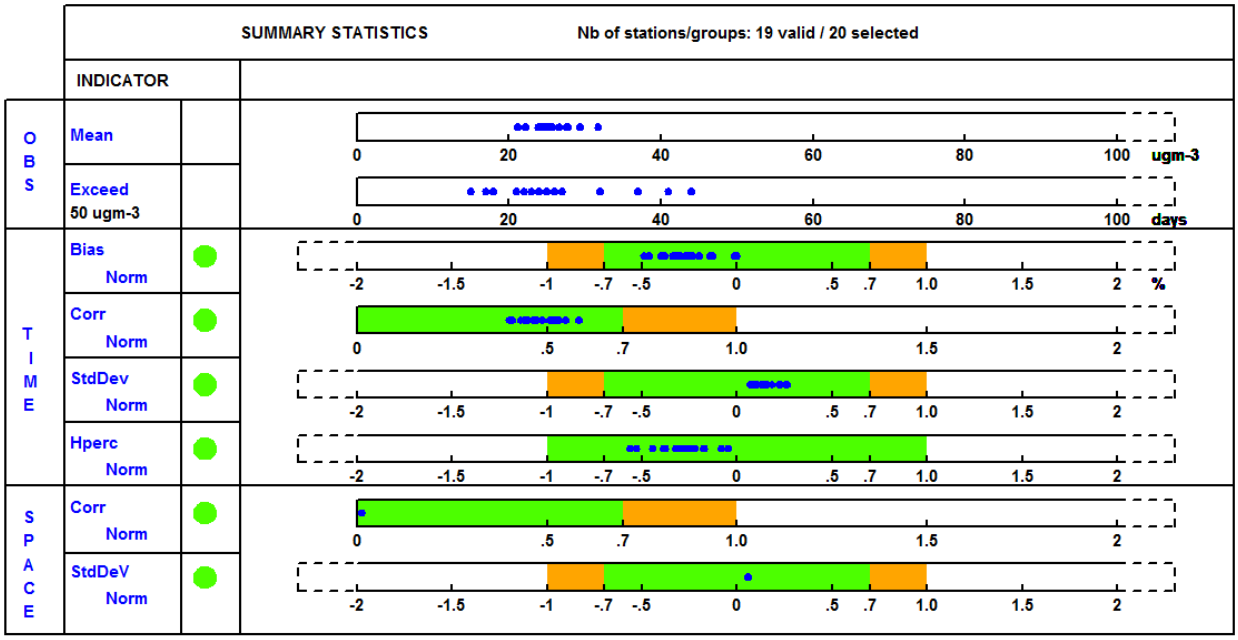


Figure 3: Example of a summary report based on hourly model results.

8.2. Yearly average data

For the evaluation and reporting of yearly averaged model results a Scatter diagram is used to represent the MQO instead of the Target plot because the CRMSE is zero for yearly averaged results so that the RMSE is equal to the BIAS in this case. The report then consists of a Scatter Diagram followed by the Summary Statistics (Figure 4).

8.2.1. Scatter Diagram

Equation (18) for yearly averaged results (i.e. based on the bias) is used as main model quality indicator. In the scatter plot, it is used to represent the distance from the 1:1 line. As mentioned above it is expected to be fulfilled (points are in the green area) by at least 90% of the available stations and a MQI value representative of the 90th percentile is calculated according to (19). The uncertainty parameters (α, β, U_r^{RV}, N_{np}, N_p and RV) used to produce the diagram are listed on the top right-hand side together with the associated model uncertainty calculated from (22).

The Scatter diagram also provides information on performances for single stations or group of stations (e.g. different geographical regions in this example below) by the use of symbols and colours. The names of the stations are given as legend below the scatterplot.

Summary Report

The summary statistics table provides additional information on the model performance. It is meant as an **additional** and **complementary** source of information to the bias-based MQI to identify model strengths and weaknesses. It is structured as follows:

- ROW 1 provides the measured observed means for the selected stations.

- ROW 2 provides information on the fulfilment of the bias-based MQO for each selected stations. Note that this information is redundant as it is already available from the scatter diagram but this was kept so that the summary report can be used independently of the scatter diagram.
- ROWS 3-4 provide an overview of spatial statistics for correlation and standard deviation. Annual values are used to calculate the spatial correlation and standard deviation. Equations (29) and (30) are used to check fulfilment of the performance criteria. The green and the orange shaded area represent the area where the model performance criterion is fulfilled. If the point is in the orange shaded area the error associated to the particular statistical indicator is dominant.

Note that for the indicators in rows 2 to 4, values beyond the proposed scale will be represented by plotting the station symbol in the middle of the dashed zone on the right/left side of the proposed scale.

The second column with the coloured circle provides information on the number of stations fulfilling the performance criteria: a green circle indicates that more than 90% of the stations fulfil the performance criterion while a red circle is used when this is less than 90% of the stations.

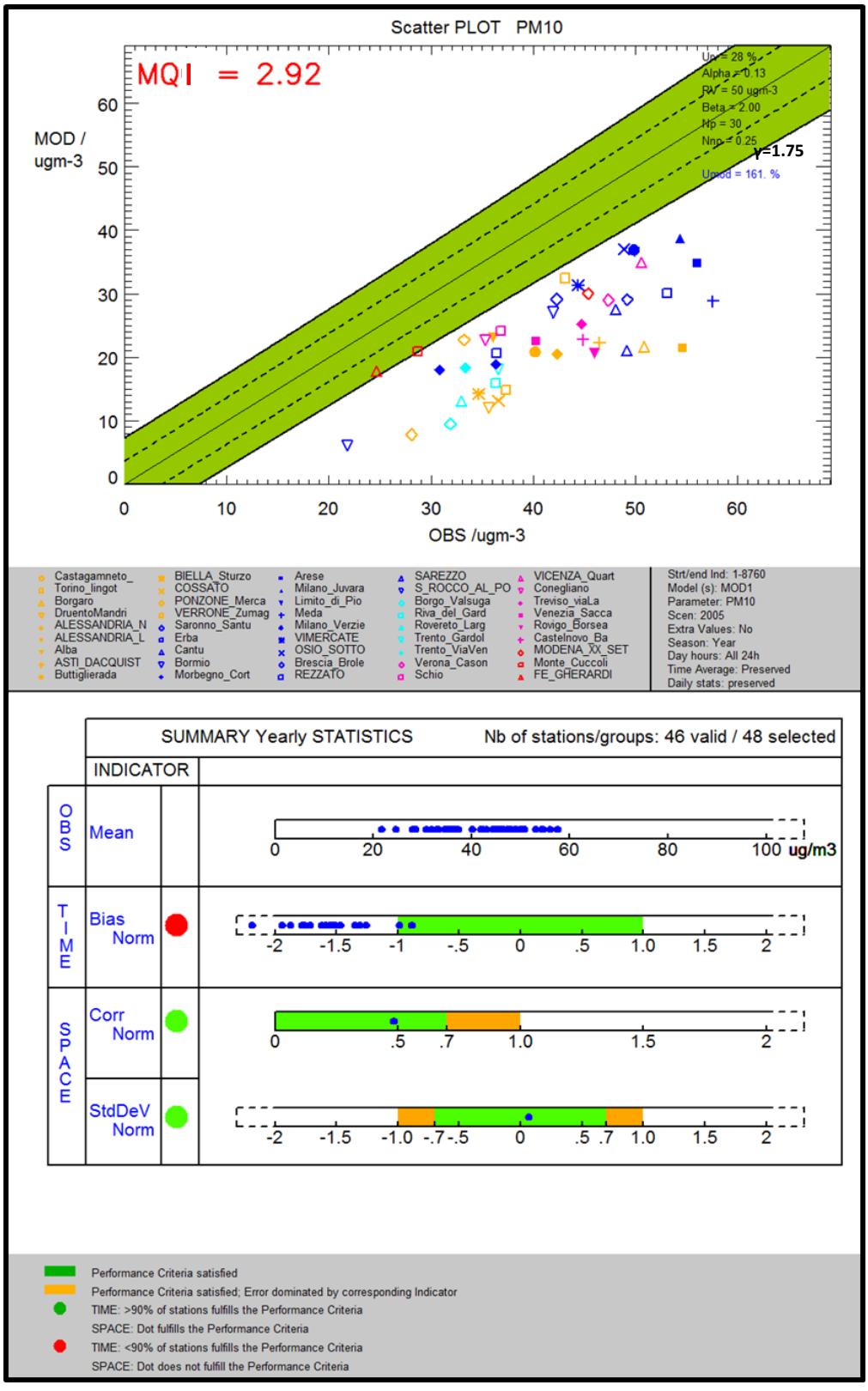


Figure 4: Example of main diagram (scatter) and a summary report based on yearly average model results.

9. OPEN ISSUES

In this section all topics are introduced on which there is no consensus yet within the FAIRMODE community and which merit further consideration.

9.1. Station representativeness

In the current approach only the uncertainty related to the measurement device is accounted for but another source of divergence between model results and measurements is linked to the lack of spatial representativeness of a given measurement station (or to the mismatch between the model grid resolution and the station representativeness). Although objectives regarding the spatial representativeness of monitoring stations are set in the AQD these are not always fulfilled in real world conditions. The formulation proposed for the MQI and MPI could be extended to account for the lack of spatial representativeness if quantitative information on the effect of station (type) representativeness on measurement uncertainty becomes available.

Current status: In order to clarify ambiguities in the definition, interpretation and assessment methodologies of spatial representativeness, an inter-comparison exercise is currently being setup within FAIRMODE. Different teams were encouraged to apply their preferred methodology and interpretation framework in a common case study. First individual results are already available and a first overall analysis is expected in the course of 2017.

9.2. Performance criteria for high percentile values

The MQI and MPI described in this document provide insight on the quality of the model average performances but do not inform on the model capability to reproduce extreme events (e.g. exceedances). For this purpose, a specific MPI indicator is proposed as:

$$MPI_{perc} = \frac{|M_{perc} - O_{perc}|}{\beta U_{95}(O_{perc})} \quad and \quad MPC: MPI_{perc} \leq 1 \tag{35}$$

where "perc" is a selected percentile value and M_{perc} and O_{perc} are the modelled and observed values corresponding to this selected percentile. The denominator, $U(O_{perc})$ is directly given as a function of the measurement uncertainty characterizing the O_{perc} value. For pollutants for which exceedance limit values exist in the legislation this percentile is chosen according to legislation. For hourly NO_2 this is the 99.8% (19th occurrence in 8760 hours), for the 8h daily maximum O_3 92.9% (26th occurrence in 365 days) and for daily PM_{10} and $PM_{2.5}$ 90.4% (36th occurrence in 365 days). For general application, when *e.g.* there is no specific limit value for the number of exceedances defined in legislation, the 95% percentile is proposed. To calculate the percentile uncertainty used in the calculation of MQI_{perc} the Eq. (10) is used with $O_i = O_{perc}$. MPI_{perc} has been included in the DELTA tool from version 5.0 on.

Current status: Apart from the extension of the MQI for percentiles as described above, a specific evaluation of model performance for episodes is now implemented in the test/expert version of DELTA vs5.3 and higher. The threshold evaluation criteria as implemented for forecast models (see further in Section 10) can also be used for the evaluation of episodes.

Jan Horalek (ETC/ACM) remarks that the uncertainty $U(O_{perc})$ is probably too large. Can we use a similar uncertainty for daily and percentile values? It could be assumed that $U(O_{perc})$ should be smaller than a U(O) of a daily value. To be further discussed…

9.3. Hourly/daily versus annual MQI

At the FAIRMODE Technical Meeting in Aveiro (June 2015) it was put forward by a number of modelling teams that there might be an inconsistency between the hourly/daily MQI and the annual MQI. Some model applications seem to pass the hourly/daily objective but apparently fail to meet the criteria when annual averaged values of the time series are used in the annual MQI procedure.

Further reflection on the topic made clear that the inconsistency between the two approaches is rather fundamental. The inconsistency is related to the auto-correlation in both the monitoring data and the model results and the way those auto-correlations are affecting the uncertainty of the annual averaged values. It became clear that a straightforward solution for the problem is not at hand. More in-depth information is available in a Working Note on the FAIRMODE website.

Current status: As a pragmatic solution to the above mentioned problem, it is suggested that model applications with hourly/daily output should also comply with the annual MQO. In DELTA vs5.3 and higher both criteria are implemented in one graph.

9.4. Data availability

Currently a value of 75% is required in the benchmarking both for the period considered as a whole and when time averaging operations are performed for all pollutants.

The Data Quality Objectives in Annex I of the AQD require a minimum measurement data capture of 90% for sulphur and nitrogen oxides, particulate matter (PM), CO and ozone. For ozone this is relaxed to 75% in winter time. For benzene the Directive specifies a 90 % data capture (dc) and 35% time coverage (tc) for urban and traffic stations and 90% tc for industrial sites. The 2004 Directive in Annex IV requires 90% dc for As, Cd and Ni and 50% tc and for BaP 90 % dc of 33% tc.

As these requirements for minimum data capture and time coverage do not include losses of data due to the regular calibration or the normal maintenance of the instrumentation the minimum data capture requirements are in accordance with the Commission implementing decision of 12 December 2011 laying down rules for the AQD reduced by an additional 5%. In case of e.g. PM this further reduces the data capture to 85% instead of 90%.

In addition, in Annex XI the AQD provides criteria for checking validity when aggregating data and calculating statistical parameters. When calculating hourly averages, eight hourly averages and daily

averages based on hourly values or eight hourly averages, the requested percentage of available data is set to 75%. For example a daily average will only be calculated if data for 18 hours are available. Similarly O_3 daily maximum eight hourly average can only be calculated if 18 eight hourly values are available each of which requires 6 hourly values to be available. This 75% availability is also required from the paired modelled and observed values. For yearly averages Annex XI of the AQD requires 90 % of the one hour values or - if these are not available - 24-hour values over the year to be available. As this requirement again does not account for data losses due to regular calibration or normal maintenance, the 90% should in line with the implementing decision above again further be reduced by 5% to 85%.

In the assessment work presented in the EEA air quality in Europe reports we can find other criteria. There, we find the criteria of 75% of valid data for PM_{10}, $PM_{2.5}$, NO_2, SO_2, O_3, and CO, 50% for benzene and 14 % for BaP, Ni, As, Pb, and Cd. In these cases you also have to assure that the measurement data is evenly and randomly distributed across the year and week days.

Current status: The 75% criteria was a pragmatic choice when the methodology was elaborated. It can be questioned if this is still a valid choice. A higher value for this criterion will limit the number of stations available for the evaluation whereas a smaller criteria value leads to truncated comparisons (only a small fraction of the year is indeed evaluated in these cases).

9.5. Data assimilation

The AQD suggests the integrated use of modelling techniques and measurements to provide suitable information about the spatial and temporal distribution of pollutant concentrations. When it comes to validating these integrated data, different approaches can be found in literature that are based on dividing the set of measurement data into two groups, one for the data assimilation or data fusion and one for the evaluation of the integrated fields. The challenge is how to select the set of validation stations.

It has been investigated within FAIRMODE's *Cross Cutting Activity Modelling & Measurements* which of the methodologies should be more robust and appropriate in operational contexts. In particular the methodology proposed by University of Brescia, based on a Monte Carlo analysis, has been tested by three groups (INERIS, U. Aveiro and VITO) and was shown to have little added value compared to the "leaving one out" approach. In addition it involves much more effort and is more time consuming. Nevertheless, additional tests are planned (by University of Aveiro) to further test the sensitivity and added value of the Monte-Carlo approach suggested.

For the time being, FAIRMODE recommends the "leaving one out" validation strategy as a methodology for the evaluation of data assimilation or data fusion results. It has to be noticed that such an approach might not be appropriate for on-line data assimilation methodologies (4D VAR, Ensemble Kalman Filter,…) due to computational constrains. In such cases, an a priori selection of assimilation and validation stations has to be made. However, the modeller should be aware of the fact that this a priori selection of validation stations will have an impact on the final result of the evaluation of the model application.

9.6. How does the 90% principle modifies the statistical interpretation of the MQO?

The requirement that the MQO and MPC should be fulfilled in at least 90% of the observation stations has some implications for the Modelling Quality Indicator (MQI).

The MQI is computed using expanded uncertainties U = k.u_c that are estimated on the basis of coverage factor k of 2 (JCGM 200, 2008) associated with level of confidence of 95 % that the true quantity values lay within the interval y±U. If the level of confidence is changed to 90%, the proper value of the coverage factor should be k=1.64 assuming that all measurements and their combined uncertainty u_c exhibit a Gaussian probability distribution. This assumption can hold for measurement uncertainties estimated with large degree of freedom as for example annual averages. From the expression of the MQI (see below), we see that the change of coverage factor will impact the stringency of the MQI (a lower value of k will increase the MQI)

$$MQI = \frac{1}{\beta}\frac{\mathrm{RMSE}}{k u_{95,r}^{RV}\sqrt{(1-\alpha^2)(\bar{O}^2 + \sigma_o^2) + \alpha^2 . RV^2}}$$

We can also note that a decrease of the coverage factor (e.g. from 2 to 1.64) when the level of confidence change for 95 to 90 % has a similar impact on the MQI than an increase of the β parameter (e.g. from 1.75 to 2.25 for PM_{10}).

9.7. Model benchmarking and evaluation for zones with few monitoring data

During the 2016 Plenary Meeting it was put forward by a number of participants that the FAIRMODE Modelling Quality Objective might not be fully applicable for urban scale applications. In many cases, only a limited set of monitoring stations is available in a single city or town. When less than 10 stations are at hand, the 90% criteria for the Target value requires that a model application fulfil the MQO criteria in all available stations, reducing the level of tolerance which is available for regional applications. In the new formulation of the MQO which implicitly takes into account the $90^{\%}$ principle (§7.3.3) accounts for this shortcoming to a certain level. It still remains questionable what a minimum number of stations should be to evaluate a modelling application for a specific (urban) region.

In addition, it is noticed that at the urban scale additional auxiliary monitoring data sets might be available (e.g. passive sampling data, mobile or temporary campaigns). Those monitoring data might be very valuable to check the quality of the urban applications but at present the FAIRMODE Benchmarking procedure is not capable to deal with those observations.

9.8. Application of the procedure to other parameters

Currently only PM, O_3 and NO_2 have been considered but the methodology could be extended to other pollutants such as heavy metals and polyaromatic hydrocarbons which are considered in the Ambient Air Quality Directive 2004/107/EC.

The focus in this document is clearly on applications related to the AQD and thus those pollutants and temporal scales relevant to the AQD. However the procedure can of course be extended to other variables including meteorological data as proposed in Pernigotti et al. (2014)

Current status: In the table below values are proposed for the parameters in (12) for wind speed and temperature data.

Table 6: List of the parameters used to calculate the uncertainty for the variables wind speed (WS) and temperature (TEMP)

	γ	U_r^{RV}	RV	A	N_p	N_{np}
WS (test)	1.75	0.260	5 m/s	0.89	NA	NA
TEMP (test)	1.75	0.05	25 K	1.00	NA	NA

When performing validation using the DELTA Tool, it is helpful to look at both NO_x as well as NO_2, as the former pollutant is less influenced by chemistry, and is therefore a better measure of the models' ability to represent dispersion processes. The NO_x measurement uncertainty is not available but could be approximated by the NO_2 uncertainty for now.

10. FORECASTING & EXCEEDANCES INDICATORS

10.1. Introduction

In this chapter, indicators and diagrams are proposed for the evaluation of model results in forecast mode. The main objective is to offer a common standardized template to facilitate the screening and comparison of model results. **It has to be stressed that this methodology is not as mature as the Modelling Quality Objective for assessment and requires further testing and fine tuning**.

10.2. Brief overview of the indicators

As we are mostly interested in forecast mode to check the model ability to accurately reproduce daily forecasts, the Target indicator is modified to consider these aspects. In place of using a normalization based on the standard deviation of the observation uncertainty, we normalize by a quantity representative of the "intra-day" variations, i.e.

$$\mathrm{Target}_{\,\mathrm{forecast}} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i^* - O_i\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_{i-j} - O_i\right)^2}} \qquad (36)$$

where N is the number of days included in the time series. In such case the Target indicator becomes one when the model forecast is as good as a persistent model (i.e. a model, which for day "i" forecasts the day "i-j" value). The indice "j" represents the forecast time length and is expressed in hours or days . M^* represent the modelled forecast values after accounting for uncertainty (see next section). Values lower than one indicates better capabilities than the persistent model whereas values larger than one indicate poorer performances.

10.3. Measurement uncertainty

The measurement uncertainty (U) is introduced to allow for some margin of tolerance. In DELTA it is now a user input parameter which can be used to test the sensitivity of the results to uncertainty. However, during the Technical Meeting in Zagreb (June 2016) it was argued that the measurement uncertainty should eventually take the same values as used in the MQI for assessment.

To account for this uncertainty each model results M_t at time *t* is transformed into a new model result M_t^* as follows :

$$\text{if } M_t < O_t \text{ then } M_t^* = \min(M_t + U * O_t, O_t)$$
$$\text{if } M_t \geq O_t \text{ then } M_t^* = \max(M_t - U * O_t, O_t)$$

This uncertainty will affect the calculation of the target indicator (36) but also the counting of:

- False Alarms (FA): Model values are above the limit value (LV) but not the observations.
- Missed Alarms (MA): Model values are below the LV but observed values are above it.
- Good values below LV (GA-): both model and observation are below the LV.
- Good values above LV (GA+): both model and observations are above the LV.

Three options to include the measurement uncertainty into the calculations are proposed: (1) Conservative, (2) Cautious and (3) as model (Note point 2). These are described in the below table.

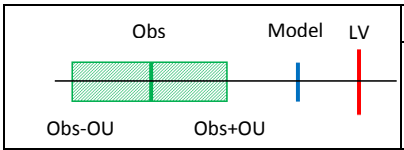
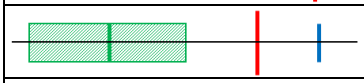
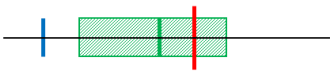
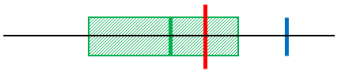
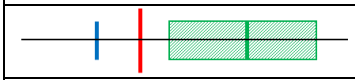
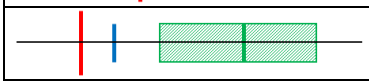
	**Observations**		**Model (M*)**		**DELTA**
	relation to LV	**Alarm?**	**relation to LV**	**Alarm?**	
	O_+<LV	No	M*<LV	No	GA-
	O_+<LV	No	M*≥LV	Yes	FA
	O_-<LV O_+≥LV	1: Yes, conservative 2: No, cautious 3: Same as model	M*<LV	No	MA GA- GA-
	O_-<LV O_+≥LV	1: Yes, conservative 2: No, cautious 3: Same as model	M*≥LV	Yes	GA+ FA GA+
	O_-≥LV	Yes	M*<LV	No	MA
	O_-≥LV	Yes	M*≥LV	Yes	GA+

Table 7: Possible cases with respect with model, observation and associated uncertainty. Please note that some "<" or ">" signs from the Note table have been changed to "≤" or "≥" to make sure all situations are included. The DELTA column indicates how DELTA considers the specific cases here described.

The measurement uncertainty can either be introduced as a percentage (from 1 to 100%) constant over the whole range of concentration or be as proposed in the frame of the assessment target approach (in this case the uncertainty varies with the concentration). This variable uncertainty can be introduced in the target by selecting "999" for the uncertainty field (see user's guidance below).

10.4. Diagrams in the delta tool

10.4.1. Target diagram

In the **Forecast target diagram** information on the following quantities (all normalised by the root mean squared intra-day variations):

- *Target forecast (RMSE)*: distance from the origin to the point (if distance inferior to one, the model behaves better than the persistent approach)
- *BIAS*: the bias can be either positive or negative and is represented along the vertical axis (Y)
- *CRMSE:* The CRMSE is always positive and given by the distance from the origin to the point along the X axis.
- *(False Alarm (FA) vs. Missed Alarm (MA)):* CRMSE is still the X axis but we use the FA/MA ratio to differentiate the negative and positive portions of the X axis. This ratio is used to differentiate the right and left parts of the target diagram:

$$\text{If} \quad \frac{FA}{MA} \leq 1 \Rightarrow \text{Left}$$

$$\text{If} \quad \frac{FA}{MA} > 1 \Rightarrow \text{Right}$$

Points are also given a characteristic colour which depends on the following ratio:

$$\frac{FA}{FA + GA_{+}} < 0.2 \Rightarrow \text{Dark green}$$

$$0.2 \leq \frac{FA}{FA + GA_{+}} < 0.4 \Rightarrow \text{Light green}$$

$$0.4 \leq \frac{FA}{FA + GA_{+}} < 0.6 \Rightarrow \text{Yellow}$$

$$0.6 \leq \frac{FA}{FA + GA_{+}} < 0.8 \Rightarrow \text{Orange}$$

$$0.8 \leq \frac{FA}{FA + GA_{+}} \Rightarrow \text{Red}$$

If more than one model forecast is used (e.g. D+1, D+2…) different symbols are used to differentiate the forecasts (see Figure 6).

Values lower than one (within the green circle) indicates better capabilities than the persistent model whereas values larger than one indicate poorer performances. An example of the forecast diagram is provided below for the case of a single model forecast (Figure 5) and in the case of two model forecasts compared to each other (Figure 6).

Note: Only stations characterized by at least one exceedance over the time period selected are considered in the analysis. Currently this filter on stations does not account for observation uncertainty (this could however be changed if necessary).

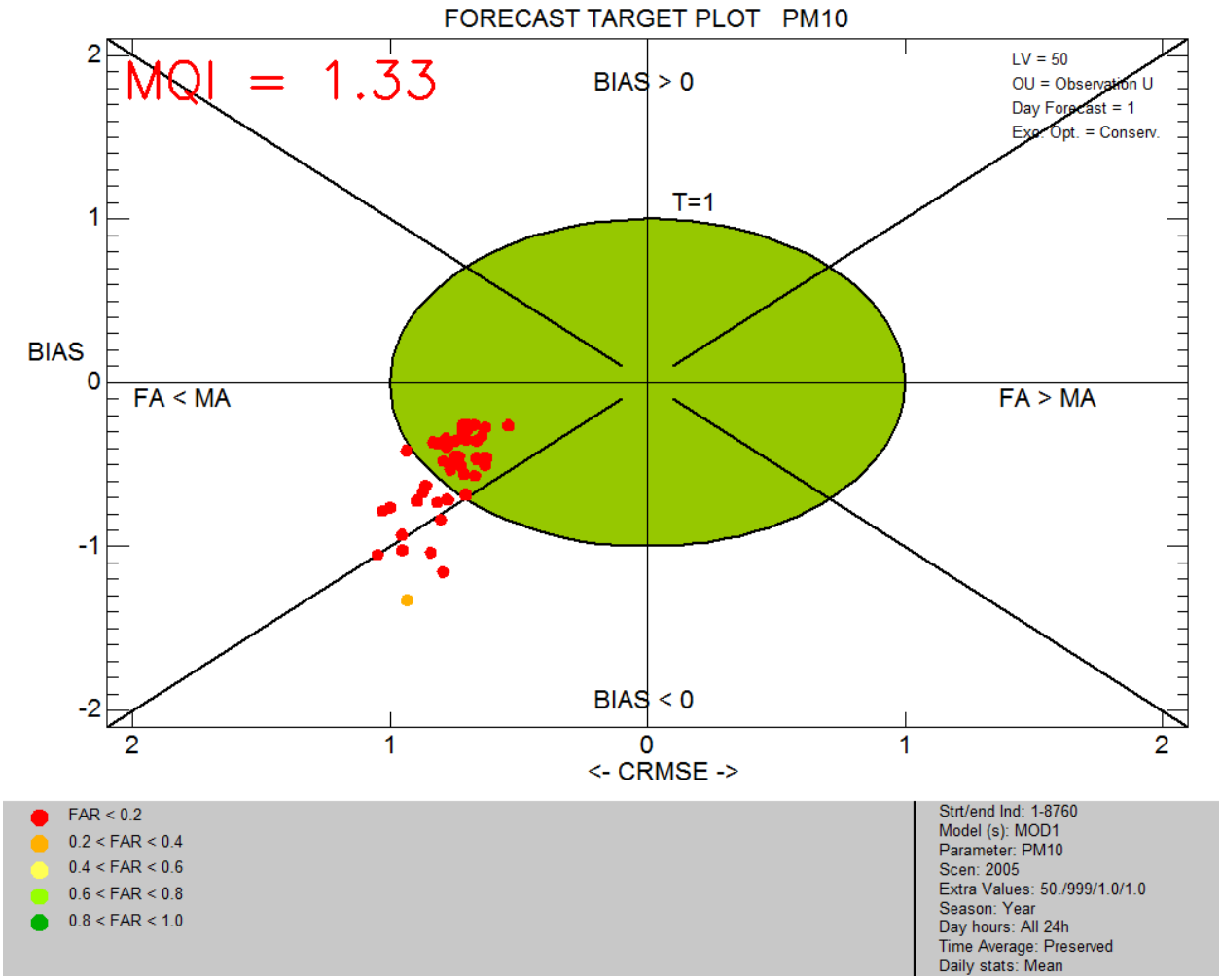


Figure 5: Target forecast for one model. The options selected (Limit Value, observation uncertainty, uncertainty flexibility option and forecast horizon) are reported in the right hand top corner of the figure. All symbols are similar and colours correspond to the value obtained for the FAR indicator.

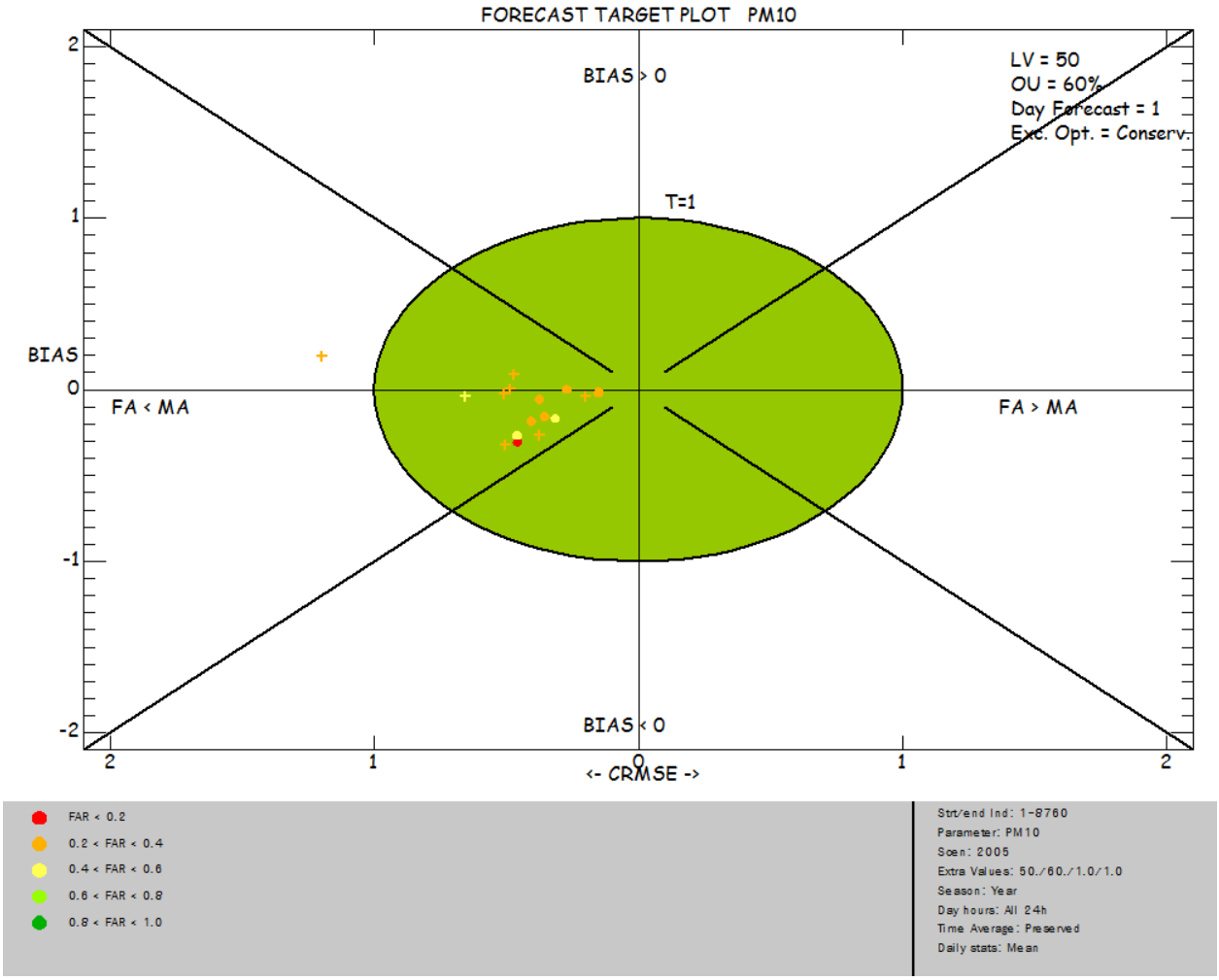


Figure 6: Same as Figure 5 but for two models. Different symbols (yet to be added in legend) are used to differentiate model results but colour still correspond to the value obtained for the FAR indicator

User's guidance: The target plot requires 4 values to be inserted as user's options under the threshold field (Analysis window interface). They will be introduced in the following formats: Val1#Val2#Val3#Val4. Some of these parameters will be hardcoded after the testing phase.

- *Val1*: Limit value
- *Val2*: Observation uncertainty – can be set as a fixed percentage (from 0 to 100%) or as the variable uncertainty used in the assessment target (use 999 as value for this option)
- *Val3*: Flexibility – Options to calculate the uncertainty (see table above). 1 for Conservative, 2 for cautious, 3 for same as model.
- *Val4*: Forecast horizon – time lag used for the comparison with the persistent model (see formula 1). The value is set in terms of a number of hours or days. Note that days will be selected only if the "Daily Stat" is not set to "preserve" in the Analysis window interface (i.e. either set to "Max", "Mean" or "Min"). Note that this value will be reset to "preserve" automatically if the diagram selection or the variable is changed.

Notes: The four user values will be reset each time a new diagram or a new variable is selected. Use CTRL-C & CTRL-V to copy/paste these values to avoid re-introducing them each time.

10.4.2. "Probability of detection" and "False alarm ratio" plots

Based on the following definitions:

- Probability of detection: DP=GA+/(MA+GA+) and
- False alarm ratio: FAR=FA/(FA+GA+),

two bar plots are created:

- The first (Figure 3) for the probability of detection plots GA+ as red dots and (MA+GA+) as grey column for each station. A good model capability would see all red dots on top of the column.
- The second (Figure 4) for false alarms is based on 1-FAR=GA+/(FA+GA+) in which again the red dots are for GA+ and the grey column for (FA+GA+). A good model again would see red dots close to the column tops.

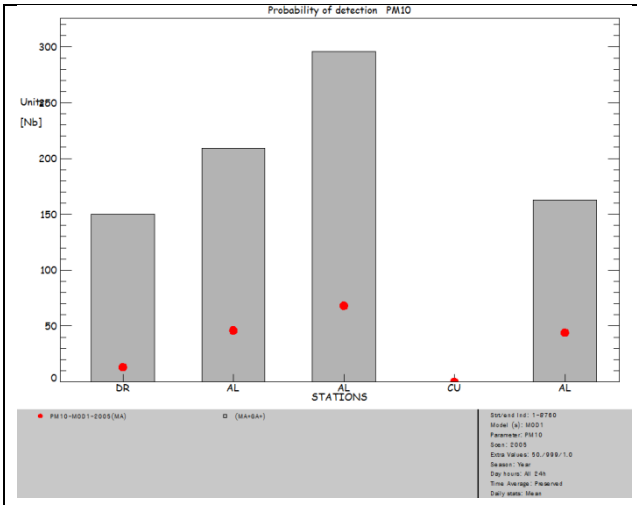


Figure 7: Probability of detection diagram: the dots (GA+) should be at top of grey columns (representing the total of observed alarms) for good model predictions (i.e. MA ≈ 0). The choices of limit value, uncertainty and flexibility options (conservative, cautious…) are listed in the lower right corner.

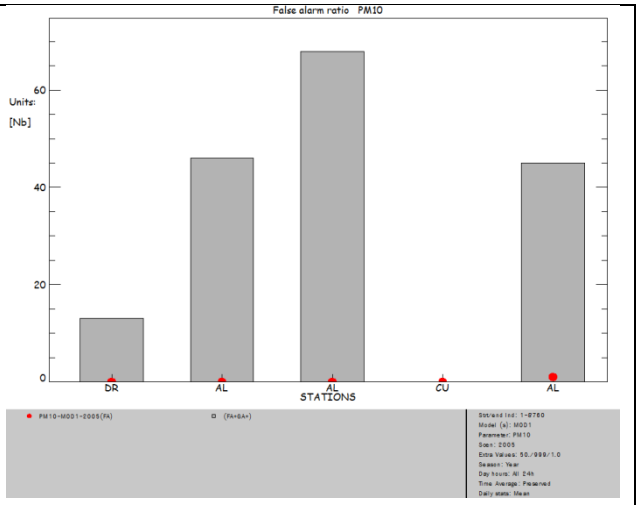


Figure 8: False alarm ratio diagram: the dots (FA) should be as close as possible from 0 (i.e. FA ≈ 0), with respect to the total of alarms predicted by the model (grey columns). The choices of limit value, uncertainty and flexibility options (conservative, cautious…) are listed in the lower right corner.

The values of FA, GA+, FA+GA+ and MA+GA+ are automatically saved in a "csv" file in the "dump" directory. The file will be over-written at each new diagram production.

10.4.3. "Exceedances indicator" bar plots

Indicator 1:

A barplot based on the delta user's guide high percentile MQO:

$$MPI_{perc} = \frac{|M_{perc} - O_{perc}|}{\beta U_{95}(O_{perc})} \quad and \quad MPC: MPI_{perc} \leq 1$$

is included in DELTA (named as "percentile indic PM, NO_2, O_3" in the barplot menu). One threshold value corresponding to the value of β is requested. This diagram is available for PM_{10}, $PM_{2.5}$, NO_2 and O_3.

Indicator 2:

It is defined as the summation combination of the previous two indicators to create a "composite exceedances indicator" as:

$$\text{CEI} \quad = 0.5(DP + 1 - FAR) = 0.5\left[\frac{GA_+}{MA + GA_+} + \frac{GA_+}{FA + GA_+}\right].$$

This indicator varies between 0 and 1 and does not allow for compensating effects between false and missed alarms. A perfect model forecast would lead to a value of 1. This ratio is available in the DELTA barplots choices under the label "exceedance ratio" (Figure 6).

The CEI indicators 2 & 3 defined as above will vary as follows:

Case	Obs Exceed.	Mod. Exceed	GA+	FA	MA	CEI_1	CEI_2
A	0	0	0	0	0	**Set to 1**	**Set to 1**
B	0	10	0	10	0	**Bound to 2**	**0**
C	10	0	0	0	10	**0**	**0**
D	10	10	10	0	0	**1**	**1**
E	10	10	5	5	5	**1**	**0.5**
F	10	5	5	0	5	**0.5**	**0.75**

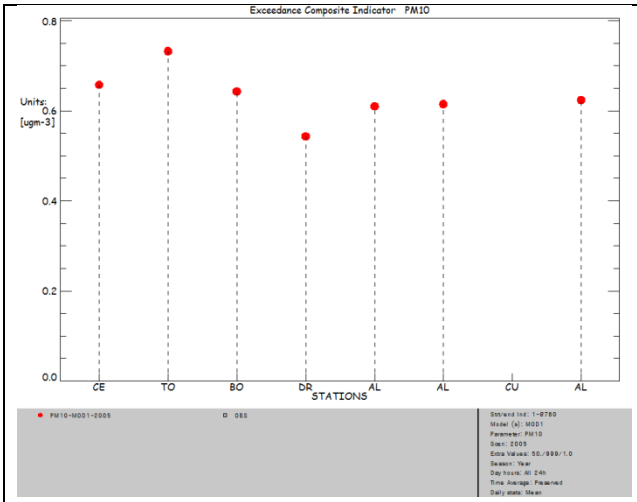


Figure 9: Composite exceedances indicator n1. The choices of limit value, uncertainty and flexibility options (conservative, cautious…) are listed in the lower right corner.

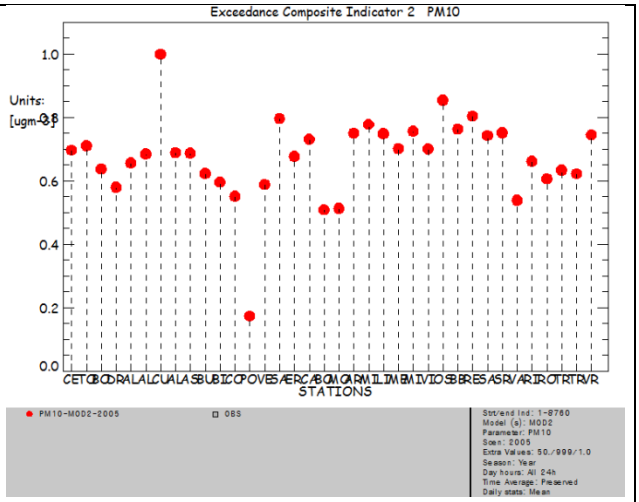


Figure 10: Composite exceedances indicator. The choices of limit value, uncertainty and flexibility options (conservative, cautious…) are listed in the lower right corner.

The values of FA, GA+, FA+GA+ and MA+GA+ are automatically saved in a “csv” file in the “dump” directory. The file will be over-written at each new diagram production.

User’s guidance:

All bar-plots require 3 values to be inserted as user’s options under the threshold field (Analysis window interface). They will be introduced in the following formats: Val1#Val2#Val3. These three values are the three first described in the Target Forecast diagram (see above). Some of these parameters will be hardcoded after the testing phase.

Note that the forecast target and barplots diagram are only available in the advanced DELTA version as this diagram is yet under development and intended for testing purposes only. To activate this functionality, go in the resource/init.ini file and uncomment the following line: ELAB_FILTER_TYPE=ADVANCED and comment this one: ELAB_FILTER_TYPE=STANDARD

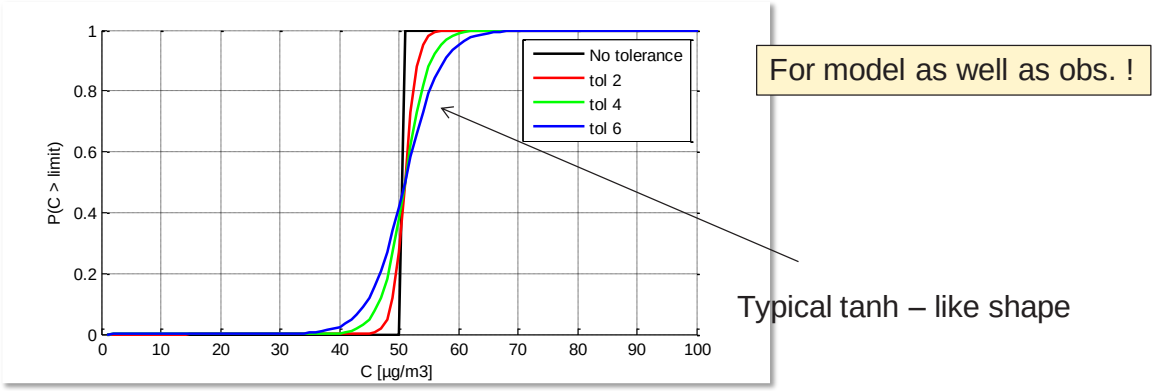
10.5. Remaining issues

This paragraph contains some points raised during the development of the methodology but for which NO IMPLEMENTATION HAS BEEN DONE in DELTA so far. Each of them needs further thinking, implementation, testing and fine tuning.

10.5.1. Probabilities

Currently, if we have a large amount of missing alerts FA/MA < 1 , it means that the model is underestimating the observations near the threshold, but introducing the OU will increase the model values, sometime to values above the threshold leading, so the ratio FA/MA would increase. Also, the opposite holds. So in the target diagram the points can switch from left to right, but in reality, you don’t really know since you are still within in the observation uncertainty. This behaviour makes it somewhat difficult to interpret the left/right labelling of the points in the target plot and relates back to the discussion above

What about introducing tolerance on the threshold? Here we don’t count hard exceedances, count probabilities i.e.:



The normal way of counting exceedances is you add 1 when the C value is e.g. above 50 µg/m3, if not you add 0. In the logic presented above, for a concentration of say 45 µg/m3, we would already add say 0.1 to the number of exceedances as there is a probability, given uncertainties that there were exceedances. This way of counting exceedances could be applied both for observations & for model values (representing observation & model uncertainty) and might be a more natural way for taking into account uncertainties in the exceedances indicators. To illustrate, say that theoretically both model & observations are 50 µg/m3 and the threshold is at 50 µg/m3 as well, there is for both model & observation a 50 % change that we have an exceedances. This would mean for FA,MA,GA+,GA- we would add in each of the classes 0.5*0.5 = 0.25 instead of picking one (which remains uncertain whether it's correct)

It is however unclear how this method would affect the target value/diagram, or how an increase in tolerance would affect the exceedances indicators such as FCF & FFA & this option would have to be further explored & perhaps presented by the next FAIRMODE meeting

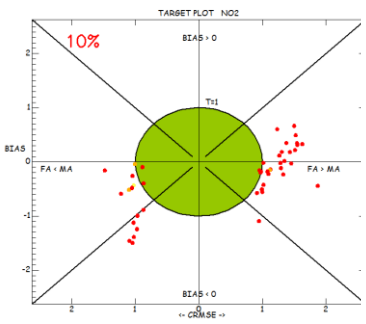
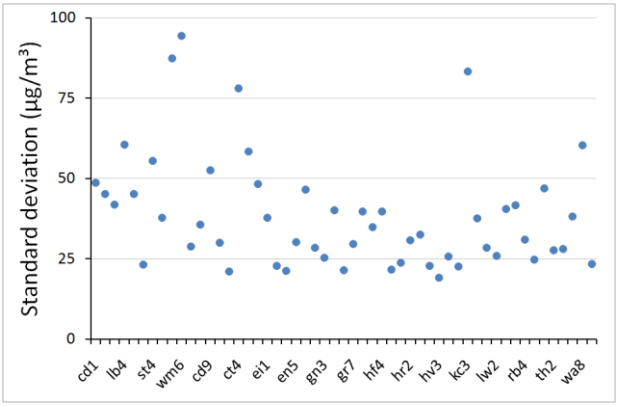
10.5.2. Standard deviation

If you had a period where the levels of pollution remained the same on a day by day basis (either constant, or varying diurnally), then

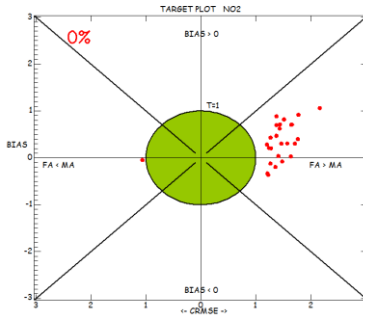
$$\frac{1}{N}\sum_{i=1}^{N}(O_{i-1}-O_i)^2 = 0$$

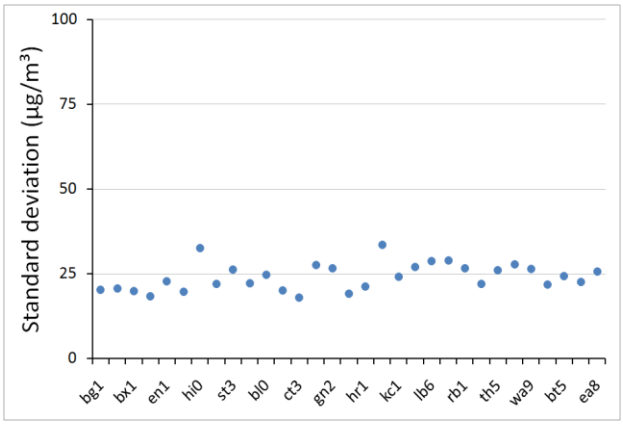
so the target → infinity. This means that at background sites where the standard deviation of the observations is relatively low compared to roadside sites, the target is harder to achieve. In fact what we are saying is that persistence is a better model at urban background sites than it is for roadside sites. Besides which, the persistence model fails mainly when there are peaks and exceedances to forecast, i.e. the main purpose of a forecasting system.

Examples of target and SD plots for roadside and background sites for London are shown below (note the different scales on the target plot):

	Target	**Standard deviation of the observations**
Roadside		

Urban background





Is it robust for the Target formulation to be independent on the standard deviation of the observations? The assessment target includes normalization by the standard deviation of the observations. One suggestion is to include:

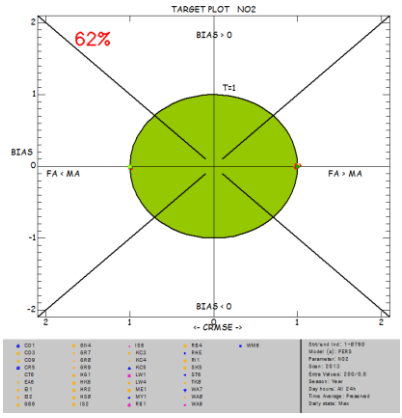
$$\sigma_o\sqrt{2(1-\alpha)}$$

where α is the autocorrelation in the observation time-series, which could be then adjusted for day+1, day+2 etc… [4]

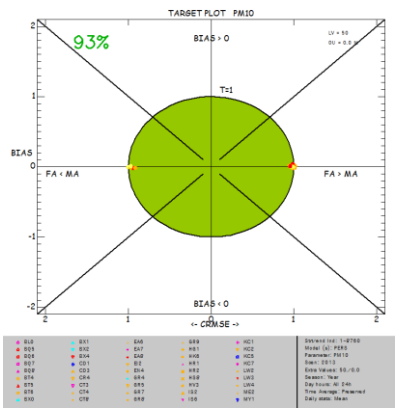
10.5.3. Persistence plots:

Running the tool with a persistence dataset and OU=0.0 gives approximately what you would expect – but it seems rounding errors are causing misleading stats, for example NO2 has 62% within the Target, PM10 has 93% yet O3 has 100% - as you would expect:

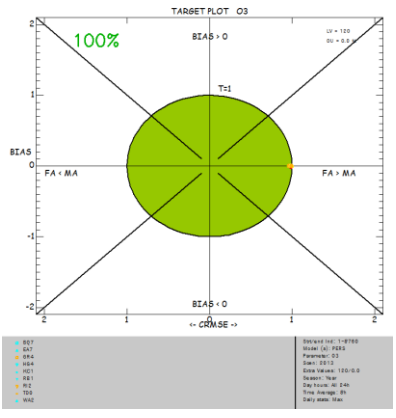
NO2 (max daily)



PM10 (daily max)



O3 (8-hour max)



[4] Bino Maiheu (VITO): I however don't think this would solve the problem mentioned above as it is simply a reformulation of the formula using O_{i-1} - O_i. But perhaps if one should replace the sigma_O buy the OU (wasn't this the strategy that was taking with the original target plot ?), this would solve the problem. Not sure how to interpret the target value then though…

Comment: This should be tested again with the latest implementation in DELTA. Note also that it is only implemented for 8h max O3, hourly NO2 and daily PM2.5 and PM10. The reason for that is the use of the observation uncertainty only available for these parameter-time average statistics.

10.5.4. Summary report:

Currently there are no forecast indicators in the summary report. Some suggestions for inclusion are the following:

Participant	Index agreement	Threshold agreement	GA+	FA	FCF	FFA
MP		✓				
JS	✓	✓			✓	✓
AM	✓	✓		✓		
BM					✓	✓

Comment: This still needs to be implemented in DELTA but a clear definition of the proposed indicators is needed. There does not seem to be consensus for all indicators.

10.5.5. X: Axis

Replace the ratio FA/MA along the x-axis by a new indicator combining the probability of good detection of threshold (FCF=GA+/(GA++ MA))

Comment: Does not seem to represent a consensus.

11. OVERVIEW OF EXISTING LITERATURE

11.1. Introduction

The development of the procedure for air quality model benchmarking in the context of the AQD has been an on-going activity in the context of the FAIRMODE community. The JRC developed the DELTA tool in which the Modelling Performance Criteria (MPC) and Modelling Quality Objective (MQO) are implemented. Other implementations of the MPC and MQO are found in the CERC Myair toolkit and the on-line ATMOSYS Model Evaluation tool developed by VITO.

In the following paragraphs a chronological overview is given of the different articles and documents that have led to the current form of the MQO and MPC. Starting from a definition of the MPC and MQO in which the measurement uncertainty is assumed constant (Thunis *et al.*, 2012) this is further refined with more realistic estimates of the uncertainty for O_3 (Thunis *et al.*, 2013) and NO_x and PM_{10} (Pernigotti *et al.*, 2013). The DELTA tool itself and an application of this tool are respectively described in Thunis *et al.*, 2013, Carnevale *et al.*, 2013 and Carnevale *et al.*, 2014, , Georgieva et al., 2015. Full references to these articles can be found at the end of this document.

11.2. Literature on how MQO and MPC are defined.

Thunis et al., 2012: Performance criteria to evaluate air quality modelling applications

This article introduces the methodology in which the root mean square error (RMSE) is proposed as the key statistical indicator for air quality model evaluation. A Modelling quality objective (MQO) and Model Performance Criteria (MPC) to investigate whether model results are 'good enough' for a given application are calculated based on the measurement uncertainty (U). The basic concept is to allow the same margin of tolerance (in terms of uncertainty) for air quality model results as for observations. As the objective of the article is to present the methodology and not to focus on the actual values obtained for the MQO and MPC, U is assumed to be independent of the concentration level and is set according to the data quality objective (DQO) value of the Air Quality Directive (respectively 15, 15 and 25% for O_3, NO_2 and PM_{10}). Existing composite diagrams are then adapted to visualize model performance in terms of the proposed MQO and MPC. More specifically a normalized version of the Target diagram, the scatter plot for the bias and two new diagrams to represent the standard deviation and the correlation performance are considered. The proposed diagrams are finally applied and tested on a real case

Thunis et al., 2013: Model quality objectives based on measurement uncertainty. Part I: Ozone

Whereas in Thunis *et al.*, 2012 the measurement uncertainty was assumed to remain constant regardless of the concentration level and based on the DQO, this assumption is dropped in this article. Thunis et al., 2013 propose a formulation to provide more realistic estimates of the

measurement uncertainty for O_3 accounting for dependencies on pollutant concentration. The article starts from the assumption that the combined measurement uncertainty can be decomposed into non-proportional (*i.e.* independent from the measured concentration) and proportional fractions which can be used in a linear expression that relates the uncertainty to known quantities specific to the measured concentration time series. To determine the slope and intercept of this linear expression, the different quantities contributing to the uncertainty are analysed according to the direct approach or GUM[5] methodology. This methodology considers the individual contributions to the measurement uncertainty for O_3 of the linear calibration, UV photometry, sampling losses and other sources. The standard uncertainty of all these input quantities is determined separately and these are subsequently combined according to the law of propagation of errors. AIRBASE data for 2009 have been used in obtaining the measurement uncertainty. Based on the new linear relationship for the uncertainty more accurate values for the MQO and MPC are calculated for O_3. MPC are provided for different types of stations (urban, rural, traffic) and for some geographical areas (Po Valle, Krakow, Paris).

Pernigotti et al., 2013: Model quality objectives based on measurement uncertainty. Part II: PM10 and NO2

The approach presented for O_3 in Thunis *et al.*, 2013 is in this paper applied to NO_2 and PM_{10} but using different techniques for the uncertainty estimation. For NO_2 which is not measured directly but is obtained as the difference between NO_x and NO, the GUM methodology is applied to NO and NO_x separately and the uncertainty for NO_2 is obtained by combining the uncertainties for NO and NO_x. For PM which is operationally defined as the mass of the suspended material collected on a filter and determined by gravimetry there are limitations to estimate the uncertainty with the GUM approach. Moreover, most of the monitoring network data are collected with methods differing from the reference one (e.g. automatic analysers), so-called equivalent methods. For these reasons the approach based on the guide for demonstration of equivalence (GDE) using parallel measurements is adopted to estimate the uncertainties related to the various PM_{10} measurements methods. These analyses result in the determination of linear expressions which can be used to derive the MQO and MPC. The Authors also generalise the methodology to provide uncertainty estimates for time-averaged concentrations (yearly NO_2 and PM_{10} averages) taking into account the reduction of the uncertainty due to error compensations during this time averaging.

Pernigotti et al., 2014: Modelling quality objectives in the framework of the FAIRMODE project: working document

This document corrects some errors found in the calculation of the NO_2 measurement uncertainty in Pernigotti *et al.*, 2013 and assesses the robustness of the corrected expression. In a second part, the validity of an assumption underlying the derivation of the yearly average NO_2 and PM_{10} MQO in which a linear relationship is assumed between the averaged concentration and the standard deviation is investigated. Finally, the document also presents an extension of the methodology for $PM_{2.5}$ and NOx and a preliminary attempt to also extend the methodology for wind and temperature.

[5] JCGM, 2008. Evaluation of Measurement Data - Guide to the Expression of Uncertainty in Measurement.

11.3. Literature on the implementation and use of the Delta tool

Thunis et al., 2012: A tool to evaluate air quality model performances in regulatory applications

The article presents the DELTA Tool and Benchmarking service for air quality modelling applications, developed within FAIRMODE by the Joint Research Centre of the European Commission in Ispra (Italy). The DELTA tool addresses model applications for the AQD, 2008 and is mainly intended for use on assessments. The DELTA tool is an IDL-based evaluation software and is structured around four main modules for respectively the input, configuration, analysis and output. The user can run DELTA either in exploration mode for which flexibility is allowed in the selection of time periods, statistical indicators and stations, or in benchmarking mode for which the evaluation is performed on one full year of modelling data with pre-selected statistical indicators and diagrams. The Authors also present and discuss some examples of DELTA tool outputs.

Carnevale et al., 2014: 1. Applying the Delta tool to support AQD: The validation of the TCAM chemical transport model

This paper presents an application of the DELTA evaluation tool V3.2 and test the skills of the chemical transport model TCAM model by looking at the results of a 1-year (2005) simulation at 6km × 6km resolution over the Po Valley. The modelled daily PM_{10} concentrations at surface level are compared to observations provided by approximately 50 stations distributed across the domain. The main statistical parameters (i.e., bias, root mean square error, correlation coefficient, standard deviation) as well as different types of diagrams (scatter plots, time series plots, Taylor and Target plots) are produced by the Authors. A representation of the measurement uncertainty in the Target plot, used to derive model performance criteria for the main statistical indicators, is presented and discussed.

Thunis and Cuvelier, 2016: DELTA Version 5.3 Concept / User's Guide / Diagrams

This is currently the most recent version of the user's guide for the DELTA tool. The document consists of three main parts: the concepts, the actual user's guide and an overview of the diagrams the tool can produce. The concepts part sets the application domain for the tool and lists the underlying ideas of the evaluation procedure highlighting that the tool can be used both for exploration and for benchmarking. The MQO and the MPCs that are applied are explained including a proposal for an alternative way to derive the linear expression relating uncertainty to measured concentrations. Examples of the model benchmarking report are presented for the cases model results are available hourly and as a yearly average. The actual user guide contains the information needed to install the tool, prepare input for the tool, and run the tool both in exploration and in benchmarking modes. Also details on how to customise certain settings (*e.g.* uncertainty) and how to use the included utility programs are given.

Carnevale et al., 2014: A methodology for the evaluation of re-analysed PM10 concentration fields: a case study over the Po valley

This study presents a general Monte Carlo based methodology for the validation of Chemical Transport Model (CTM) concentration re-analysed fields over a certain domain. A set of re-analyses is evaluated by applying the measurement uncertainty (U) approach, developed in the frame of

FAIRMODE. Modelled results from the Chemical Transport Model TCAM for the year 2005 are used as background values. The model simulation domain covers the Po valley with a 6 km x 6 km resolution. Measured data for both assimilation and evaluation are provided by approximately 50 monitoring stations distributed across the Po valley. The main statistical indicators (i.e. Bias, Root Mean Square Error, correlation coefficient, standard deviation) as well as different types of diagrams (scatter plots and Target plots) have been produced and visualized with the Delta evaluation Tool V3.6.

12. RELATED TOOLS

12.1. The DELTA Tool

The DELTA tool is an IDL evaluation software developed at EC-JRC, Ispra within the FAIRMODE activities. It was built upon the assets of the EuroDelta, CityDelta and POMI tools (Cuvelier et al., 2007, Thunis et al., 2007) and was designed for rapid diagnostics of air quality modelling applications under the EU Air Quality Directive 2008. The tool is based on pairs of measurement and modelled data at given location. It allows the user to perform two types of analysis: exploratory, looking at various statistical parameters, diagrams, pollutants and time intervals and benchmarking, when preselected model performance indicators for some regulated pollutants are compared to model quality objective and model performance criteria. The main concept of the methodology is based on taking into account the measurement uncertainty while calculating model performance indicators related to RMSE, Correlation, BIAS and standard deviation. The main model performance indicator, called modelling quality indicator (MQI), is expected to fulfil a criteria (the model quality objective) easily viewable at the target diagram, part of the benchmarking template.

The tool (current version 5.4, version 5.5 expected in March 2017) is available upon request via the FAIRMODE website.

12.2. The ATMOSYS benchmarking tool

The ATMOSYS (Policy support system for atmospheric pollution hotspots) system that was developed and evaluated in the context of a LIFE+ project (2010 – 2013) is an integrated Air Quality Management Dashboard that can be used for air pollution management and policy support in accordance with the 2008 EU CAFÉ Directive. ATMOSYS (http://www.atmosys.eu) offers different tools to support air pollution forecasting and assessment one of which is an air quality model benchmarking tool that is based on the methodology developed in the context of FAIRMODE. The tool allows the user to upload comma separated (csv) text files with hourly modelled and observed concentration values and use these to calculate the target plot and summary statistics (see chapter 8). The benchmarking functionality is currently limited to hourly values. As ATMOSYS is based on a generic web-based interface it can easily be adopted in other regions and in 2015 the ATMOSYS model benchmarking tool was updated and implemented as the model evaluation service for the French national air quality monitoring system (http://www.lcsqa.org/).

12.3. The MyAir Model Evaluation Toolkit

The MyAir Model Evaluation Toolkit has been designed to evaluate air quality models in terms of general performance. In addition, the MyAir Toolkit has specific features that assess the models' ability to calculate metrics associated with air quality forecasting, for example exceedances of daily limit values. The toolkit was developed as part of the GMES downstream service project,

PASODOBLE, which produced local-scale air quality services for Europe under the name 'Myair'. The MyAir Toolkit consists of four tools: a questionnaire tool offering structured advice on the advisability of the proposed evaluation; a data input tool able to import a wide range of modelled and in-situ monitored data formats; a model evaluation tool that analyses the performance of the model at predicting concentrations and pollution episodes; and a model diagnostics tool that compares modelled and monitored data at individual stations in more detail. The Myair Toolkit is easy to use, produces statistical data and attractive graphs, and has a comprehensive User Guide. The tool is downloadable from http://www.cerc.co.uk/MyAirToolkit and further information can be found in Stidworthy *et al.* (2013).

13. REFERENCES

13.1. Peer reviewed articles

Carnevale C., G. Finzi, A. Pederzoli, E.Pisoni, P. Thunis, E.Turrini, M.Volta (2014), Applying the Delta tool to support AQD: The validation of the TCAM chemical transport model, Air Quality, Atmosphere and Health , 10.1007/s11869-014-0240-4

Carnevale C., G. Finzi, A. Pederzoli, E. Pisoni, P. Thunis, E. Turrini, M. Volta. (2015), A methodology for the evaluation of re-analysed PM10 concentration fields: a case study over the Po Valley, Air quality Atmosphere and Health, 8, p.533-544

Georgieva, E., Syrakov, D., Prodanova, M., Etropolska, I. and Slavov, K. (2015), Evaluating the performance of WRF-CMAQ air quality modelling system in Bulgaria by means of the DELTA tool, Int. J. Environment and Pollution, 57, p.272–284

Lagler, F., Belis, C., Borowiak, A., 2011. A Quality Assurance and Control Program for PM2.5 and PM10 Measurements in European Air Quality Monitoring Networks, EUR - Scientific and Technical Research Reports No. JRC65176.

Pernigotti D., P. Thunis, C. Belis and M. Gerboles, (2013) Model quality objectives based on measurement uncertainty. Part II: PM10 and NO2, Atmospheric Environment, 79, p.869-878.

Stidworthy A., D. Carruthers, J. Stocker, D. Balis, E Katragkou, J Kukkonen, (2013), Myair toolkit for model evaluation, Proceedings of the 15th International Conference on Harmonisation, Madrid, Spain, 2013.

Thunis P., D. Pernigotti and M. Gerboles, (2013), Model quality objectives based on measurement uncertainty. Part I: Ozone, Atmospheric Environment, 79, p.861-868.

Thunis P., A. Pederzoli, D. Pernigotti (2012), Performance criteria to evaluate air quality modelling applications. Atmospheric Environment, 59, p.476-482

Thunis P., E. Georgieva, A. Pederzoli (2012), A tool to evaluate air quality model performances in regulatory applications, , Environmental Modelling & Software 38, p.220-230

13.2. Reports/ working documents / user manuals

P. Thunis, A. Pederzoli, D. Pernigotti (2012) FAIRMODE SG4 Report Model quality objectives Template performance report & DELTA updates, March 2012. http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_March2012.pdf

D. Pernigotti, P. Thunis and M. Gerboles (2014), Modeling quality objectives in the framework of the FAIRMODE project: working document http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Working%20note_MQO.pdf

P. Thunis, E. Georgieva, A. Pederzoli (2011), The DELTA tool and Benchmarking Report template Concepts and User guide Joint Research Centre, Ispra Version 2 04 April 2011 http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_April2011.pdf

P. Thunis, E. Georgieva, S. Galmarini (2011), A procedure for air quality models benchmarking Joint Research Centre, Ispra Version 2 16 February 2011 http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/WG2_SG4_benchmarking_V2.pdf

P. Thunis, C. Cuvelier, A. Pederzoli, E. Georgieva, D. Pernigotti, B. Degraeuwe (2016), DELTA Version 5.3 Concepts / User's Guide / Diagrams Joint Research Centre, Ispra, September 2016

ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparison.

JCGM 200, International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 2008.

13.3. Other documents/ e-mail

D.Brookes, J. Stedman, K. Vincent, B. Stacey, Ricardo-AEA, 18/06/14, Feedback on Model Quality Objective formulation

Mail correspondence between RIVM – The Netherlands (J. Wesseling) and JRC (P.Thunis)