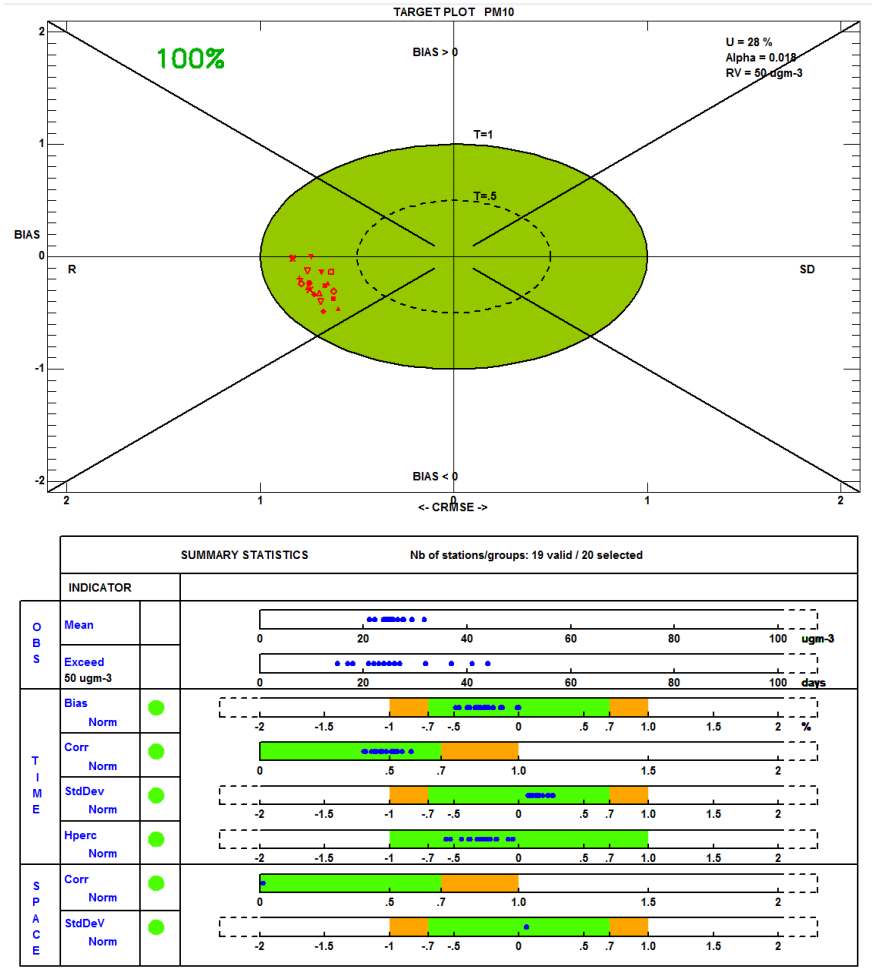
Guidance Document on Model Quality Objectives and Benchmarking

Peter Viaene, Stijn Janssen, Philippe Thunis, Elke Trimpeneers, Joost Wesseling, Alexandra Montero,, Ana Miranda, Jenny Stocker, Helge Rørdam Olesen, Cristina Guerreiro, Gabriela Sousa Santo, Keith Vincent, Claudio Carnevale, Michele Stortini, Giovanni Bonafè, Enrico Minguzzi and Marco Deserti

Version 1.0 - February, 2015



FAIRMODE
Forum for air quality modelling in Europe

Table of contents

1. INTRODUCTION

The objective of this guidance document is twofold:

1. to summarize the contents of different documents that have been produced in the context of FAIRMODE with the aim to define a methodology to evaluate air quality model performance for policy applications, especially related to the Ambient Air Quality Directive 2008/50/EC (AQD). Air quality models can have various applications (forecast, assessment, scenario analysis, …). The focus of this document is only on the use of air quality models for the assessment of air quality.
2. to present user feedback based on a number of examples in which this methodology has been applied.

2. BENCHMARKING: A WORD OF CAUTION

UNESCO[1] defines benchmarking 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 intercomparison 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 within the observation uncertainty of monitoring results. In many cases, this is a reasonable assumption. However, it is important to recognize that this is not always the case, so you should proceed with caution when you interpret benchmarking results. Here are two 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.

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 you wish to avoid drawing unwarranted conclusions from benchmarking results, then it is not sufficient to inspect benchmarking results. You should acquire some background information on the underlying data and consider the challenges they represent.

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.

[1] Vlăsceanu, 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

3. SCOPE AND FOCUS

The focus of this Guidance Document and the work performed within FAIRMODE is on producing a model quality objective (MQO) and model 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 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 at least of 75% for 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 O3 daily maximum 8-hour means is only calculated for the 8 hour periods in which 6 hourly values are available. In open issues (0) the choice of the data availability criterion is further elaborated.

2. Model performance criteria

The model performance criteria (MPC) are in this document only defined 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. MPC 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, by the limit value (or target value in the case of ozone), without taking into account the timing of the events. For benchmarking 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. As the number of stations is an integer value this means that sometimes more than 90% of the available stations will need to fulfil the criteria and for example in the specific case that there are less than 10 observation stations, all stations will need to fulfil the criteria. In the open issues (0) an alternative interpretation of the fulfilment criterion is presented.

4. OVERVIEW OF EXISTING LITERATURE

4.1. Introduction

The development of the procedure for air quality model benchmarking in the context of the AQD has been an ongoing activity in the context of the FAIRMODE[2] community that has been led by JRC. The JRC has also developed the DELTA tool in which the Model Performance Criteria (MPC) and Model 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 Model Performance Criteria and Model Quality Objective. 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. Full references to these articles can be found at the end of this document.

4.2. Literature on how these model performance criteria and model quality objectives are defined.

Thunis et al., 2011: A procedure for air quality model benchmarking

This document was produced in the context of the work done in the Subgroup 4 (SG4) of Working Group 2 (WG2) of FAIRMODE. The objective was to develop a procedure for the benchmarking of air quality models in order to evaluate their performances and indicate a way for improvements. The document first gives a global overview of the proposed approach by presenting the prerequisites, the four key elements envisioned (the DELTA tool, the ENSEMBLE tool, an online benchmarking service and an extraction facility) and a description of the procedure that focuses on how the different facilities could help in the model performance evaluation.

Some key concepts underlying the procedure are presented next: 1) the application domain which is the EU Air Quality Directive (AQD, 2008), 2) the need for input data consistency checks, 3) not only model to observation comparison but also model intercomparison and model response evaluation, 4) use of a limited set of model performance indicators that are assessed with respect to criteria and goals, 5) the aim of the procedure which is to provide model user with feedback and 6) the automatic reporting system of the benchmarking service.

[2] 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.

The final section of the article is devoted to the methodology for the benchmarking service, the different testing levels and the goals, criteria and the observation uncertainty considered in the evaluation as well as a proposal for the automatic report. The document concludes with a number of annexes on the application domain (pollutants and scales), the statistics and charts, the different spatial and temporal aggregations for model results and performance criteria and goals.

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. Model Performance Criteria (MPC) to investigate whether model results are 'good enough' for a given application are calculated based on the observation 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 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 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 proposes 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[3] 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. Based on the new linear relationship for the uncertainty more accurate values for the MQO and MPC are calculated for O_3.

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

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 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 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.

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

Thunis et al., 2013: 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 V3.2and test the skills of DELTA tool by looking at the results of a 1-year (2005) simulation performed using the chemical transport model TCAM 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 observation uncertainty in the Target plot, used to derive model performance criteria for the main statistical indicators, is presented and discussed.

Thunis et al., 2014: DELTA Version 4 User's Guide

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 observed 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 observation 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 kmx6 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.

5. MODEL QUALITY OBJECTIVE (MQO)

5.1. Statistical performance 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 performance 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 set of statistical indicators can be used 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:

Root Mean Square Error ($RMSE$ RMSE)

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2} \quad (1)$$

correlation coefficient (R)

$$R = \frac{\sum_{i=1}^{N}(M_i-\bar{M})(O_i-\bar{O})}{\sqrt{\sum_{i=1}^{N}(M_i-\bar{M})^2}\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2}} \quad (2)$$

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.

Normalised Mean Bias (NMB)

$$NMB = \frac{BIAS}{\bar{O}} \text{ where } BIAS = \bar{M} - \bar{O} \quad (3)$$

Normalised Mean Standard Deviation (NMSD)

$$NMSD = \frac{(\sigma_M - \sigma_O)}{\sigma_O} \quad (4)$$

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.

5.2. Model performance criteria (MPC) and formulation of the model quality objective (MQO)

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. This is the reason why Model Performance Criteria (MPC), defined as the minimum level of quality to be achieved by a model for policy use, are also needed.

To derive performance criteria for the selected statistical indicators we take into account the observation uncertainty. We define RMS_U as the quadratic mean of the measurement uncertainty:

$$RMS_U = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(U(O_i)\right)^2} \quad (5)$$

where $U(O_i)$ denotes the uncertainty for the *i*-th observed concentration level, O.

With the simple principle of allowing the same margin of tolerance to both model and observations we can define the Model Quality Objective (MQO) as:

$$MQO = \frac{1}{2}\frac{RMSE}{RMSU} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}}{2RMS_U} \leq 1 \quad (6)$$

With this formulation for the MQO the error between observed and modelled values (numerator) is compared to the absolute measured uncertainty (denominator). Three cases can then be distinguished:

1. $MQO \leq 0.5$: the model results are within the range of observation uncertainty (U) and it is not possible to assess whether further improvements to the model are closer to the true value;
2. $0.5 < MQO \leq 1$: $RMSE$ is larger than RMS_U, but model results could still be closer to the true value than the observation
3. $1 < MQO$: the observation and model uncertainty ranges do not overlap and model and observation are more than 2U apart. Observation is closer to the true value than the model value in this case.

This is illustrated in Figure 1 in which examples of these three different cases occur respectively on day 3, day 13 and day 10.

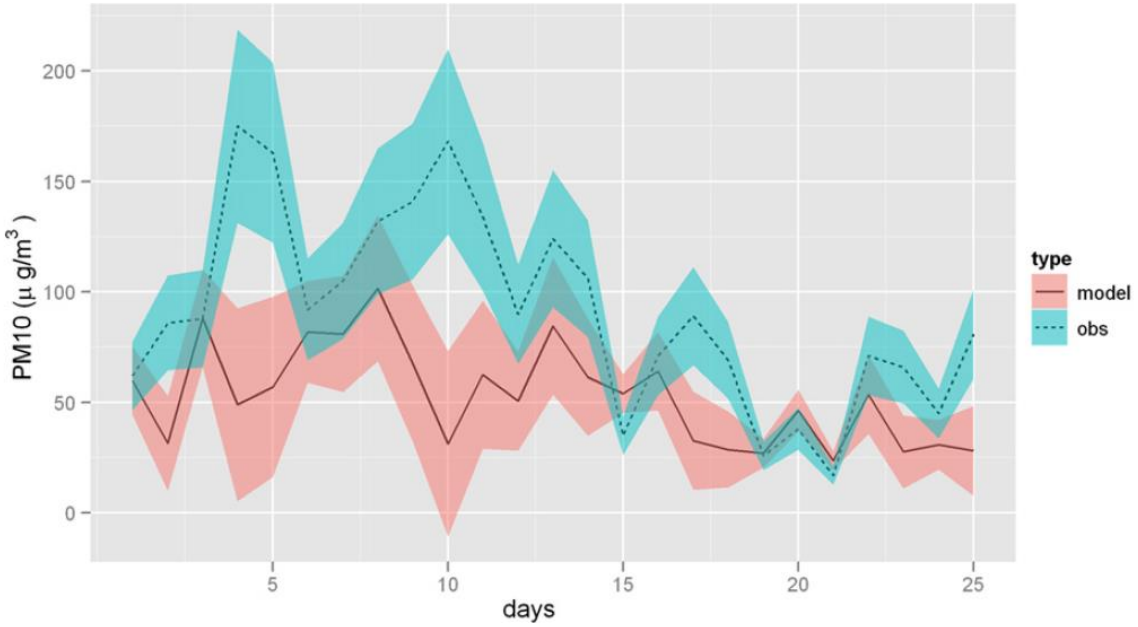


Figure 1 Example PM_{10} time series (measured and modelled concentrations) for a single station, together with a coloured area representative of the model and observed uncertainty ranges. (from Thunis *et al.*, 2012)

The proposed MQO has the advantage that it allows for introducing more detailed information on observation uncertainty when this becomes available.

For annual average values, the MQO simplifies to:

$$MQO = \frac{|BIAS|}{2RMS_U} \leq 1 \qquad (7)$$

5.3. Additional model performance criteria (MPC) for Bias, R and standard deviation

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

$$MQO^2 = \frac{RMSE^2}{(2RMS_U)^2} = \frac{BIAS^2}{(2RMS_U)^2} + \frac{(\sigma_M - \sigma_O)^2}{(2RMS_U)^2} + \frac{2\sigma_O\sigma_M(1-R)}{(2RMS_U)^2} \qquad (8)$$

By considering the ideal cases where $R = 1$, $\sigma_O = \sigma_M$ and $BIAS = 0$ separate MPC can be derived from (8) for each of these three statistics:

Statistic	Model Performance Criterion
BIAS $(R = 1, \sigma_O = \sigma_M)$	$\frac{\lvert BIAS \rvert}{2\,RMS_U} \leq 1$ (9)
R $(BIAS = 0, \sigma_O = \sigma_M)$	$\frac{(1-R)}{2}\frac{{\sigma_o}^2}{{RMS_U}^2} \leq 1$ (10)
Standard deviation $(BIAS = 0, R = 1)$	$\frac{\lvert \sigma_M - \sigma_O \rvert}{2\,RMS_U} \leq 1$ (11)

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

If one of the terms in equation 8 is larger than 0.5 the error type (BIAS, standard deviation or R) associated with this term will be predominant. This allows us to distinguish the following three cases:

Statistic	Model Performance Criterion	
BIAS	$0.5 < \frac{BIAS^2}{(2\,RMS_U)^2} \leq 1$	(12)
R	$0.5 < \frac{(1-R)}{2}\frac{\sigma_o\sigma_M}{{RMS_U}^2} \leq 1$	(13)
Standard deviation	$0.5 < \frac{(\sigma_M - \sigma_O)^2}{(2\,RMS_U)^2} \leq 1$	(14)

Finally, the MPC can also be derived for the individual statistics based on (8) for the case where the MQO ≤ 0.5 and thus the error between modelled and observed values lies within the measurement uncertainty range:

Statistic	Model Performance Criterion	
BIAS	$\frac{\lvert BIAS \rvert}{RMS_U} \leq \sqrt{2}$	(15)
R	$(1-R)\frac{\sigma_o\sigma_M}{{RMS_U}^2} \leq 1$	(16)
Standard deviation	$\frac{\lvert \sigma_M - \sigma_O \rvert}{RMS_U} \leq \sqrt{2}$	(17)

5.4. Observation uncertainty

5.4.1. General expression

In Thunis *et al.*, 2013 a general expression for the observation uncertainty is derived by considering that the combined uncertainty, $u_c(O_i)$ 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, $u_{np}(O_i)$:

$$u_c{}^2(O_i) = u_p{}^2(O_i) + u_{np}{}^2(O_i) \qquad (18)$$

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 estimated relative 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 (0-1) of the uncertainty at the reference value:

$$u_{np}{}^2(O_i) = \alpha(u_r{}^{RV}RV)^2 \qquad (19)$$

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

$$u_p{}^2(O_i) = (1-\alpha)\,(u_r{}^{RV} O_i)^2 \tag{20}$$

From the combined uncertainty, $u_c(O_i)$ an expanded uncertainty $U(O_i)$ can be estimated by multiplying with a coverage factor k:

$$U(O_i) = k u_c(O_i) \tag{21}$$

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 = 1.4, k = 2.0 and k = 2.6 correspond to confidence levels of around respectively 90, 95 and 99%.

Combining (18) – (21) the uncertainty of a single observation value can be expressed as:

$$U(O_i) = k u_c(O_i) = k u_r^{RV} \sqrt{(1-\alpha) O_i^2 + \alpha . RV^2} \tag{22}$$

From Equation (23) it is possible to derive an expression for RMS_U (equation 5) as:

$$RMS_U = \sqrt{\frac{\sum_{i=1}^{N} (U(O_i))^2}{N}} = k u_r^{RV} \sqrt{(1-\alpha)(\bar{O}^2 + \sigma_o^2) + \alpha . RV^2} \tag{23}$$

where $\bar{O}$ and σ_0 are respectively the mean and the standard deviation of the measured time series.

5.4.2. Derivation of parameters for the uncertainty

To be able to apply (24) it is necessary to estimate $u_r{}^{RV}$, the relative uncertainty around a reference value and α, the non-proportional fraction around the reference value. If we define the relative expanded uncertainty as $U_r^{RV} = k . u_r{}^{RV}$, equation 23 can be rewritten as

$$U(O_i)^2 = (U_r^{RV})^2 \left[(1-\alpha) O_i^2 + \alpha\, RV^2\right] = \alpha\, (U^{RV})^2 + \left(\frac{U^{RV}}{RV}\right)^2 (1-\alpha) O_i^2 \tag{24}$$

with $U^{RV} = RV . U_r^{RV}$ the absolute expanded uncertainty around the reference value, RV. This is a linear relationship with slope, $m = (1-\alpha)\left(\frac{U^{RV}}{RV}\right)^2$ and intercept, $q = \alpha\, (U^{RV})^2$ which can be used to derive values for U^{RV} and α by fitting measured squared uncertainties $U(O_i)^2$ to squared observed values $(O_i)^2$.

An alternative procedure for calculating U^{RV} and α can be derived by rewriting (25) as

$$U(O_i)^2 = (U^L)^2 + \frac{(U^{RV})^2 - (U^L)^2}{RV^2 - L^2} (O_i)^2 \tag{25}$$

where L is a low range concentration value (*i.e.* close to zero) and U^L its associated absolute expanded uncertainty. Comparing the two formulations we obtain:

$$\alpha = \left(\frac{U^L}{U^{RV}}\right)^2 \tag{26}$$

$$(U^L)^2 = (U^{RV})^2 - \left(\frac{U^{RV}}{RV}\right)^2 (1-\alpha)(RV^2 - L^2) \tag{27}$$

The two above relations (26) and (27) allow switching from one formulation to the other. The first formulation (24) requires defining values for both α and $U_r^{RV} = k.u_r{}^{RV}$ around an arbitrarily fixed reference value (RV) and requires values of $U(O_i)^2$ over a range of observed concentrations, while the second formulation (27) requires defining uncertainties around only two arbitrarily fixed concentrations (RV and L).

For air quality models that provide yearly averaged pollutant concentrations, the MQO is modified into a criterion in which the mean bias between modelled and measured concentrations is normalized by the expanded uncertainty of the mean concentration (equation 7). For this case, Pernigotti et al (2013) derive the following expression for the uncertainty:

$$U(\bar{O}) = ku_r^{RV}\sqrt{\frac{(1-\alpha)}{N_p^*}(\bar{O}^2 + \sigma_o^2) + \frac{\alpha.RV^2}{N_{np}}} \cong ku_r^{RV}\sqrt{\frac{(1-\alpha)}{N_p}\bar{O}^2 + \frac{\alpha.RV^2}{N_{np}}} \quad (28)$$

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. In equation (28) the standard deviation term is assumed to be linearly related to the observed mean value in the annual average formulation (*i.e.* $\sigma_o = \mathbf{I}\bar{O}$). The calculation of the N_p coefficient accounts for the correction resulting from this assumption. To determine N_p and N_{np} a similar procedure is used as for α and U^{RV} above. Once α and U^{RV} are known from the uncertainties on the hourly observations, values N_p and N_{np} are derived from the uncertainties for the yearly averaged values by using a linear fit between $U(\bar{O})^2$ and $(\bar{O})^2$. You could also again simplify the fitting of the coefficients for the annual expression using the same methodology using two arbitrarily fixed concentrations as presented by equations 26 and 27 above.

The following values are currently proposed for the parameters in (22) and (28) based on Thunis *et al.* (2012), Pernigotti *et al.* (2013) and Pernigotti *et al.* (2014). Note that the value of α for $PM_{2.5}$ referred to in the Pernigotti *et al.* (2014) working note has been arbitrarily modified from 0.018 to 0.050 to avoid larger uncertainties for PM_{10} than $PM_{2.5}$ in the lowest range of concentrations.

Table 1: List of the parameters used to calculate the uncertainty

	k	u_r^{RV}	RV	α	N_p	N_{np}
NO_2	2.00	0.120	200 µg/m^3	0.040	5.2	5.5
O_3	1.40	0.090	120 µg/m^3	0.620	NA	NA
PM10	2.00	0.140	50 µg/m^3	0.018	40	1
PM25	2.00	0.180	25 µg/m^3	0.050	40	1

5.5. Open issues

In this section a few topics are introduced on which there currently is no consensus yet but which merit further consideration.

5.5.1. 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 which are based on dividing the set of measurement data into two groups, one for the integration and one for the evaluation of the integrated fields. The challenge is how to select the set of validation stations. By repeating the procedure e.g. using a Monte Carlo approach until all stations have been included at least once in the evaluation group, validation is then possible for all stations. As a specific case the 'leaving one out' method can be mentioned in which all stations are included in the integration except for the single station that we want to validate. By repeating this procedure for each station in turn, all stations can be validated 'Leaving one out' therefore requires as many re-analyses as there are stations. It is currently investigated within FAIRMODEs Cross Cutting Activity Modelling & Measurements which of the methodologies is most robust and applicable in operational contexts.

5.5.2. 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. 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 MQO and MPC 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.

5.5.3. Handling changes in observation data uncertainty

As defined in 5.2 the MQO depends on the observation data uncertainty. As measurement techniques improve this observation data uncertainty will likely reduce over time. A consequence of this could be that a model that produced results that complied to the MQO based on a set of measurements could have a problem fulfilling the MQO for a new set of measurements obtained using the improved technique. A clear procedure is thus needed on how to define and update the different parameters needed for quantifying the observation data uncertainty.

5.5.4. Performance criteria for high percentile values

The model quality objective described above provides insight on the quality of the model average performances but does not inform on the model capability to reproduce extreme events (e.g. exceedances). For this purpose, a specific MQO indicator is proposed as:

$$MQO_{perc} = \frac{|M_{perc} - O_{perc}|}{2U(O_{perc})} \leq 1 \qquad (29)$$

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 $PM_{2.5}$ 90.1% (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 MQO_{perc} the equation 22 is used with $O_i = O_{perc}$.

5.5.5. 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 PM10, PM2.5, NO2, SO2, O3, 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.

5.5.6. MPC fulfilment criteria: improved statistical basis for the MQO

By considering the requirement that the MPC should be fulfilled in at least 90% of the observation stations as a requirement for the confidence interval of the differences between observed and modelled values an alternative basis for the MQO can be derived.

The Model Quality Objective (MQO) is derived above with the simple principle of allowing a similar margin of tolerance to both model and observations. Assume a set of normal distributed data pairs consisting of observations and model calculations. The standard deviations of the observations and modelled concentrations are σ_O and σ_M. As the observations and model calculations are assumed to follow a normal distribution, their difference does too. For the MQO we define that 90% of the differences between observations and model results must be between $-2\ \sigma_O$ and $+2\ \sigma_O$. In statistical terms: the 90% CI[4] is $4\sigma_O$. From the above it follows that the 95%CI of the concentrations differences, defined by $2\sigma_d$, is given by (2.0/1.64) 4 σ_O. The factor 2.0/1.64 = 1.22 takes into account the difference between the 90%CI and 95%CI. The standard deviation (σ_d) of the distribution of the differences can be expressed as $\sigma_d = \gamma\ \sigma_O$, with $\gamma = 2.44$.

It is furthermore evident that

$(\sigma_d)^2 = (\sigma_O)^2 + (\sigma_M)^2$

or: $(\gamma\ \sigma_O)^2 = (\sigma_O)^2 + (\sigma_M)^2$

or: $(\gamma^2 - 1)\ \sigma_O^2 = \sigma_M^2$

and, finally: $\sigma_M = \sqrt{(\gamma^2 - 1)}\ \sigma_O$

Numerically:

$\sigma_M = \sqrt{(2.44^2 - 1)}\ \sigma_O \approx 2.3\ \sigma_O$.

Conclusion: the present Model Quality Objective (MQO) implies that the uncertainty in the model result can be up to roughly twice as high as the measurement uncertainty.

5.5.7. 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 off course be extended to other variables including meteorological data as proposed in Pernigotti et al. (2014)

In Table 2 below values are proposed for the parameters in (23) and (29) for wind speed and temperature data.

---

[4] Confidence Interval

Table 2 List of the parameters used to calculate the uncertainty for the variables wind speed (WS) and temperature (TEMP)

	k	u_r^{RV}	RV	α	N_p	N_{np}
WS (test)	2.00	0.130	5 m/s	0.800	NA	NA
TEMP (test)	2.00	0.025	25 K	1.000	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 NOx uncertainty is not available but could be approximated by the NO_2 uncertainty for now. (Table 1).

6. REPORTING MODEL PERFORMANCE

6.1. The proposed template

In the reporting composite diagrams (e.g. Taylor, Target,…) are favoured. Benchmarking reports are currently available for the hourly NO_2, the 8h daily maximum O_3 and daily PM_{10} and $PM_{2.5}$. There are different reports for the evaluation of hourly and yearly average model results. Below we present details for these two reports.

6.1.1. Hourly

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

Target Diagram (Figure 2)

The MQO as described by equation (6) is used as main indicator. In the normalised Target diagram, the MQO represents the distance between the origin and a given station point. The performance criterion for the target indicator 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.

In the Target diagram the X and Y axis correspond to the BIAS and $CRMSE$ and are normalized by the observation uncertainty, U. The $CRMSE$ is defined as:

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(M_i - \bar{M}) - (O_i - \bar{O})]^2} \quad (30)$$

and is related to $RMSE$ RMSE and BIAS as follows:

$$RMSE^2 = BIAS^2 + CRMSE^2 \quad (31)$$

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

$$CRMSE^2 = \sigma_o^2 + \sigma_m^2 - 2\sigma_o\sigma_m R \quad (32)$$

For each point representing one station on the diagram the abscissa is then bias/2U, the ordinate is $CRMSE$/2U and the radius is proportional to $RMSE_U$. The green area on the Target plot identifies the area of fulfilment of the MQO.

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 (32) 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:

$$\frac{CRMSE(R=1)}{CRMSE\ (\sigma_M=\sigma_O)} = \frac{|\sigma_M-\sigma_O|}{\sigma_O\sqrt{2(1-R)}} \begin{Bmatrix} > 1: \ \sigma\ dominates\ R: right \\ < 1: \ R\ dominates\ \sigma: left \end{Bmatrix} \tag{33}$$

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 percentage of stations fulfilling the target criterion is indicated in the upper left corner and is meant to be used as the main indicator in the benchmarking procedure. As mentioned above, values higher than 90% must be reached. The uncertainty parameters ($u_r^{RV}, \alpha\ and\ RV$) used to produce the diagram are listed on the top right-hand side.

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.

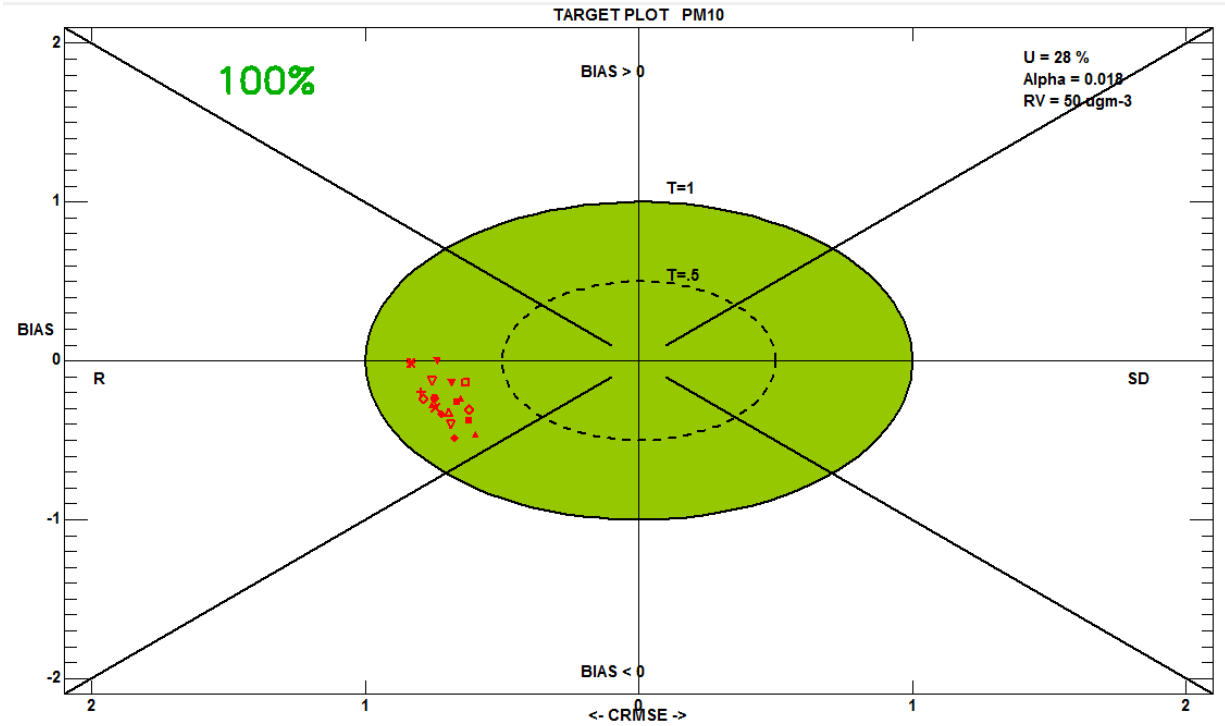


Figure 2 Target diagram to visualize the main aspects of model performance

Summary Report (Figure 3)

The summary statistics table provides additional information on model performances. It is meant as a **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, 120 and 200 µg/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…) for which no threshold exists, the value is set to 1000 so that no exceedance will be 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 via equations (9), (10), (11) and (29) respectively. 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 RMSE 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. Fulfilment of the performance criteria (8) and (9) is then checked for these values. 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

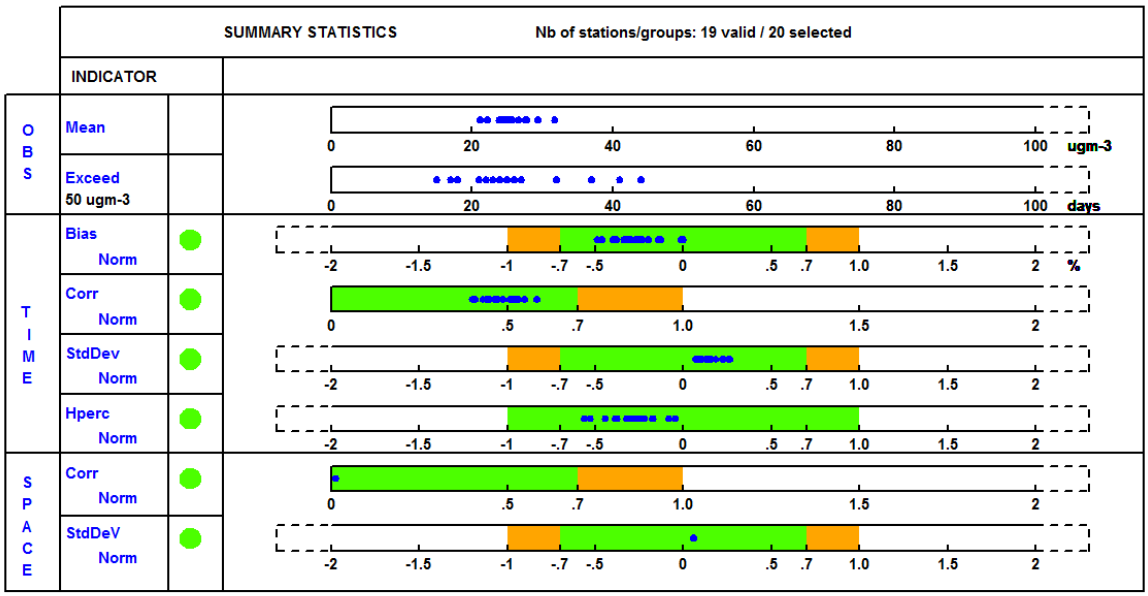


Figure 3 Summary table for statistics

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%.

6.1.2. Yearly average

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)

Scatter Diagram

For yearly averaged results the MQO based on the BIAS (equation 7) is used as the main indicator. In the scatter plot, it is used to represent the distance from the 1:1 line. The MQO is expected to be fulfilled by at least 90% of the available stations. The uncertainty parameters ($u_r^{RV}, \alpha, RV, N_p$ and N_{np}) used to produce the diagram are listed on the top right-hand side

The Scatter diagram also provides information on the performance for single stations or group of stations by presenting these with different symbols and colours.

Summary Report

The summary statistics table provides additional information on the model performance. It is meant as a **complementary** source of information to the bias-based MQO 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 (10) and (11) are used to check fulfilment of the performance criteria. Points that are within the green and the orange shaded area represent those stations where the model performance criterion is fulfilled. For the points that are 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.

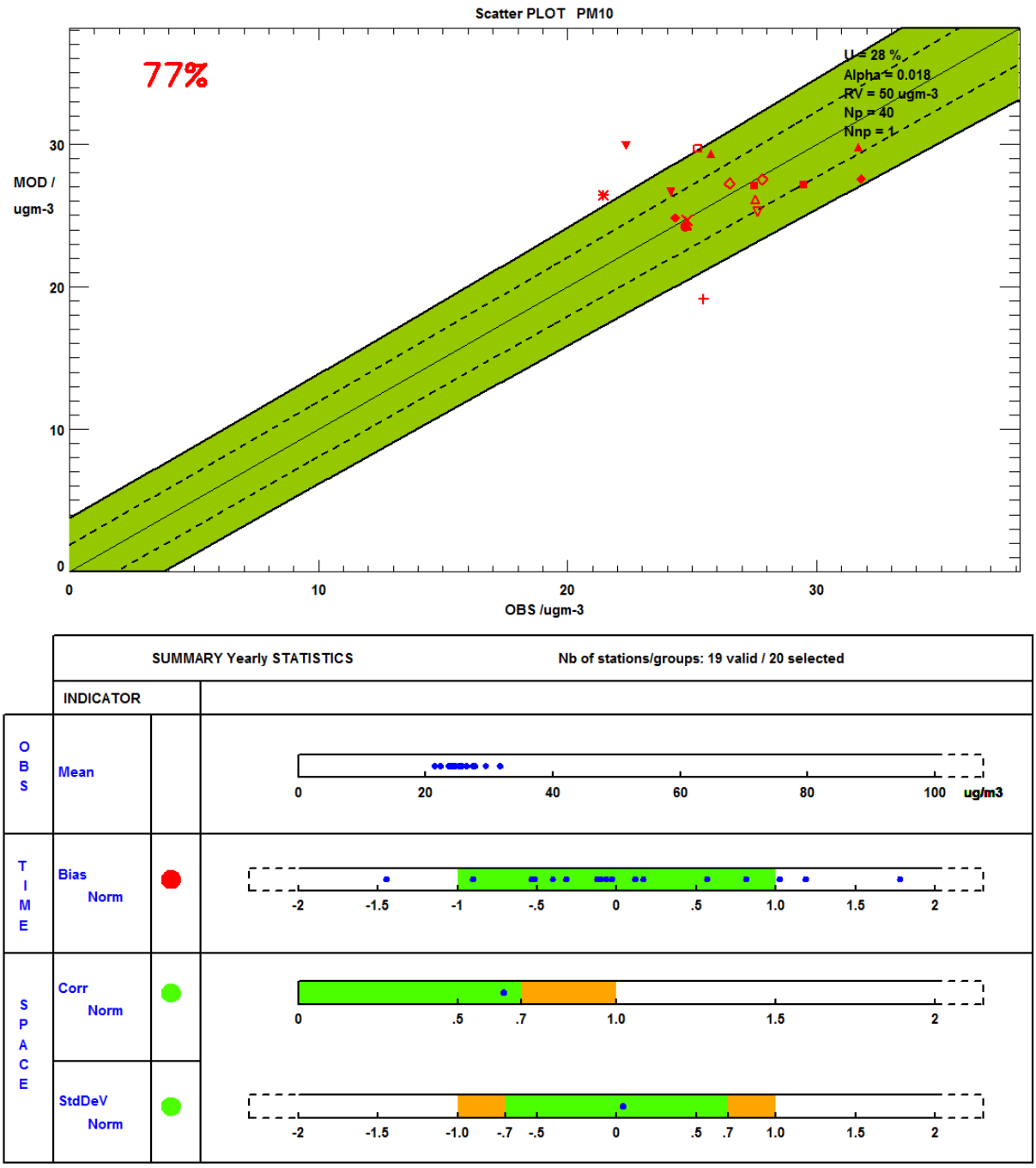


Figure 4 Example of a scatterplot and summary report based on yearly averaged model results.

6.2. Open issues

Based on user feedback the following improvements are proposed to the template:

- In the Summary Report the name of the pollutant indicator for which the report was generated is missing.
- A single symbol is used for the stations in the Summary Report: would it not be possible to reuse the symbols used in the Target Plot/Scatter Diagram to identify the different stations?

- In the Target Plot/Scatter Diagram the colour coding by site type is useful, but a key to the colour coding that is used would be helpful.
- In the summary plots for the observations, the green colour could be used to designate the area where the observations are within the limit values.
- The definitions of the different indicators should be included in the report to make apparent that these are not the 'standard' definitions for bias, correlation and standard deviation but that these have been normalised with the measurement uncertainty.

7. EXAMPLES OF GOOD PRACTICE

In this section we present a number of examples provided to us by the following parties:

- Regional Agency for Environmental Protection and Prevention (ARPA) Emilia Romagna, Italy
- Cambridge Environmental Research Consultants (CERC), United Kingdom
- University of Aveiro, Portugal
- Belgian Interregional Environment Agency (IRCEL), Belgium
- University of Brescia (UNIBS), Italy
- Ricardo AEA, UK

7.1. CERC experience

Jenny Stocker, CERC

Background Information

1. What is the context of your work:
   a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

   Model verification exercise

   b. Scope of the exercise (pollutants, episodes…)

   Pollutants: NO_x, NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2 interested in annual averages and model performance statistics, e.g. correlation, standard deviation. Also plots of results, such as scatter plots, bar-charts, Q-Q plots.

2. Model
   a. Model name

   ADMS-Urban

   b. Main assumptions

   Advanced three dimensional quasi-Gaussian model calculating concentrations hour by hour, nested within a straight line Lagrangian trajectory model which is used to calculate background concentrations approaching the area of interest. Road, industrial and residual sources can be modelled in detail using a variety of options such as terrain, buildings, street canyons and chemistry.

   c. I/O

   Input: *Emissions data, such as a grid of emissions over modelling domain, with detailed road and industrial source emissions; hourly meteorological data and*

measured/modelled background concentrations in text file format; text files containing the variation of terrain heights and roughness lengths over the domain; source parameters such as widths and street canyon geometry for roads, stack heights for industrial sources; and building dimensions.

Output: *Concentrations may be output in an hourly average format over a 2D or 3D grid of receptor points and/or at specified receptor points.*

d. Reference to MDS if available

Reference to MDS: http://pandora.meng.auth.gr/mds/showshort.php?id=18

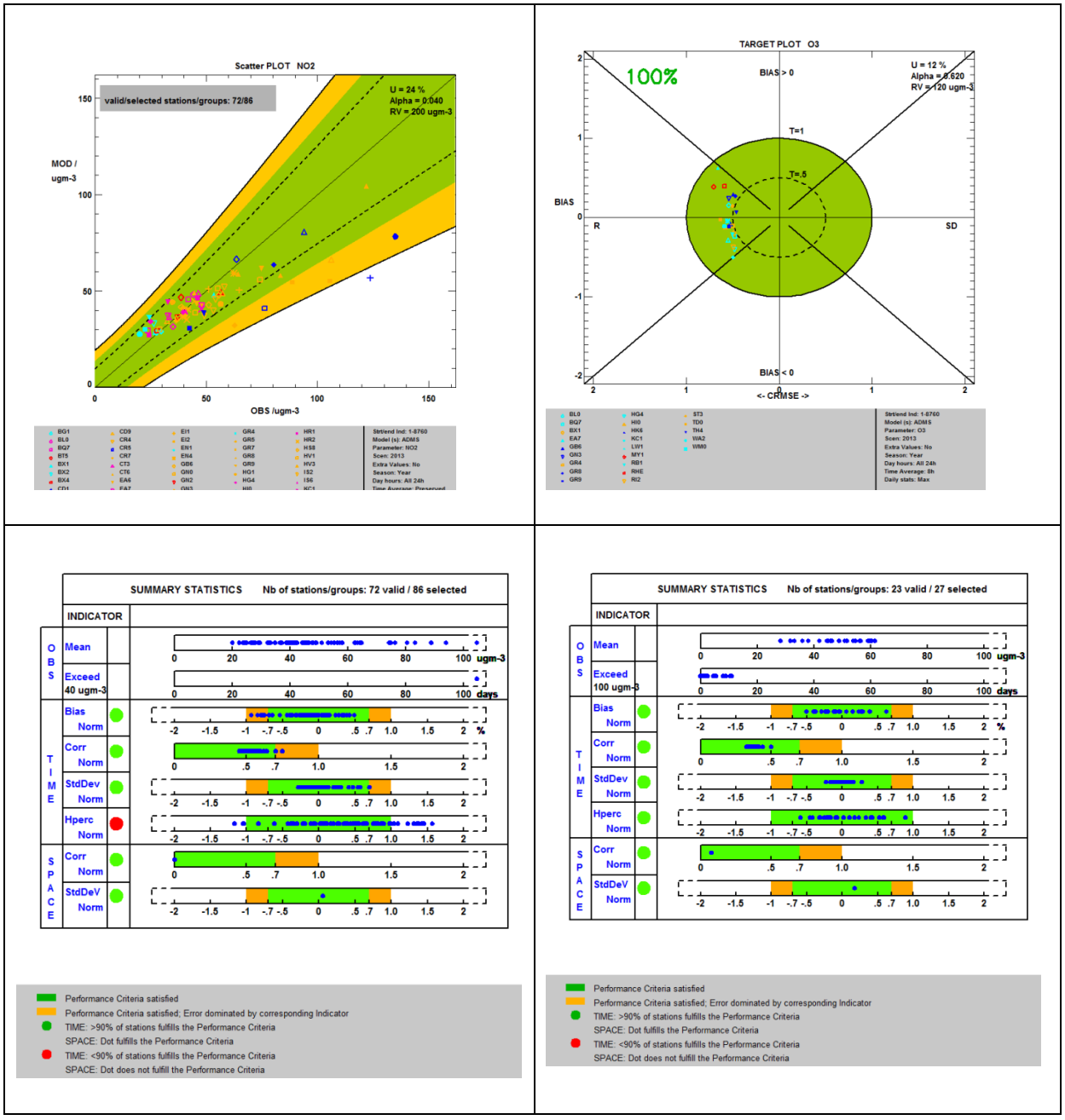


Figure 5 Results of the CERC test case

1. Test-case
    a. Spatial resolution and spatial domain

    Modelling covered Greater London (approx. 40 km x 50 km) with variable resolution, i.e. finer resolution near roadside areas, regular grid in background areas.

    b. Temporal resolution

    Hourly average data for one year at multiple receptor points

    c. Pollutants considered

    NO_x, NO_2, O_3, PM_{10}, $PM_{2.5}$, SO_2

    d. Data assimilation, if yes methodology used

    Not used

Evaluation

1. How did you select the stations used for evaluation?

All available monitoring stations with data for the modelled year have been included in the analysis.

2. In case of data-assimilation, how are the evaluation results prepared?

Not applicable.

3. Please comment the DELTA performance report templates

Looking at the summary statistics report given below both the scatter and target plots (

Figure 5*), it is not clear which station performed well as all sites use the same symbol, so we cannot see which stations are performing well and which are underperforming from this summary plot alone. Even if the individual symbols are not used the colour coding by site type may be useful here.*

Scatter plot

The scatter plot (

Figure 5*) shows the bands well with the bold colouring and the individual symbols for each monitoring site is useful. The colour coding by site type is very useful, it would be good to have the key to this colour coding included in the plot.*

Target plot

Most of the comments for the scatter plot apply to the Target Plot. It is useful to see the individual symbols for each site modelled, the colour coding by site type is useful, but a key to the colour coding used would be helpful.

Further it is stated the left and right hand side of the target plot distinguishes between points that have errors dominated by correlation and those dominated by standard deviation. However, in terms of 'reading' the plot, it is hard to know how close, in terms of accuracy, the points on either side of the plot are; could the plot be replaced by a semi-circle, and the information regarding correlation and standard deviation be presented separately? So that there is a smooth transition between the values, rather than a jump?

Feedback

1. What is your overall experience with DELTA?

CERC argues that there is little justification for insisting that models and measurements are subject to the same degree of error as this would mean that models need to improve as measurement uncertainty becomes smaller. Model objective criteria need to be developed which ensure the model has a performance appropriate for the task for which it is being used, both in terms of application (for example compliance assessment, policy, local planning or research) and scale (for example regional, urban or roadside).

When performing validation, 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.

Furthermore CERC provides feedback on using the DELTA Tool implementation as provided by JRC. Points of improvement to the DELTA Tool implementation provided by CERC relate to the use of IDL and the different file formats that are used for observed and modelled data.

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

CERC currently uses the benchmarking procedure described here but with the Myair toolkit developed during EU FP7 PASODOBLE project.

CERC currently uses the benchmarking procedure described here but with the Myair toolkit developed during EU FP7 PASODOBLE project. Some advantages and additional features of the Myair Toolkit compared to the DELTA tool are:

- *It is more flexible in terms of concentration data input so that for a typical project, much less time is spent re-formatting the data for input into the processing tool.*
- *It includes some additional statistics: the number of valid observations and the observed and modelled maximum concentrations.*
- *It allows statistics to be binned according to site type or pollutant while in the Delta Tool statistics are only given by site type.*

- *It can process many pollutants and datasets together which is very useful for the inter-comparison between different modelled datasets in model validation.*
- *It can produce Box and Whisker plot.*

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

CERC would like to see statistics for each receptor point, and each pollutant in a numerical table. The statistics plot could use a different colour for each site type.

7.2. Applying the DELTA tool v4.0 to NINFA Air Quality System

Michele Stortini, Giovanni Bonafè, Enrico Minguzzi, Marco Deserti

ARPA Emilia Romagna (Italy), Regional Agency for Environmental Protection and Prevention

Background Information

1. What is the context of your work:
   a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

   The Emilia-Romagna Environmental Agency has implemented since 2003 an operational air quality modelling system, called NINFA, for both operational forecast and regional assessment. NINFA recently has been used for the assessment of regional air quality action plan.

   b. Scope of the exercise (pollutants, episodes…)

   O3, PM10, PM25 and NO2

2. Model
   a. Model name

   Chimère version 2008c

   b. Main assumptions

   Not provided

   c. I/O

   Meteorological inputs are from the COSMO-I7, the meteorological Italian Limited Area Model. Chemical boundary conditions are provided by Prev'air data and emission input data are based on regional Emilia-Romagna Inventory (INEMAR), national (ISPRA) and European inventories (MACC).

   d. Reference to MDS if available

   MDS link for Chimère: http://pandora.meng.auth.gr/mds/showlong.php?id=144

3. Test-case
   a. Spatial resolution and spatial domain

   *The simulation domain (640*410 km) covers the northern Italy, with a horizontal resolution of 5 km.*

   b. Temporal resolution

   Hourly resolution: the model runs daily at ARPA and provides concentration for the previous day (hind cast) and the following 72 hours (forecast).

c. Pollutants considered

Concentration maps of PM10, Ozone and NO2 are produced

d. Data assimilation, if yes methodology used

Not used

Evaluation

1. How did you select the stations used for evaluation?

All the observations from the active Emilia-Romagna regional background stations have been used in this study. 13 monitoring station are rural, 13 are urban and 10 suburban.

2. In case of data-assimilation, how are the evaluation results prepared?

Not applicable

3. Please comment the DELTA performance report templates.

Often the station names in bar plot diagrams are not readable because they overlap. The results for PM10 are presented in the figures below (Figure 6, Figure 7 and Figure 8)

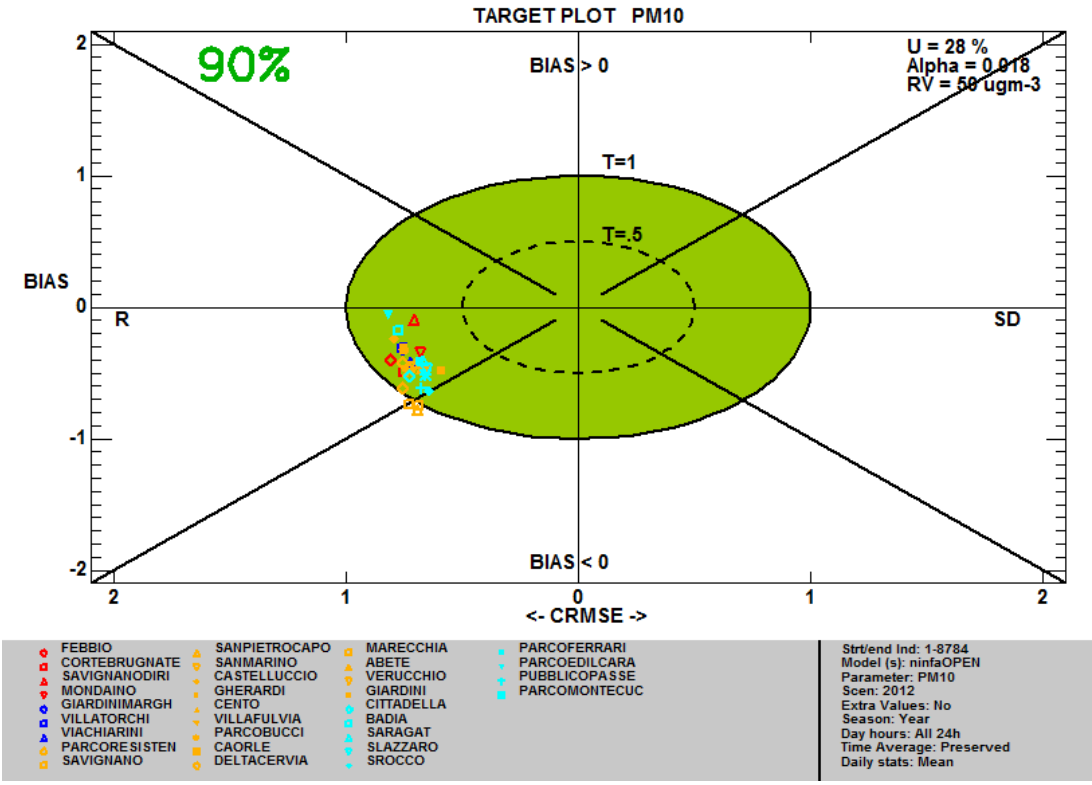


Figure 6 ***Target diagram for daily average PM10 concentrations. Model NINFA, year 2012. Red stations are located in the hills, blue in Bologna area, orange in the east, cyan in the west***

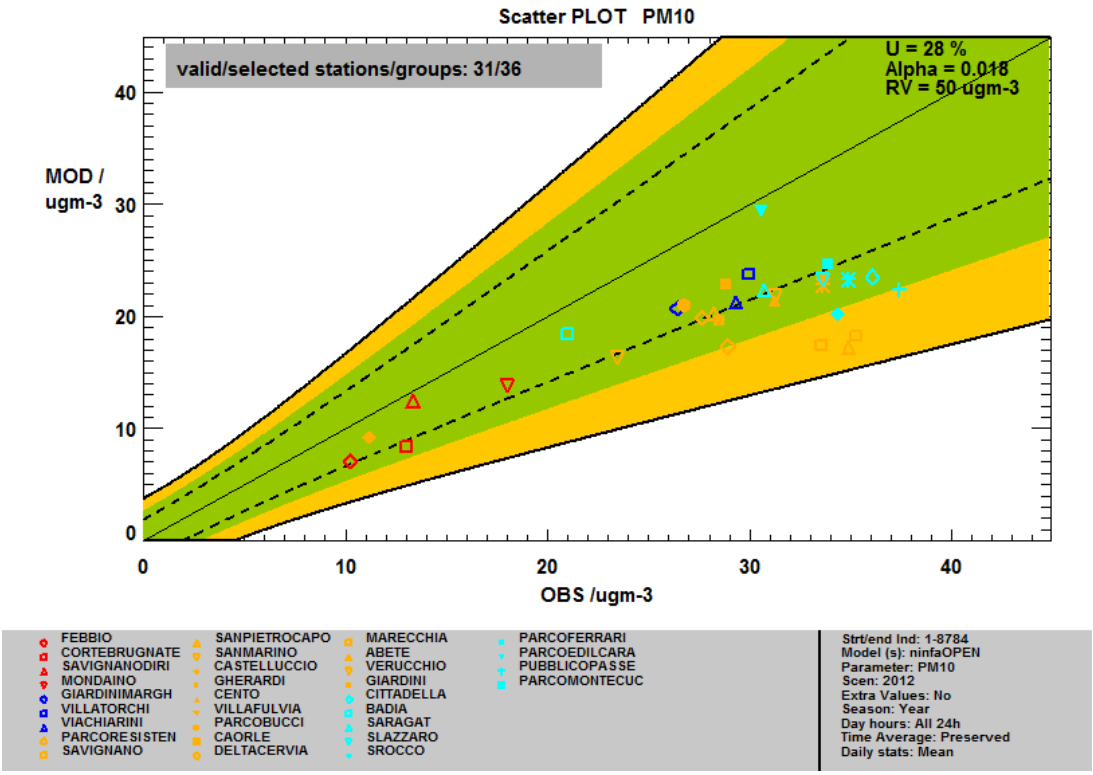


Figure 7 *Scatter plot of the modelled versus measured PM10 concentrations. NINFA, year 2012*

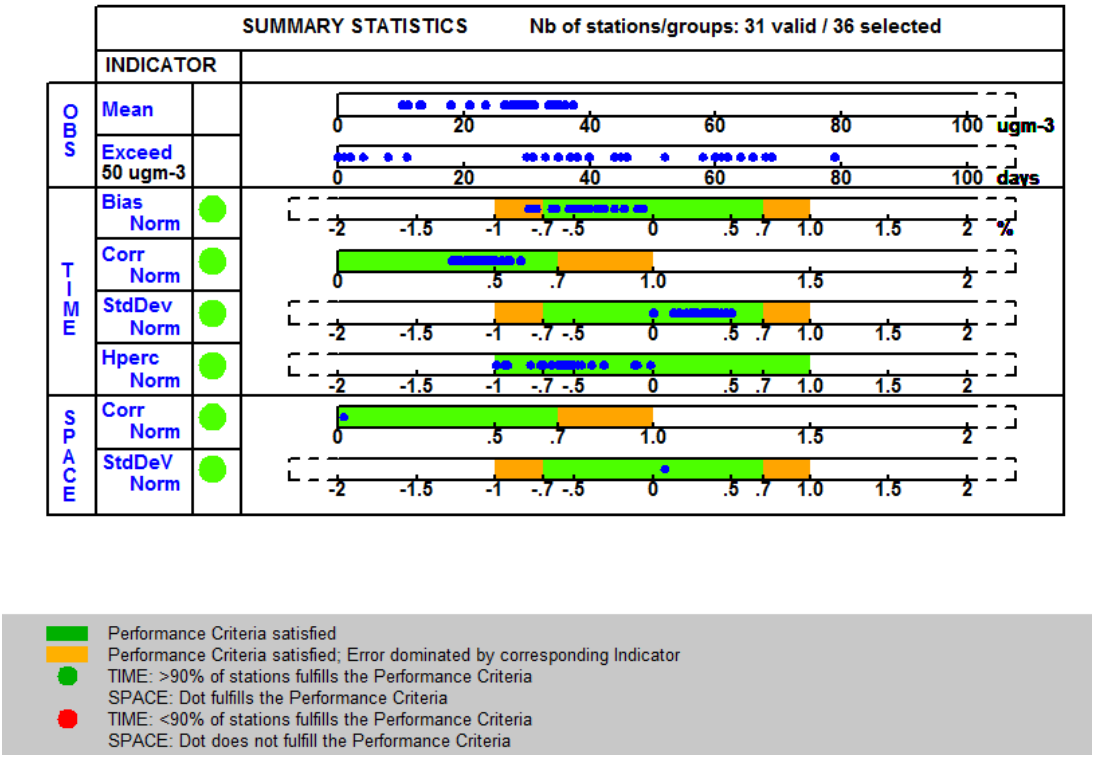


Figure 8 *Summary statistics for daily PM10. NINFA, year2012*

Feedback

1. What is your overall experience with DELTA?

The tool is useful for assessing air quality models, especially because in this way is it possible to use standard methodologies to intercompare air quality model performances. Other comments relate to the implementation of the method: it would be useful to be able to use the tool in batch mode as well as on other operating systems (e.g. Linux).

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

The evaluation is usually performed on statistical index (Bias, correlation, rmse).

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

It could be useful to have time series for a group of stations as well as time series of mean daily values both for individual stations and station groups.

7.3. JOAQUIN Model comparison PM10 NW Europe

Elke Trimpeneers (IRCEL, Belgium)

Background Information

1. What is the context of your work:
   a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

   Joaquin (Joint Air Quality Initiative) is an EU cooperation project supported by the INTERREG IVB North West Europe programme (www.nweurope.eu). The aim of the project is to support health-oriented air quality policies in Europe.

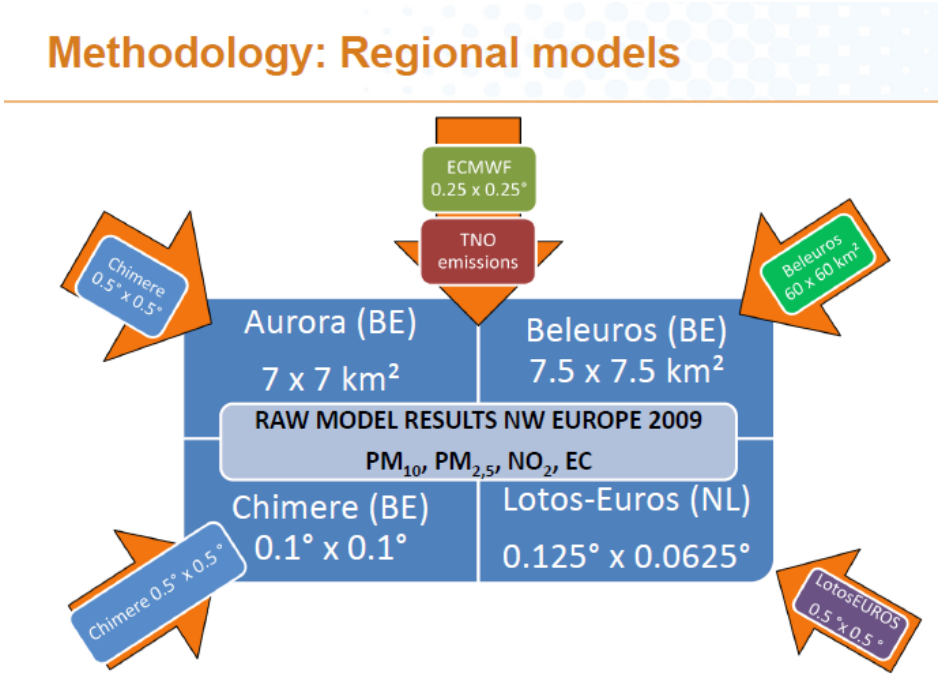
   b. Scope of the exercise (pollutants, episodes…)

   The scope of the exercise is to compare model performances for the pollutant PM_{10} for the NW-Europe domain.

2. Model
   a. Model name

   Four models are used in the exercise: Chimère, Aurora, LotosEuros and Beleuros.

   b. Main assumptions: *see figure below*
   c. I/O : *see figure below*



   d. Reference to MDS if available

- *Chimère: http://pandora.meng.auth.gr/mds/showlong.php?id=144*
- *Aurora : http://pandora.meng.auth.gr/mds/showlong.php?id=167*
- *BelEuros: http://pandora.meng.auth.gr/mds/showlong.php?id=166*
- *LotosEuros: http://pandora.meng.auth.gr/mds/showlong.php?id=57*

3. Test-case
    a. Spatial resolution and spatial domain

    b. Temporal resolution:

Both hourly and yearly data were produced for the 2009.

    c. Pollutants considered

PM_{10}.

    d. Data assimilation, if yes methodology used

No data assimilation used only raw model results.

Evaluation

1. How did you select the stations used for evaluation?

    We selected all background stations within the NW Europe (Joaquin) domain from Airbase data 2009. This resulted in 300 stations to be used for the model comparison.

2. In case of data-assimilation, how are the evaluation results prepared?

    Raw model results were used, no data-assimilation was applied.

3. Please comment the DELTA performance report templates

    IRCEL provides feedback on the evaluation of the model results using the DELTA tool.

    The evaluation is based on the 'raw (=not calibrated, data assimilated) model results' of the four models. None of the models not meet the model quality objective (=target value ≤ 1) in 90% of the stations for the PM10 daily mean model evaluation (Chimère 81 %, Aurora 54 %, BelEuros 80 %, LotosEuros 62 %). The target plots are presented in Figure 9.

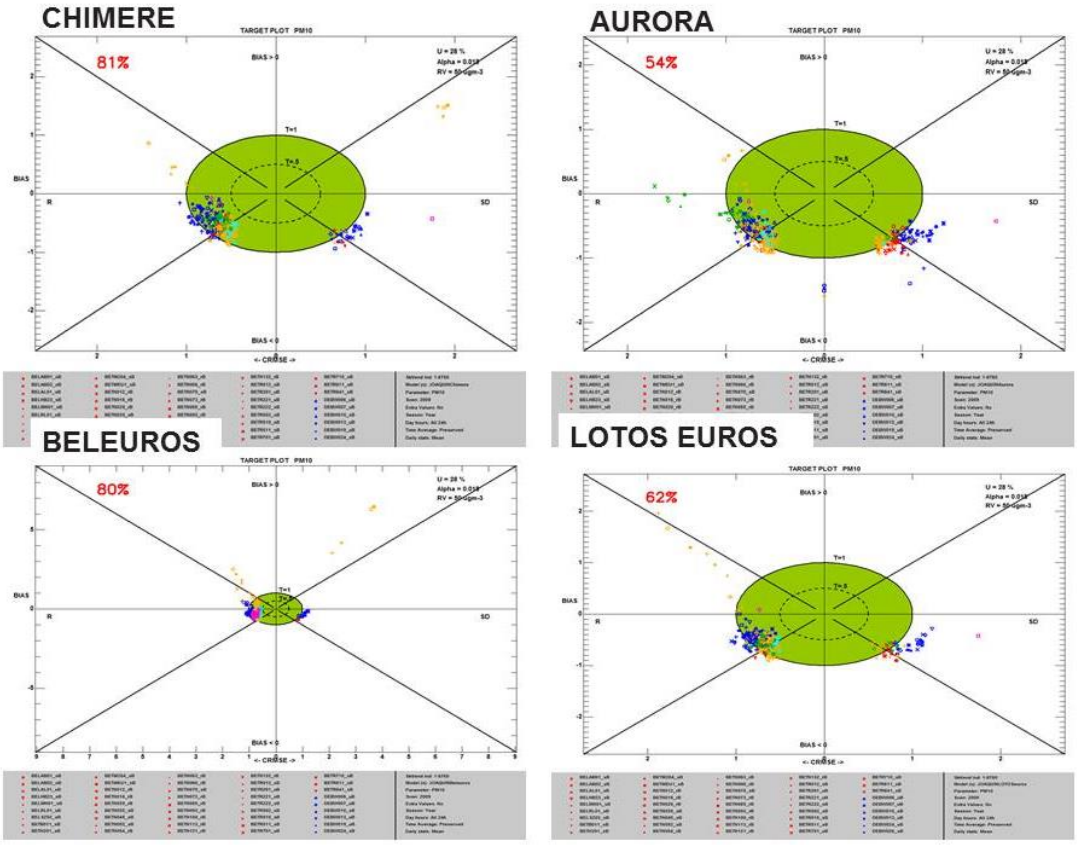


Figure 9 Target plots for the daily average results CHIMERE, BELEUROS, AURORA and LOTOS EUROS.

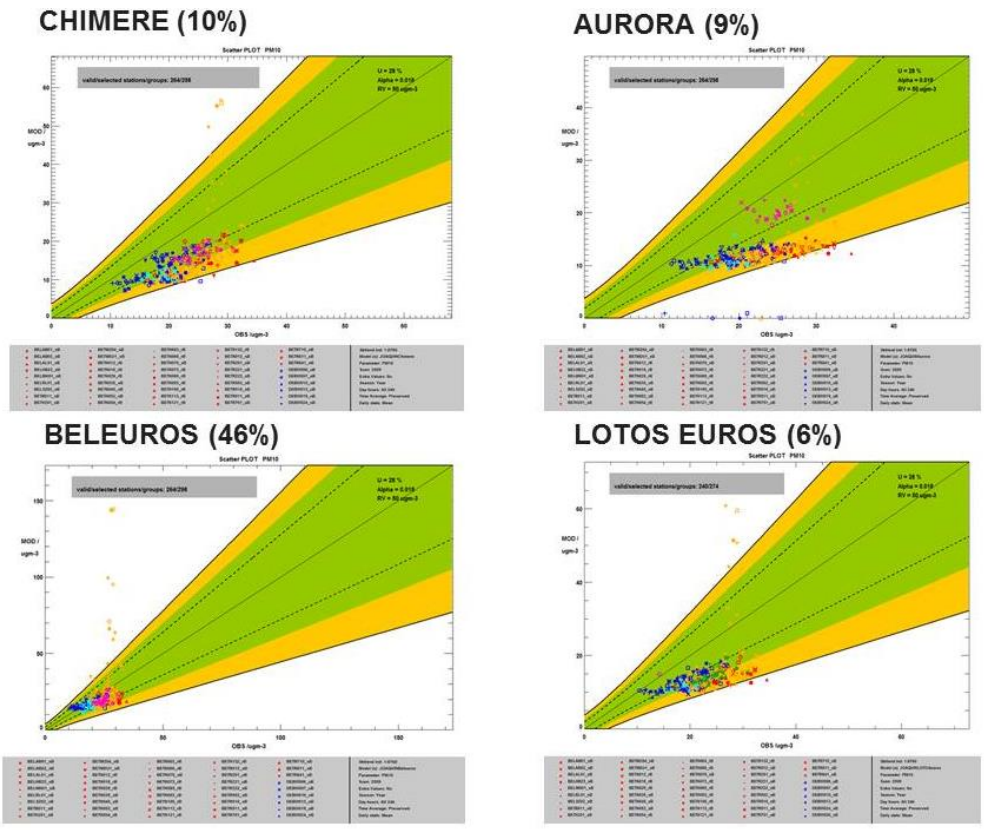


Figure 10 Scatterplots for yearly average results for CHIMERE, BELEUROS, AURORA and LOTOS EUROS.

Noticeable is that the model quality objective for yearly average model results is apparently even harder to comply to in this particular case. For all models the evaluation result based on yearly average model values is worse than the evaluation based on the daily average values. This can be seen from the annual mean scatterplots (Figure 10) where the MQO is only met in respectively 10 % (Chimère), 9 % (Aurora), 46 % (BelEuros)and 6 % (LotosEuros) of the stations. This might seem strange but can be explained by the measurement uncertainty which is lower for the annual mean observed PM10 than for the daily mean values.

Feedback

1. What is your overall experience with DELTA? (5L)

   Most of the feedback is on the actual implementation of the method. Special about this exercise is that so many points (300) are considered. The DELTA Tool implementation considered was able to handle such a large amount of stations but it is difficult to interprete individual station results in this case as legends become cluttered and in practice useless.

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

   IRCEL was already using another implementation of DELTA, the ATMOSYS tool.

   *Concerning the daily mean PM10 results, two models perform relatively well considering the model quality objectives as set in the Delta tool. The results for these same models based on the annual PM10 values are however a lot worse (*Figure 10*).*

   In the latest template an indicator (MQOperc) was added to assess whether a model can correctly calculate exceedances. It was noticed in this specific example that even though the model would apparently comply to the MQOperc objective it still significantly underestimates the number of exceedances. For example in the Belgian station BETR012 that measures the suburban background concentration the 50 µg/m^3 PM10 daily limit value was exceeded 24 times in 2009 while Chimère or Beleuros predict respectively only 4 and 0 exceedances. Both models however comply with the MQOperc model quality objective for the station BETR012.

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

   IRCEL would like to see additional output with statistics for individual stations. This is also useful to be able to do some complementary calculations.

7.4. UAVR experience with DELTA

Alexandra Monteiro and Ana Miranda

Background Information

1. What is the context of your work:
   a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

   Air quality assessment; Air Quality Plans and also research work for publications

   b. Scope of the exercise (pollutants, episodes…)

   PM_{10}, $PM_{2.5}$, NO_2 and O_3 have been considered in the scope of the annual air quality assessment delivered to the Portuguese Agency for Environment and PM_{10} and NO_2 have been worked within AQP. Research activities include all PM_{10}, $PM_{2.5}$, NO_2 and O_3. If other pollutants will be included in DELTA Tool, we would consider them too.

2. Model
   a. Model name

   Different models are used EURAD-IM, CHIMERE, CAMx, TAPM

   b. Reference to MDS if available

   EURAD-IM: http://pandora.meng.auth.gr/mds/showshort.php?id=169

   CHIMERE: http://pandora.meng.auth.gr/mds/showlong.php?id=144

   CAMx: http://pandora.meng.auth.gr/mds/showshort.php?id=141

   TAPM: http://pandora.meng.auth.gr/mds/showlong.php?id=120

3. Test-case
   a. Spatial resolution and spatial domain

   Portugal (9 km x9 km; 3 km x3 km); Porto and Lisbon urban areas (1 km x 1 km)

   b. Temporal resolution

   1 hour

   c. Pollutants considered

   NO2, O3, PM10, PM2.5

   d. Data assimilation, if yes methodology used

   Not used

Evaluation

1. How did you select the stations used for evaluation?

Stations were select according to the data collection efficiency (> 75%) and the type of environment: traffic stations were only included in the urban scale model validation (Porto and Lisbon domains with 1x1 km2). For the other regional scale application we just use background stations (representative of the model grid).

2. In case of data-assimilation, how are the evaluation results prepared?

Not applicable

3. Please comment the DELTA performance report templates

Report templates are an excellent product of DELTA but they still need some improvements to be clearly understood, in particular by the air quality managers, more specifically with respect to the identification of stations and the inclusion of more pollutants to the analysis.

Feedback

1. What is your overall experience with DELTA?

The UAVR experience with the DELTA Tool is based on several model validation exercises that we performed, together with some intercomparison modelling work. This experience involves several model types (EURAD, CHIMERE, CAMx, TAPM), besides all regional scale models, for different type of pollutants (O_3, PM_{10}, $PM_{2.5}$, NO_2) and different spatial domains (Portugal; Porto; Lisbon; Aveiro; …).

Our experience with DELTA is quite positive and we are using it more and more often. DELTA is well documented and relatively easy to apply. The chance to have a common evaluation framework is very well acknowledged and our national air quality management entities receive now model evaluation results based on DELTA and accept these with confidence.

About the things to be improved, we think DELTA should cover all the evaluation aspects included in the Directive:

- *Extend the tool to all pollutants of the Directive*
- *Consider a section for AQ assessment prepared to work with all Directive thresholds;*
- *Consider a section for AQP and its scenarios evaluation (incorporating the Planning Tool that is being developed in work group 4 (WG4) of FAIRMODE);*
- *Consider a section for forecasting purposes with specific model skill/scores (which is already being prepared by INERIS).*

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

Before the DELTA Tool, UAVR performed their model validations using a group of three main statistical parameters (namely BIAS, correlation factor and RMSE) following the work of Borrego et al.[5] produced in the scope of the AIR4EU project (http://www.air4eu.nl/).

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

   The following are missing according to UAVR:

   - *Other pollutants, like CO, SO2, benzene, …*
   - *Distinction of the monitoring sites (difficult to identify the different sites in some graphs/table summary report*
   - *Easy to confuse the traditional parameters and the new ones, since the name is the same (BIAS, Standard Deviation and correlation)*

[5] BORREGO, C., MONTEIRO, A., FERREIRA, J., MIRANDA, A.I., COSTA, A.M., CARVALHO, A.C., LOPES, M. (2008). Procedures for estimation of modelling uncertainty in air quality assessment. Environment International 34, 613–620.

7.5. TCAM evaluation with DELTA tool

Claudio Carnevale (UNIBS, Brescia, Italy)

Background Information

1. What is the context of your work:
   a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

   FAIRMODE Work Group 1 and an internal project at the UNIBS

   b. Scope of the exercise (pollutants, episodes…)

   Application of the methodology to a "real" modelling case. A sensitivity analysis is also performed on the parameters used for the computation of the observation uncertainty.

2. Model
   a. Model name

   TCAM (Transport Chemical Aerosol Model)

   b. Main assumptions

   Horizontal Transport: Chapeau Function (+ Forester FIlter), Vertical Transport: Crack-Nicholson hybrid scheme, Deposition: Wet & Dry, Gas Chemistry: SAPRC mechanism, Aerosol: Condensation/Evaporation, Nucleation, Acqueous Chemistry

   c. I/O

   Emission Inventory: POMI project, 2005
   Meteorology: MM5 2005 output provided by JRC in the frame of POMI project,
   Boundary Condition: Chimère 2005 BC provided in the frame of POMI project

   d. Reference to MDS if available

   http://pandora.meng.auth.gr/mds/showlong.php?id=147

3. Test-case
   a. Spatial resolution and spatial domain

   6 kmx6 km resolution over Northern Italy

   b. Temporal resolution

   Daily

   c. Pollutants considered

   PM10

d. Data assimilation, if yes methodology used

Not used

Evaluation

1. How did you select the stations used for evaluation?

Observations from approximately 50 monitoring sites located in the Po Valley have been used. The sites have been classified in terms of station type (suburban, urban, and rural). The orography (hilly, plane, valley) is also specified. Monitoring data are the same as those used in the model intercomparison exercise (POMI) performed for year 2005.

2. In case of data-assimilation, how are the evaluation results prepared?

Not used

3. Please comment the DELTA performance report templates

No feedback ("Target plot and MQO plot used only")

Feedback

1. What is your overall experience with DELTA?

Comments on tool not the procedure

"Useful for visualizing all main statistical indicators and for summarizing the results of the evaluation in specific statistic tables. It also provides a wide range of plots (scatter, time series, Taylor and target diagrams), which helps to tell whether the overall model response is actually acceptable for regulatory purposes according to the AQD (2008) guidelines."

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

Without the DELTA tool, the evaluation is usually performed in our cases on statistical indexes (correlation, RMSE, bias etc…) and on exceedance days modelling without considering the uncertainty in the measurements.

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

No feedback

7.6. UK feedback Ricardo AEA

Keith Vincent

The feedback is based on a comparison that was made between the DELTA 4.0 implementation by JRC and a spreadsheet calculation.

Background Information

1. What is the context of your work?
    a. Frame of the modelling exercise (Air Quality Plan, research project, …?)

    Evaluation of the PCM modelled results produced as part of the annual AQ compliance for 2013 for the UK

    b. Scope of the exercise (pollutants, episodes…)

    This evaluation is carried out for NO_2, PM_{10} and $PM_{2.5}$ concentrations.

2. Model
    a. Model name

    The Pollution Climate Mapping (PCM) model is a collection of models designed to fulfil part of the UK's EU Directive (2008/50/EC) requirements to report on the concentrations of particular pollutants in the atmosphere.

    b. Main assumptions

    Not provided

    c. I/O

    Not provided

    d. Reference to MDS if available

    Not provided

3. Test-case
    a. Spatial resolution and spatial domain

    The modelling is for the UK, the resolution is 1km x 1km.

    b. Temporal resolution

    Annual average concentrations

    c. Pollutants considered

    NO_2, PM_{10} and $PM_{2.5}$

d. Data assimilation, if yes methodology used

Not used

Evaluation

1. How did you select the stations used for evaluation?

This evaluation is carried out for NO2, PM10 and PM2.5 concentrations predicted at both non-traffic (background + industrial) and traffic locations. This is because different models are used to predict concentrations for the respective locations.

2. In case of data-assimilation, how are the evaluation results prepared?

Not applicable

3. Please comment the DELTA performance report templates (10L per report)

No feedback

Feedback

1. What is your overall experience with DELTA?

Ricardo-AEA has for a number of years played a supporting role in assessing and understanding the usefulness of the MQOs based on measurement uncertainty. A spreadsheet tool (spreadsheet_deltatool_v4.xls) has been developed by Ricardo-AEA and this replicates some of the functionality provided by the Delta tool. This has provided a degree of confidence in how the Delta tool has been applied. (drawbacks/advantages of the method are not provided)

2. How do you compare the benchmarking report of DELTA with the evaluation procedure you normally use? Please briefly describe the procedure you normally use for model evaluation?

No information is given on the normal procedure at Ricardo AEA. The results of the latest implementation of DELTA are compared to those of spreadsheet_deltatool_v4.xls. There seems to be a slight difference in how the fulfilment criteria is calculated between the two implementations. It is noticed that the N_p and N_{np} parameters seem to be treated as integers in DELTA 4.0. The parameters used for PM (u_r^{LV} , α) should also be changed depending on the measurement technique that is used.

3. What do you miss in the DELTA benchmarking report and/or which information do you find unnecessary

No feedback

8. REFERENCES

8.1. Peer reviewed articles:

1. Applying the Delta tool to support AQD: The validation of the TCAM chemical transport model, C. Carnevale, G. Finzi, A. Pederzoli, E.Pisoni, P. Thunis, E.Turrini, M.Volta, Air Quality, Atmosphere and Health , 10.1007/s11869-014-0240-4, 2014
2. Model quality objectives based on measurement uncertainty. Part I: Ozone, P. Thunis, D. Pernigotti and M. Gerboles, 2013. Atmospheric Environment, 79 (2013) 861-868.
3. Model quality objectives based on measurement uncertainty. Part II: PM10 and NO2, D. Pernigotti, P. Thunis, C. Belis and M. Gerboles, 2013. Atmospheric Environment, 79 (2013) 869-878.
4. Performance criteria to evaluate air quality modelling applications. P. Thunis, A. Pederzoli, D. Pernigotti. Atmospheric Environment, 59, 476-482, 2012.
5. A tool to evaluate air quality model performances in regulatory applications, P. Thunis, E. Georgieva, A. Pederzoli, Environmental Modelling & Software, 38, 220-230, 2012
6. A methodology for the evaluation of re-analysed PM10 concentration fields: a case study over the Po Valley, C. Carnevale, G. Finzi, A. Pederzoli, E. Pisoni, P. Thunis, E. Turrini, M. Volta. Air quality Atmosphere and Health, in press

8.2. Reports/ working documents / user manuals:

7. FAIRMODE SG4 Report Model quality objectives Template performance report & DELTA updates P. Thunis, A. Pederzoli, D. Pernigotti March 2012. http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_March2012.pdf
8. Modeling quality objectives in the framework of the FAIRMODE project: working document. D. Pernigotti, P. Thunis and M. Gerboles, 2014.
9. (http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Working%20note_MQO.pdf)
10. The DELTA tool and Benchmarking Report template Concepts and User guide P. Thunis, E. Georgieva, A. Pederzoli Joint Research Centre, Ispra Version 2 04 April 2011http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_April2011.pdf
11. A procedure for air quality models benchmarking P. Thunis, E. Georgieva, S. Galmarini Joint Research Centre, Ispra Version 2 16 February 2011 http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/WG2_SG4_benchmarking_V2.pdf
12. DELTA Version 4.0 Concepts / User's Guide / Diagrams P. Thunis, C. Cuvelier, A. Pederzoli, E. Georgieva, D. Pernigotti, B. Degraeuwe Joint Research Centre, Ispra, September 2014

8.3. Other documents/ e-mail:

13. Feedback on Model Quality Objective formulation D.Brookes, J. Stedman, K. Vincent, B. Stacey, Ricardo-AEA, 18/06/14.
14. Mail correspondence between RIVM – The Netherlands (J. Wesseling) and JRC (P.Thunis)