# FAIRMODE SG4 Report

# Model quality objectives
# Template performance report
# &
# DELTA updates

**P. Thunis, A. Pederzoli, D. Pernigotti**

**March 2012**

# Table of Contents

# 1 Introduction

Two documents have been distributed so far within FAIRMODE Sub Group 4 to support the activities on model benchmarking. In the first document (Thunis et al., 2010; referred here after as **D1**) a general overview of the benchmarking procedure has been presented while in the second document (Thunis et al., 2011; referred hereafter as **D2**) details about the DELTA tool (one component of the benchmarking procedure) have been given. In **D2** a template for reporting model performances has been presented as well. These two documents have been discussed during the SG4 Fairmode meetings.

This third document must be seen as a follow up of the second one. It builds on the comments/suggestions/requests received during SG4 meetings, as well as on the experience gained in the meantime with new model/measurements datasets. It is structured around the following issues:

1. Current Model Quality Objectives (**MQO**) and alternatives.
2. Suggested modifications to the template for reporting model performances
3. Links between MQO and performance criteria for selected statistical indicators
4. Updates on recent developments of the Delta Tool
5. Generalization of the report template to other applications/scales

First we review the MQO currently proposed in the Air Quality Directive (**AQD**, 2008) using available model/measurement datasets highlighting some of their limitations. An alternative set of MQO based on RMSE (Target indicator) has been proposed in **D2** and its advantages are briefly recalled here.

We then discuss some possible modifications to the template for reporting model performances. These modifications mostly deal with 1) the choice of indicators which should be retained in the core set of statistics to perform the evaluation of a given model application. This core set should be large enough to capture the main model performance aspects but limited enough to serve still as a summary overview; 2) the visual representation of these indicators, 3) an adaptation of the Target indicator to relate model performances to observation uncertainty (this last point is detailed in Section 3.2.3) and 4) the setting of consistent performance criteria for the core set of statistical indicators.

This document is distributed together with the release of a new version of DELTA (version 2) which includes some significant improvements (i.e. batch mode, benchmarking service, Group modes, Google Earth visualization…) detailed in Section 5. Some of these DELTA updates allow additional types of applications as for example the possibility of dealing with statistical models providing only annual averages. A new version of the target plot based on observation uncertainty has also been introduced. Other extensions (e.g. regional scale applications) are briefly discussed as well.

**The aim of this document is to serve as a basis for discussions at the next FAIRMODE meeting to be held in May 2012.**

# 2 Model Quality Objectives

Modelling Quality Objectives (MQO) are mentioned in the AQD (2008) but the wording of the text remains ambiguous and open to interpretation. The Guidance on the use of models for the EU AQD (Denby, 2010) reviews different interpretations and recommends the use of the Relative Directive Error (RDE) indicator to provide a quantitative estimate of the model uncertainty. Despite these recommendations, the current AQD MQO retain limitations and ambiguity that are inherent to their formulation. In particular MQO only focus on the model behaviour around the appropriate limit (or target) value and do not account for the timing of events. In addition they are restricted to assessment applications while the use of models for assessing the impacts of air quality plans (planning applications) is becoming widespread and requires a proper evaluation as well. The current MQO therefore only provide a partial view of the strengths and weaknesses for a given model application and do not guarantee that models reach good performances for the good reasons. This becomes problematic when models are applied outside the range of applications for which they have been evaluated.

A proposal for new MQO has been made with the view of providing more insight on how well a model performs for a given AQD application and indicate some directions on how to improve model performances. This is discussed in more details in the next section.

## *2.1 Current definition*

In this section we refer to what is stated in the 2008 AQD and in Denby (2010). The MQO described in the AQD along with the monitoring quality objectives, are given as a relative uncertainty (%). The AQD defines the modeling uncertainty 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. The uncertainty for modelling shall be interpreted as being applicable in the region of the appropriate limit value (or target value in the case of ozone). The fixed measurements that have to be selected for comparison with modelling results shall be representative of the scale covered by the model."*

According to this definition, the AQD states that the uncertainty will be determined from the maximum of 90% of the available monitoring stations. Furthermore, it will be computed "without considering the timing of the event", which means that any temporal correspondence between modeled and observed values will be disregarded. The official statistical indicator currently used for estimating modeling uncertainty is the Relative Directive Error (RDE) (Denby, 2010). It is mathematically defined at a single station as follows:

$$RDE = \frac{|O_{LV} - M_{LV}|}{LV}$$

where $O_{LV}$ is the observed concentration closest to the limit value concentration ($LV$) and $M_{LV}$ is the correspondingly ranked modeled concentration. The maximum of this value found at 90% of the available stations is then the Maximum Relative Directive Error (MRDE).

This formulation is similar to the one recommended (Stern and Flemming, 2004) called the Relative Percentile Error (RPE), which is defined at a single station as:

$$RPE = \frac{|O_P - M_P|}{O_P}$$

where $O_P$ and $M_P$ are the observed and modeled concentrations at the percentile (P), used to define the exceedance percentile.

## 2.2 Limitations

There are arguments against the adoption of RDE as official statistical indicator of the AQD, due to the limitations related to the definition itself:

- The AQD defines a range of applications for which models can be applied for assessment purposes instead of, or in combination with fixed measurements. Denby (2010) review these possible model usages. As stated in this document the use of models is encouraged in all situations to supplement fixed measurements and models can even be used as a substitute to fixed measurements when air quality levels are below the lower assessment threshold. On the other hand the MQO as currently defined in the AQD are required to be satisfied only around the limit/target values. It therefore provides information only about model performances around this value and leaves space for models to be used for a range of applications for which quality is not assessed properly.

- This approach, which does not take into account the timing of events, potentially allows models to perform in a completely uncorrelated way with respect to observations (e.g. a model could wrongly deliver max $O_3$ values in wintertime rather than in summertime and still fulfill the current MQO). Models are therefore given the possibility to fulfill the quality objectives although they have incomplete or wrong parameterizations of the physical and chemical processes. As a consequence the RDE as MQO does not provide insight on why model performances are good or bad and therefore no insight on whether model performances need to be improved or not.

To illustrate these points we have selected a few examples extracted from three different datasets: 1) a one year CHIMERE simulation (EC4MACS dataset) covering all Europe with a 7 km resolution (urban scale modeling) 2) a modeling exercise performed over the Po Valley (POMI dataset) and 3) a simulation over the London area (local scale modeling with the ADMS model).

Figure 1 shows for these three examples how RDE and RPE can be influenced by outliers and lead to ambiguous or wrong interpretations. In the first example the model performs poorly as clearly seen from the time series and the quantile-quantile (Q-Q) plot of $NO_2$ concentrations. A large overestimation is visible over the whole range of values. The main statistical indicators confirm this with a mean fractional bias (MFB) exceeding 70%. However, given the very good agreement of the model with the observed value close to the limit value (200 $\mu gm^{-3}$), the model performs well for both AQD indicators, i.e. RDE and RPE with values of 10% and 21% respectively.

In the second example, an opposite behavior is seen. Here the model performs very well most of the time and correlation, bias and standard deviation indicators show good values but the model performs very poorly around the limit value, leading to a very large RDE indicator.

The third example illustrates how a model could perform relatively poorly both below and above the limit value but still keep good RDE and RPE statistics because of a very good behavior around the limit value. Of course in this specific case all performance statistical indicators are relatively good (R=0.48, MFB=8% and Target=1.05) but this example illustrates well why the RDE and RPE indicators cannot be used to gain insight into model performances and identify ways to improve these performances.

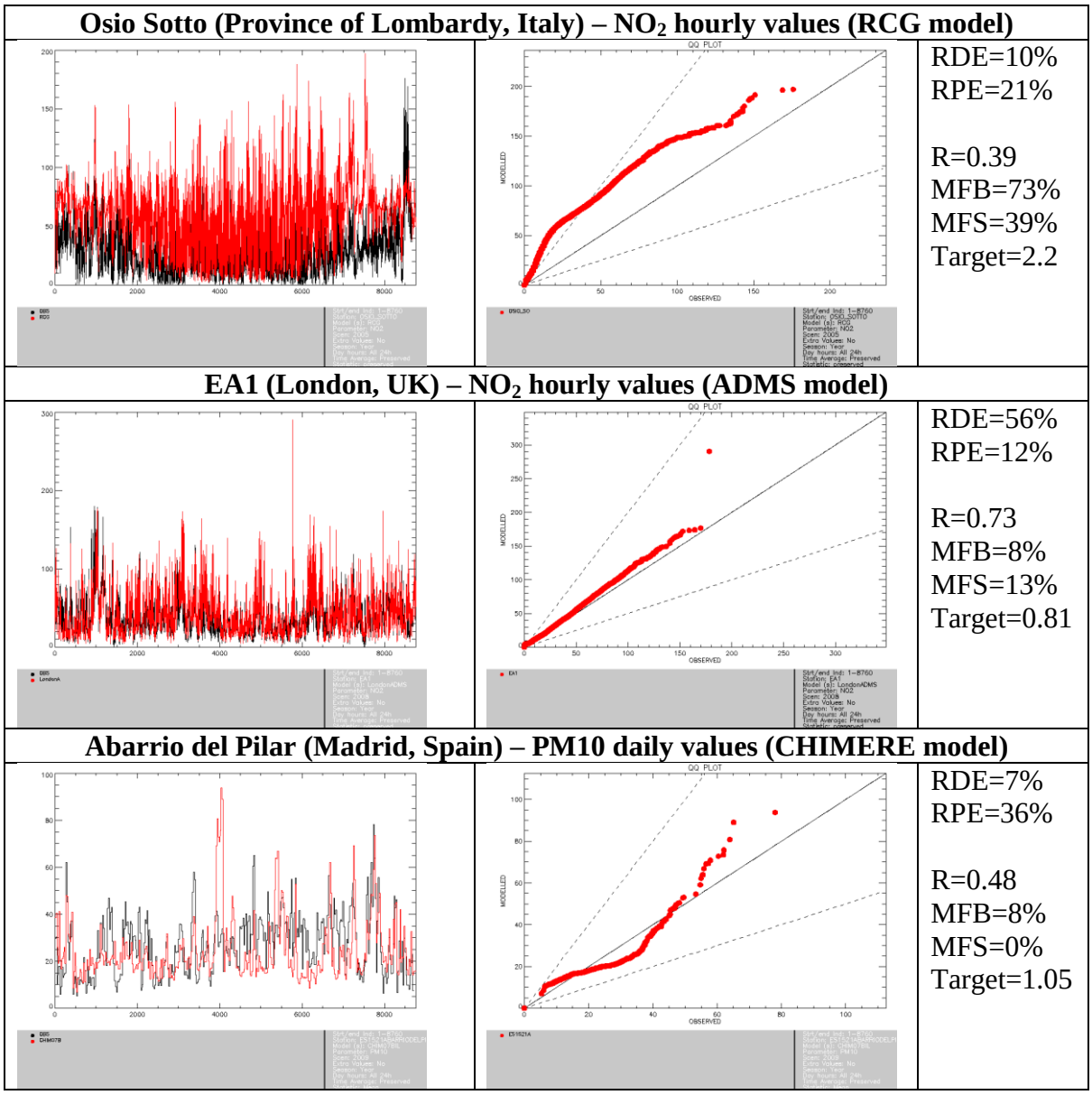



**Figure 1: Examples of stations statistics. In the quantile-quantile diagram, modelled and observed concentrations are ranked, from lowest to highest and plotted one versus each other. Timing of the events is therefore not considered in this type of graphical representation.**

In addition to these three specific examples Figure 2 shows how the RDE compares to three statistical indicators: R (correlation), NMB (Normalized Mean Bias) and NMSD (Normalized Mean Standard Deviation) defined as

$$NMB = \frac{BIAS}{\overline{O}}; \qquad NMSD = \frac{(\sigma_M - \sigma_O)}{\sigma_O}$$

where $\sigma_M$ and $\sigma_O$ are the standard deviation of the model results and observations respectively. Here all stations across Europe are considered (rural, suburban, urban and traffic stations) and performances of the CHIMERE 7 km model are evaluated. The objective is here to understand what a good RDE means in terms of currently used indicators and vice-versa. As expected from its definition (timing of the events not considered), no evidence of correlation between RDE and R is seen and a significant number of stations for which the correlation is very low (e.g. below 0.3 for $NO_2$) still fulfill the RDE criteria (50%). The same happens for PM10. For $O_3$ however, there is a quite good agreement between RDE and other indicators. For bias and standard deviation) some link seems to exist, especially for $NO_2$ and PM10 but there is still a large proportion of stations characterized by large errors in these two indicators which yet fulfill the RDE criteria.

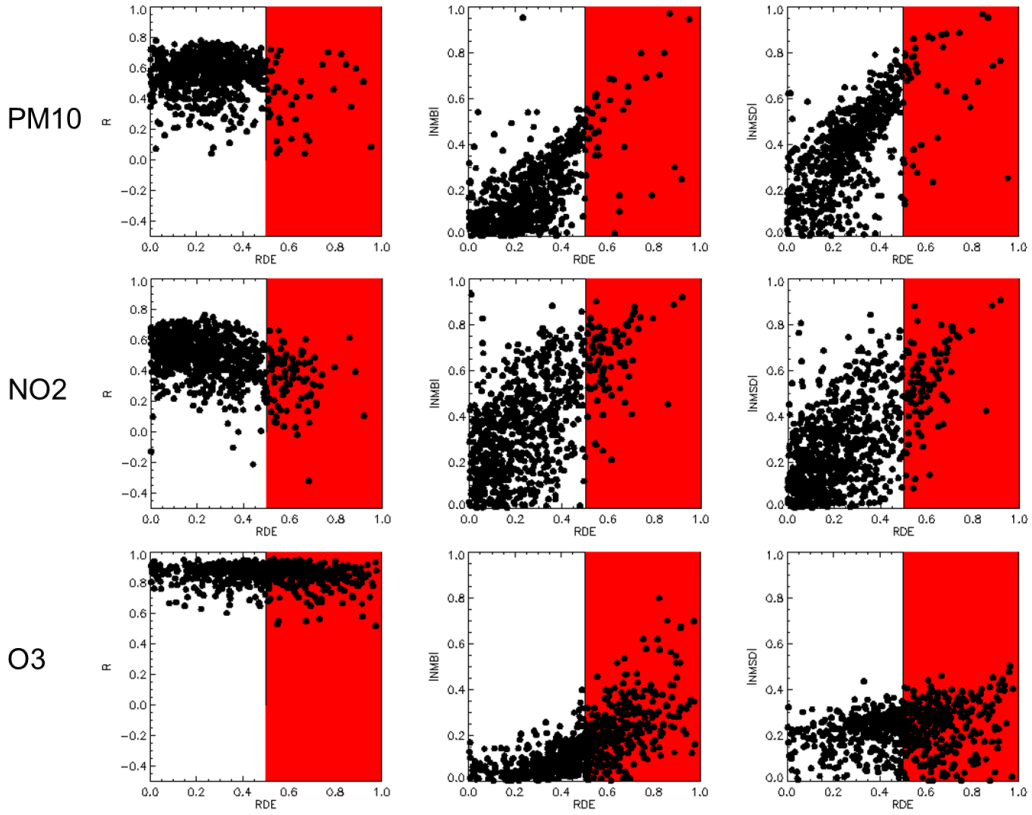



**Figure 2: Links between RDE and other performance indicators: R(left column) , |NMB| (center column)and NMSD (right column) for CHIMERE 7 km resolution concentrations at all EC4MACS monitoring stations.**

Given these limitations we come back in the following section to the MQO which have been introduced in **D2**, highlighting their advantages and proposing some modifications. The scope is to provide more information about model performances and relate these performances to measurement uncertainty.

## 2.3 An alternative to the current AQD MQO

As mentioned in **D2** we propose to replace the MQO currently used in the AQD by a condition on the Target indicator which can be expressed as follows:

$$\max_{90\%\text{stations}} \langle \text{Target Indicator} \rangle \leq \text{Target criterion}$$

where the selected Target indicator is defined as RMSE/*Norm* with *Norm* as normalization factor (e.g. in Jolliff et al., (2009) the selected *Norm* is $\sigma_O$). Similarly to the current MQO the maximum value of the Target indicator found among 90% of the available monitoring stations (in the following called P90) should be tested against the Target criterion. Note that this criterion could be pollutant specific but also scale or time dependent.

The Target indicator based on RMSE synthetizes the model performances in terms of phase, amplitude and bias for each single monitoring station. One disadvantage of this formulation is that all this information is contained in a single number. This is why the Target indicator is complemented by a diagram and a summary statistics table in the performance report. This diagram and table are meant to help the user with the interpretation of the Target indicator and highlight weaknesses and strengths of a given model application.

It is also important to note that within this formulation, 1) all values are treated identically regardless of their absolute concentration level (no emphasis on the limit values as it is for the current AQD MQO) and 2) the timing of the event is considered. A comparison between the current AQD MQO and the Target criterion is proposed in Section 4.2

# 3 Model performances template report

## *3.1 Reference version*

A model performance template has been presented in **D2** and discussed during the FAIRMODE SG4 meeting in May 2011. We start the following discussion with this template, reproduced in Figure 3. For more information on this template please refer to **D2**.

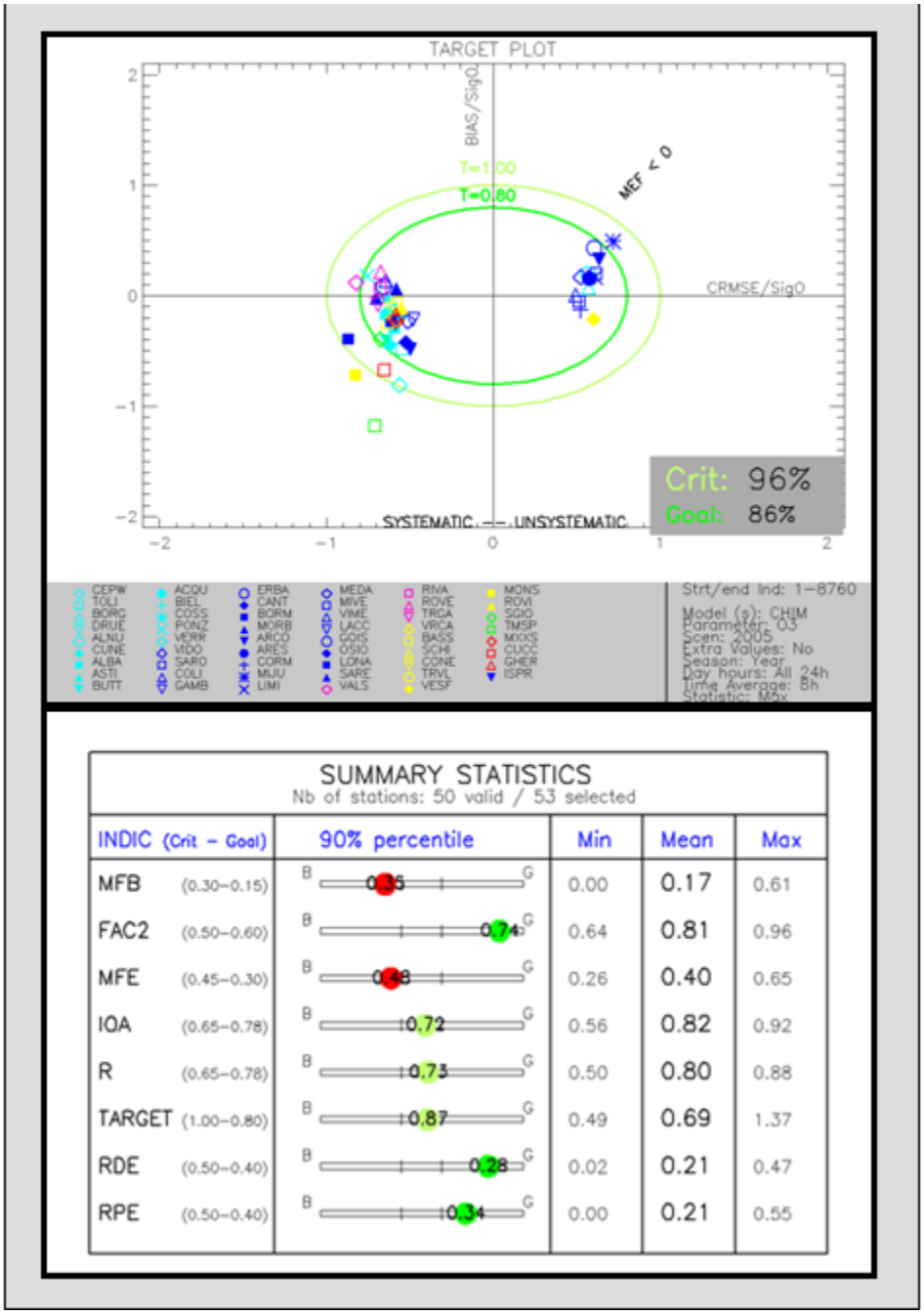



**Figure 3: Reference template (as proposed in D2)**

Before detailing the proposed changes (next section) to the template let's emphasize its main aspects which will remain unchanged:

- The report is divided into two sections. On top the Target diagram aims at delivering a qualitative overview of model performances providing some visual insight on how well the model performs in terms of bias, RMSE and CRMSE (unbiased root mean square error) for each individual station. At bottom a summary of the model performances for a core set of statistical indicators is presented. This part of the report is intended to support the interpretation of the Target indicator by detailing the performance statistics with the aim to help the modeler improving a particular model application.

- The MQO i.e. a minimum level of performance expected to be reached for policy applications has been introduced within the Target diagram. As mentioned above P90 of the available monitoring stations should lie between the origin and the criterion radius. The number of stations actually fulfilling the criterion is indicated in the bottom right part of the diagram and should therefore be larger than 90% for a model to be applied for policy for a given application.

- Information on observation uncertainty is provided on the Target diagram to provide some insight to the modeler on how well his model performs when accounting for the margin of uncertainty related to observations.

The following section presents some modifications to the performance report based on the discussions held during the FAIRMODE SG4 meeting in May2011 and on further investigations made on available datasets. It highlights the main differences between the version of the report as presented in **D2** and the new one.

## *3.2 Proposed updates to the report template*

### 3.2.1 Core set of statistics in the summary statistics table

Regarding the core set of statistics available in the summary table (bottom part of the report), we propose to organize them into three main categories: temporal, spatial and AQD.

- For the **temporal group** (related to time series), we propose to simplify the previous selection of indicators by retaining only three main ones to look how well the model performs in terms of phase, amplitude and bias: R, NMB and NMSD. We propose to eliminate from the list proposed in **D2** the Target indicator (already available in the top part), the Index of Agreement (generally very close to R), FAC2 (in general providing redundant information to the bias) and the mean fractional error (MFE) which also deliver information captured (at least in great part) in the RMSE or target

indicator. We also propose to substitute the mean fractional Bias (MFB) by the Normalised Mean Bias (NMB).

- A new group of **spatial statistics** is introduced to address how well the model performs spatially (i.e. the model capability to reproduce both areas of high and low concentrations as well as the spatial variability of concentrations.

- Regarding the **AQD indicators**, we propose to reduce them to RDE only.

The above changes will result in a simplified summary statistics Table where redundancy among indicators is reduced and where some aspects related to spatial performances are introduced. Note that the core set as currently defined for the temporal indicators is very similar to the one proposed by Borrego et al. (2008).

The performance criteria used to identify the fulfillment zones in the summary table are detailed in Section 4.

### 3.2.2 Explicit station representation in the summary statistics table

One of the drawbacks of the performance report proposed in **D2** is that information is provided only in relative terms. Most indicators are indeed expressed in terms of percentage or are normalized by some quantity (e.g. the Target diagram has all indicators normalized by the standard deviation of the observations). No information is therefore available on absolute values measured or modeled at stations (e.g. mean values, number of exceedances…). We therefore added a specific section in the summary statistic table (Figure 5) to provide information on mean and exceedance values observed at monitoring stations (see first two rows in the summary table of Figure 5).

In the summary table stations are now represented individually (each point corresponding to the value measured at a specific station and not as the 90P indicator value as in **D2**. With this type of visualization stations (and especially the outlying ones) are more easily identified. A green/orange/red light then indicates whether the model performs well enough with respect to the performance criteria for 90%, between 75 and 90% or for less than 75% of the available stations (note that this choice of 75% as intermediate fraction is arbitrary).

For spatial statistics one value is attributed corresponding to the spatial correlation (or the value of the spatial NMSD) calculated on 100% of the available stations.

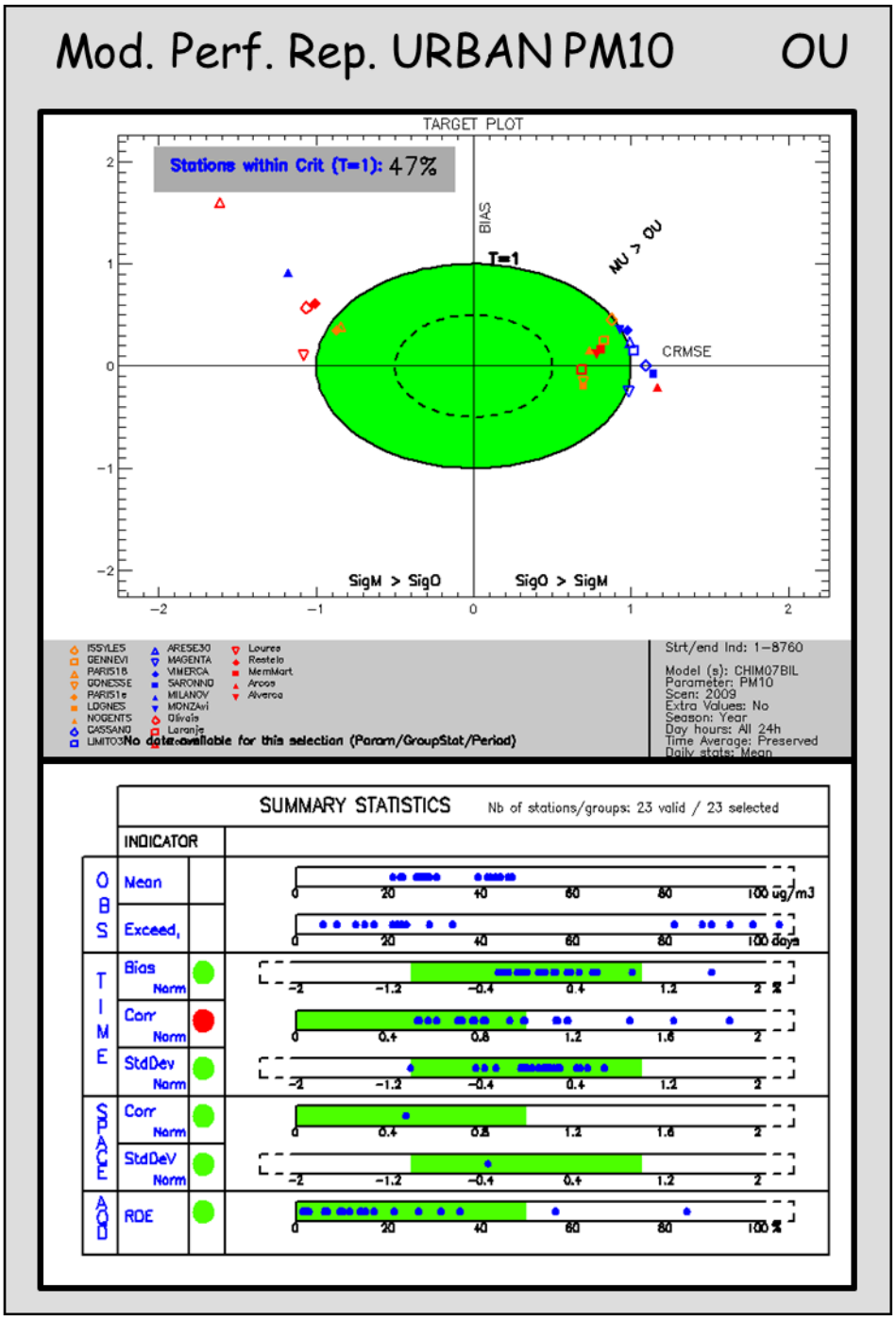



**Figure 4: Modified template based on normalization by observation uncertainty**

### 3.2.3 Representation of the observation uncertainty in the Target plot

In **D2** the proposed Target diagram followed the original work proposed by Jolliff et al. (2009). The distance between the origin and a given point (representing the performance of a model for a given station) is equal to the RMSE normalised by the standard deviation of the observations $\sigma_o$. The performance criterion was set to one, i.e.

$$\text{Target} = \frac{\text{RMSE}}{\sigma_O} = \frac{\sqrt{\sum (O_i - M_i)^2}}{\sqrt{\sum (O_i - \overline{O})^2}} < 1 \qquad (1)$$

where the sum over *i* represents the sum of daily or hourly measurements in the time series. With this normalisation a "less or equal to one" radius has the following meaning in terms of model performances:

a. The normalized bias and normalized CRMSE are less or equal to 1
b. Model and observations data are positively correlated
c. *A better than average model efficiency is achieved.* This means that the model is a better predictor of the monitoring data compared to the mean of the monitoring data (Stow et al., 2009). This can be easily seen by substituting $M_i = \overline{O} \quad \forall i$ in formula (1).

As stated by Jolliff et al. (2009), *"it is possible to include in the Target diagram some information on uncertainty. If the average relative observational uncertainty ($U_r$) is expressed as a percentage, then $U_r\overline{O}$ may be used as an estimation of the average value of uncertainty for the time series"*. The Authors then use $U=U_r\overline{O}/\sigma_O$ as an estimate of the normalized observation uncertainty in the Target diagram. For example a value of $U_r$=±15% and an average observed value of 50ppb would yield an average uncertainty of ±7.5 ppb. If the model to observations RMSE is smaller than this value of 7.5ppb the model result *"is within the observational uncertainty range and further improvement may not be meaningful"*.

This formulation has some limitations: every station would have its own $U_r\overline{O}/\sigma_O$ as it depends on the station standard deviation and average concentration level of each station. To represent the observation uncertainty in the Target diagram for more than one station at a time an averaged value for $\bar{O}/\sigma_o$ need to be selected. This will be valid only if this average value can be considered as representative for all stations.

## A new normalization for the Target plot

We propose here a new definition for the Target indicator, with the aim to account for observation uncertainty in a more generalised way. **The main motivation is to build an indicator which qualifies model performance in terms of the observation uncertainty**.
Let's illustrate this with an example. Figure 5 shows the model and observed PM10 time series at a monitoring station. We have assumed the relative observation uncertainty $U_r$ to be 25% which is the maximum value set in the AQD for PM10 (highlighted in cyan). We then request as a model quality criterion that for every time step *i* the model result has at maximum a similar absolute uncertainty, i.e. $U_rO_i$ (highlighted in light red).

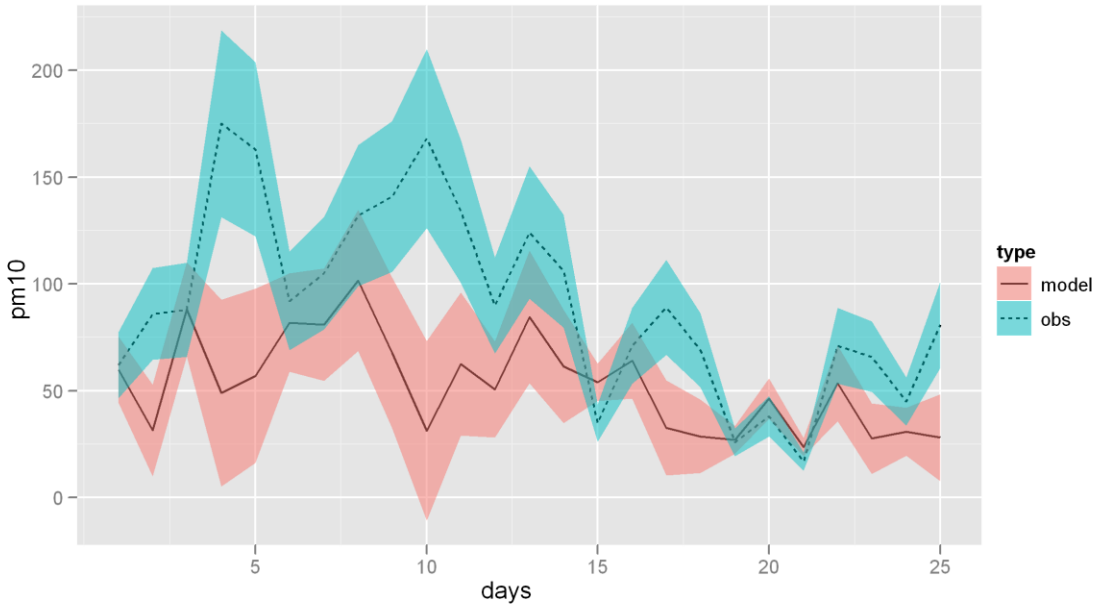



**Figure 5: Example of a station PM10 time series (measured and modeled), each plotted with the same absolute error $U_rO(i)$ with $U_r$ =25% (shaded areas).**

Each day in the time series corresponds to one of the following cases:

1) **Model results are within the range of observation uncertainty** (e.g. Day 3 or 6 in Figure 5). Model and observations are then distant by less than $U_rO_i$;
2) Observations and model uncertainties ranges overlap (e.g. day 13): the model to observation distance is between $U_rO_i$ and 2 $U_rO_i$ which means that the **model still might be closer to the "true value" than the observation**;
3) Observations and model uncertainties ranges do not overlap: model and observation are further distant than 2 $U_rO_i$. **Observation is closer to the "true value" than the model result**.

In order to include these considerations within the Target indicator and diagram we substitute formula (1) with:

$$\text{Target} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}}{2U} < 1 \qquad (2)$$

where $U$ is the observation uncertainty. In general $U$ for a temporal series of N measurements is a function of the concentration level and can be expressed as:

$$\text{U} = \frac{\sum_{i=1}^{N} U_r(O_i) * O_i}{N} \qquad (3)$$

where $U_r$ is the relative uncertainty for a given concentration level and for a given species. In the remaining of the text, as in Figure 5, we will consider in first approximation that $U_r$ is independent of the concentration level and is set to the Data Quality Objective (DQO) reported in the AQD.

If we use CRMSE/2U and BIAS/2U as X and Y axis in the Target diagram, the three cases mentioned above respectively translate as follows:

1) Target $\leq$ 0.5. In this case the normalized RMSE between the observed and modeled values is less than the observation uncertainty. Model results are in average within the range of the observation uncertainty for that station and it is meaningless to further improve model performances.

2) 0.5<Target≤1. In this case the normalized RMSE between observed and modeled values is in average larger than the range of observation uncertainty but the model might still be a better predictor of the "true value" than observations.

3) Target>1. In this case model results are further away from the "true value" than observations.

As mentioned above we have assumed here that $U_r$ is equal to the maximum value allowed in the AQD, i.e. 25% for PM and 15% for $O_3$ and $NO_2$ for the whole range of concentration levels, although DQO are set only around the limit value in the AQD. But the approach proposed here is flexible and would allow introducing more detailed information on observation uncertainty as it becomes available. For example the measurements performed in the frame of the AQUILA network (http://ies.jrc.ec.europa.eu/aquila-homepage.html) could be used to build a concentration dependency in definition (3) resulting in higher uncertainty at low concentration levels for some compounds (e.g. PM). This would then allow for a larger tolerance margin on model results at the lower concentration range. Other source of uncertainties than instrumental, e.g. the uncertainty resulting from the lack of representativeness of monitoring stations might as well be considered with this formulation.

Other advantages of this approach are:

- Although its formulation simplifies, the same indicator (e.g. Target) can be used for models delivering hourly, daily or annual average values (See Annexes)

- The formulation could be adapted for approaches combining measurements and modeling results. The Target circles (0.5 and 1) correspond to the observation and model uncertainty respectively. The largest the fraction of observations used for model calibration, the more stringent the performance criteria becomes (closer to 0.5).

# 4 Performance criteria for other indicators

In **D2** model performances were proposed to be assessed in terms of both criteria and goals. Performance criteria indicate the level of accuracy considered to be acceptable for regulatory applications while performance goals would represent the maximum level of accuracy a model would be expected to achieve for a given application. In **D2** a tentative set of criteria has been proposed while goals had arbitrarily been fixed to a 20% more stringent level due to the lack of information to fix them more precisely. Since setting adequate and coherent values for performance criteria already is a challenge, we have dropped here (temporarily) the goals from the target diagram and summary statistics table.

With the new formulation based on observation uncertainty, the scale or pollutant dependency is included in the normalization factor **and the Target performance criterion is always unity regardless of the pollutant or scale considered**.

As mentioned earlier the core set of statistical indicators is meant to provide additional and complementary information to the Target indicator. We develop here below an approach to set performance criteria for the correlation, bias and standard deviation, which are consistent with the performance criterion set for the Target indicator.

We start from two basic equations (Jolliff et al., 2009) which relate statistical indicators among themselves.

$$\mathrm{RMSE}^2 = \mathrm{CRMSE}^2 + \mathrm{BIAS}^2 \quad (4)$$

$$\mathrm{CRMSE}^2 = {\sigma_0}^2 + {\sigma_M}^2 - 2\sigma_0\sigma_M R \quad (5)$$

These two equations can be used to relate the Target criterion (RMSE/2U < 1) to corresponding criteria for R, NMB and NMSD.

Three cases are analyzed below to derive performance criteria for R, NMB, NMSD.

**Case 1: perfect correlation (R=1), perfect $\sigma_M$ ($=\sigma_O$). Identification of a performance criteria for NMB:**

In this case: CRMSE=0 and from equation (4) the Target criteria becomes:

$$\left(\frac{RMSE}{2U}\right)^2 = \left(\frac{BIAS}{2U}\right)^2 = \left(\frac{BIAS}{2U_r\overline{O}}\right)^2 < 1 \Rightarrow |NMB| < 2U_r \quad (6)$$

In which we assumed that $U_r$ is independent of the concentration level. In the case of PM10 for which $U_r$ is required to be 25% around the limit value, the NMB would then become 50%. This value is close to the 60% value proposed by Boylan and Russell

(2006) for MFB. Similarly for $O_3$ or $NO_2$ the required uncertainty value of 15% would lead to a NMB of 30%, close to the value proposed in Chemel et al. (2010).

**Case 2: no bias, perfect correlation (R=1). Identification of a performance criteria for NMSD:**

In this case RMSE=CRMSE and the criterion for NMSD is obtained from Eq. (**5**) with R=1 as follows:

$$\frac{\mathrm{CRMSE}^2}{(2U)^2} = \frac{{\sigma_0}^2 + {\sigma_M}^2 - 2\sigma_0\sigma_M}{(2U)^2} < 1 \Rightarrow (\sigma_0 - \sigma_M)^2 < (2U)^2 \Rightarrow$$

$$|NMSD| = \left|\frac{(\sigma_M - \sigma_O)}{\sigma_O}\right| < 2\frac{U}{\sigma_O} \quad (7)$$

As seen from this formula the performance criteria for NMSD (consistent with the Target indicator) depends on the ratio $U/\sigma_0$ . This ratio is a comparison between the absolute observation uncertainty and the observed standard deviation. The higher this ratio becomes (i.e. low observed standard deviation or large observation uncertainty) the less stringent the performance criteria becomes.

**Case 3: no bias, perfect $\sigma_M$ ($\sigma_M$=$\sigma_O$). Identification of a performance criteria for R:**

In this case RMSE=CRMSE and the criteria for R is obtained from Eq. (5) with $\sigma_M$=$\sigma_O$ as follows:

$$\frac{\mathrm{CRMSE}^2}{(2U)^2} = \frac{2{\sigma_O}^2 - 2{\sigma_O}^2 R}{(2U)^2} < 1 \Rightarrow R > 1 - 2\left(\frac{U}{\sigma_O}\right)^2 \quad (8)$$

As for NMSD, the criterion for correlation is function of the ratio $U/\sigma_O$ (but squared) and the largest the uncertainty is, the less stringent the performance criterion becomes. In the extreme case of $U$ equivalent to the observed standard deviation, the performance criterion for R becomes R>-1 (everything is then allowed for model results due to the fact that the observed variations are in average not distinguishable from the observed uncertainty).

Values for these performance criteria based on the AIRBASE monitoring data for daily PM10, hourly $NO_2$ and 8h daily maximum $O_3$ in 2009 are shown in Table 1, with $U_r$ set to the AQD DQO values (25% for PM10, 15% for both $NO_2$ and $O_3$). For each pollutant and station type, the mean value for the performance criteria obtained over all available stations is given. For PM10 the performance criteria show little variation with respect to the station type whereas for $O_3$ significant changes are visible. For $NO_2$ the performance criteria are relatively stringent as a consequence of the relatively small observation

uncertainty and of the large observation standard deviation. It must be noted however that these criteria values will be modified when the concentration dependency or the impact of the station representativeness will be included in $U_r$.

| Pollutant | Station type | Performance criteria | | |
|---|---|---|---|---|
| | | **NMB<** | **R>** | **NMSD<** |
| Daily PM10 | Rural bckg. | 50% | 0.55 | 0.88 |
| | Urban bckg. | 50% | 0.58 | 0.91 |
| | Traffic | 50% | 0.54 | 0.95 |
| | | | | |
| 8h-max $O_3$ | Rural bckg. | 30% | 0.59 | 0.89 |
| | Urban bckg. | 30% | 0.73 | 0.70 |
| | Traffic | 30% | 0.86 | 0.72 |
| | | | | |
| Hourly $NO_2$ | Rural bckg. | 30% | 0.93 | 0.39 |
| | Urban bckg. | 30% | 0.89 | 0.45 |
| | Traffic | 30% | 0.85 | 0.54 |

**Table 1: Performance criteria for specific statistical performance indicators**

Since the performance criteria for R, NMSD and NMB are station and time dependent (through $\sigma_O$), we also define normalized criteria from Eq.(6), (7) and (8) as follows:

Bias $$\left|\left(\overline{M}-\overline{O}\right)/2U\right|<1 \qquad (9)$$

Correlation $$\left(1-R\right)\Big/ 2\left(\frac{U}{\sigma_O}\right)^2<1 \qquad (10)$$

Standard deviation $$\left|\left(\sigma_M-\sigma_O\right)/2U\right|<1 \qquad (11)$$

Relations (9) to (11) are used in the summary statistic table as main performance criteria in DELTA V2.0. The performance criteria for R, NMB and NMSD represent necessary but not sufficient conditions to ensure that the Target main criterion (based on RMSE) is fulfilled. They are proposed here to indicate the aspects of the modeling application which need to be improved.

One of the main advantages of this approach is to allow the setting of a consistent set of performance criteria for different indicators based on one single input: the observation uncertainty. Equations (9) to (11) provide a single value for the performance criteria regardless of the type of stations, compounds… dependencies which are implicitly accounted for in the normalization factor. More detailed information on monitoring uncertainty (dependency on concentration level, estimation of the systematic and random components…) can then progressively be built in to provide more realistic values for the performance criteria. In the next section a summary of key diagrams, key indicators and associated performance criteria is provided.

## 4.1 Indicators, diagrams and performance criteria in DELTA

In the previous section we related in a consistent manner three main statistical indicators to the Target indicator based on RMSE: NMB, NMSD and R. In this section we adapt existing diagrams to associate to each indicator a specific diagram on which performance criteria can be visualized. Figure 6 relates the main indicators to the type of diagram used to highlight model performances.

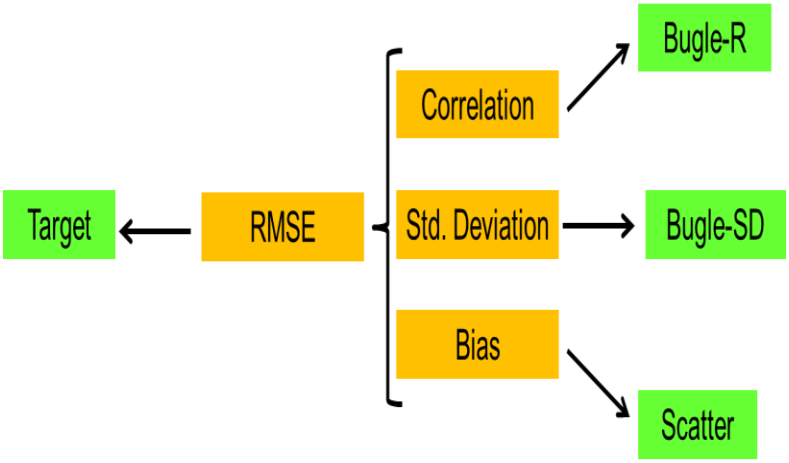
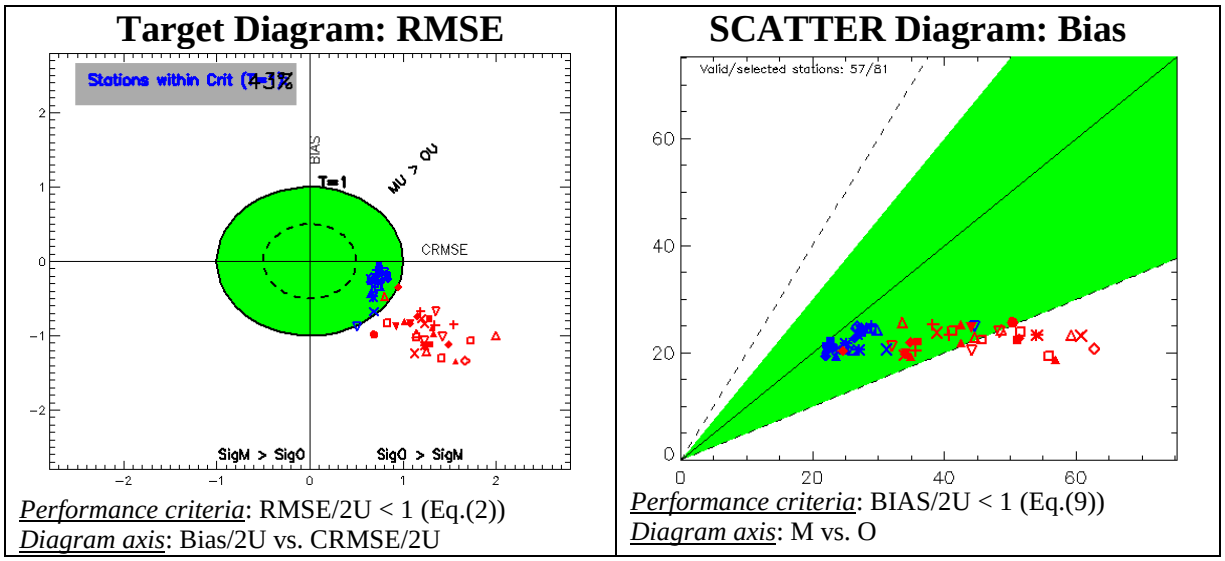



**Figure 6: Schematic overview of main diagrams and related indicators in DELTA**

For most of these diagrams, a change has been required to account for the normalization by the observation uncertainty. Figure 7 provides an example of each of the five diagrams together with their new axis normalization and associated performance criteria.




*Performance criteria*: RMSE/2U < 1 (Eq.(2))
*Diagram axis*: Bias/2U vs. CRMSE/2U

*Performance criteria*: BIAS/2U < 1 (Eq.(9))
*Diagram axis*: M vs. O

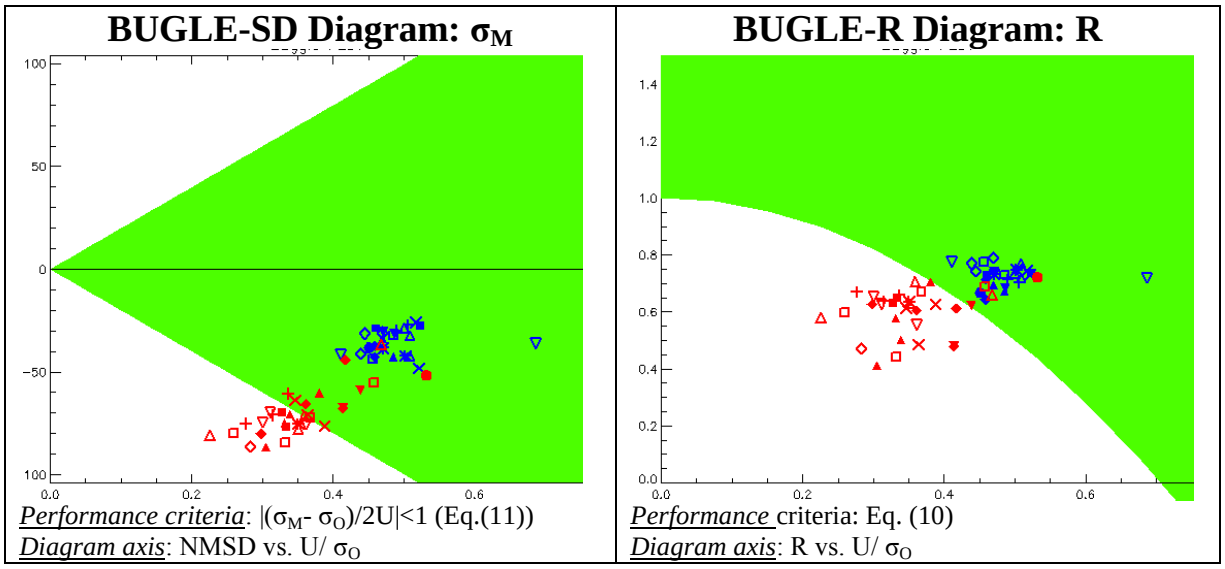



**Figure 7: Main diagrams (adapted from Target and Bugle) with associated performance criteria and fulfillment zones (in green) for Krakow (red) and Paris (blue) for the whole year 2009**

Other diagrams are available in DELTA to analyze model performances. The table below summarizes the main functionalities of each of them.

| **Diagram Name** | **Available Statistics** | **Representation** | **Performance criteria Available** |
|---|---|---|---|
| Target | RMSE, BIAS, CRMSE | Bias vs. CRMSE | Yes |
| Summary Report | BIAS, R, MNSD, RDE, Spatial R and MNSD | Performance Indicator value per station | Yes |
| Bugle | R, NMSD | R vs. U/ $\sigma_O$<br>NMSD vs. U/ $\sigma_O$ | Yes |
| Scatter | Bias | M vs. O | Yes |
| Category | Bias, R, NMSD | Performance Indicator value per station classified in terms of regions/category | Yes |
| Barplot | All | Indicator vs. stations | No |
| Time series | Raw values | Value vs.Time | No |
| Taylor | R, NMSD | σ vs. σ<br>σ/2U vs. σ/2U | No |
| Soccer | Bias, RMSE | RMSE vs. Bias | Yes (for meteo) |
| Quantile | Bias | M vs. O (sorted values) | |
| Google Earth | All | | No |

**Note**: In DELTA, diagrams which require information on observation uncertainty, i.e. the diagrams where performance criteria are available, are labeled with (OU). Those diagrams will work only if a value is set for the observation uncertainty in the configuration files. This is currently done only for PM10 daily averages, $O_3$ daily 8h max and $NO_2$ hourly values as well as for WS and TEMP.

## 4.2 Comparison of AQD MQO and proposed criteria

A comparison of the AQD criterion (i.e. RDE) with the new set of criteria is proposed in Table 2. For this we used the CHIMERE results performed over the whole Europe at 7 km resolution for a full year in 2009. The study focuses on 30 different urban areas over which the Airbase stations have been retrieved to perform the comparison. As seen from Table 2, for $NO_2$ and PM10 the new set of criteria is a more stringent condition to fulfill than RDE. While the results are good in terms of bias and standard deviation, they show too low correlations. On the contrary for $O_3$ the new set of criteria is a less stringent condition than RDE and most of the stations fulfill all criteria. It is important to note that this model application covers the whole of Europe and the analysis presented in Table 2 should be refined in terms of the spatial region to highlight the areas where the model performs well and those where improvements are required.

As mentioned earlier the correlation performance criterion for $NO_2$ is very stringent. CHIMERE with its current resolution does not capture well these effects and none of the stations fulfill the criteria indicating that finer resolution modeling is probably necessary to address $NO_2$ correctly. As $NO_2$ observations generally depend on the local environment of a given station, these concentrations do most probably vary significantly within a 7x7 $km^2$ resolution grid cell, even at background stations. It might be therefore important for $NO_2$ to broaden the meaning of the observation uncertainty beyond instrumental uncertainty and include station spatial representativness (in terms of model), as suggested in Section 3.2.3. In this way the $NO_2$ criteria would become less stringent and better reflect the limitations of modeling due to spatial resolution. This information on data representativeness might be less relevant for PM10 and $O_3$ characterised by smoother concentration fields and where a 7x7 $km^2$ resolution seems better suited.

| | PM10 | NO2 | O3 |
|---|---|---|---|
| Number of stations (urban – suburban – rural backgrounds) | 464 | 556 | 588 |
| Stations percentage fulfilling Target < 1 | 40% | 0% | 73% |
| Stations percentage fulfilling RDE < 50% | 94% | 96% | 56% |
| Stations percentage fulfilling R criteria | 46% | 0% | 96% |
| Stations percentage fulfilling NMB Criteria | 98% | 93% | 92% |
| Stations percentage fulfilling NMSD Criteria | 93% | 77% | 100% |

**Table 2: Percentage of stations fulfilling RDE criterion and the new performance criteria**

# 5 Delta Tool: Updates

In this section the main updates made to DELTA are presented briefly.

## *5.1 Batch mode - Benchmarking service*

Although the batch mode was already available in the first release (Version 1) this functionality was not very user friendly. In Version 2 it is menu-driven offering the possibility to the user to set-up procedures which will automatically process data in the chosen way. Figure 8 illustrates this set-up menu. One or more diagrams can be organized in an A4 format either in postscript or Bitmap format. This function has been used to create benchmarking procedures which are available from the top menu of the tool to automatically create postscript or bitmaps reports (as illustrated in Section 3 and 6). The reports for the following applications are available:

- Assessment $O_3$ (daily 8h max) for urban and local scales
- Assessment PM10 (daily and annual averages) for urban and local scales
- Assessment $NO_2$ (hourly and annual values) for urban and local scales

*Note*: the layout of the postscript reports is not yet fully developed and improvements on how information is displayed are yet required.

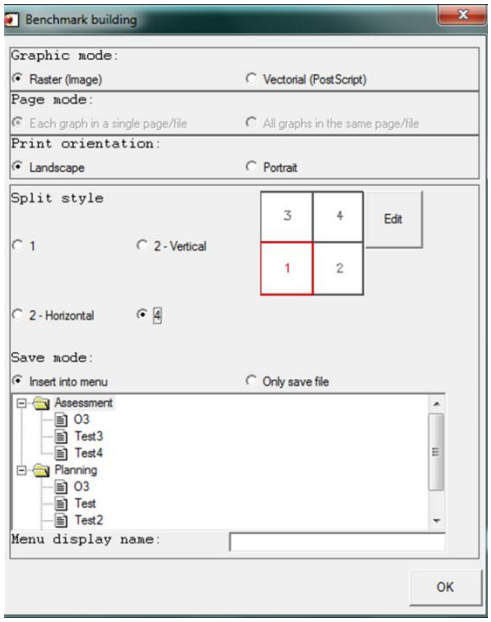



**Figure 8: Batch mode/ benchmarking menu widget**

## 5.2 Group vs. single observation mode

In version 2 of DELTA, the possibility of grouping stations has been added. Once these stations are grouped they will be treated as a single entity and can be compared either to other groups of stations or to a selection of single stations. This group mode option is activated for most of the DELTA diagrams. A choice is left to the user to decide how this group should be created, i.e. on which statistical basis it will be formed. Currently two options are offered:

- Statistical parameters are calculated for each station and the value assigned is the mean of all indicators.
- Statistical parameters are calculated for each station and the value assigned is the P90 value. This means we select the P90 worst value for the indicator. When a group built in this way lies within a green fulfillment zone, it means that 90% of the stations belonging to the group fulfill the criteria. This is the reason why fulfillment zones are visible only when this selection is activated.

Note that only one of these two possibilities can be activated at a time when more than one group is selected. The option selected for the last group will be automatically applied to all previous groups.

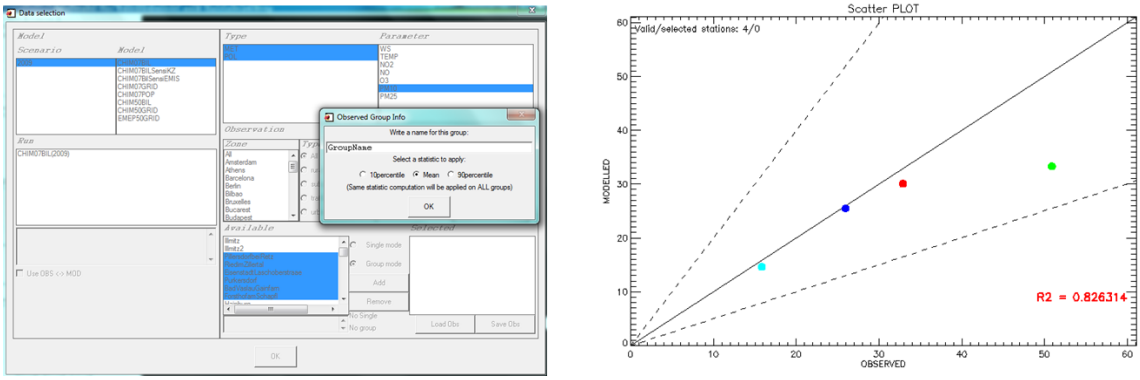



**Figure 9: Example of station/grouping setup and corresponding scatter diagram**

## 5.3 Google Earth application

The “geomap functionality” is now replaced by a graphical option based on Google Earth. Once the choice of the models, scenarios, parameters and stations is made, the Google Earth interface is launched and the selected stations can be visualized. More information about the selected stations and statistical performance indicators can be obtained by clicking on the stations. It must be noted that this interface will refresh itself according to any new selection. It remains also possible to run other DELTA tool diagrams and analysis in parallel.

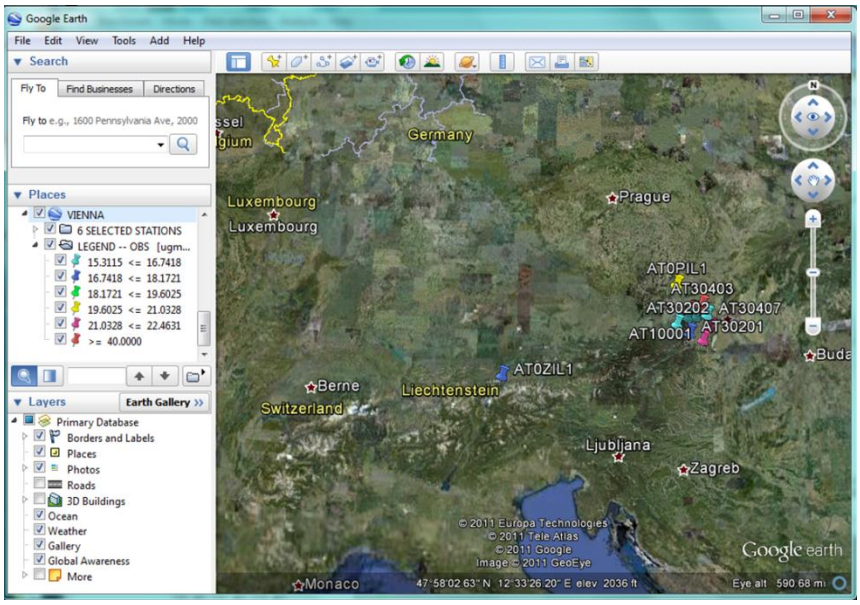



**Figure 10: Example of Google Earth interface**

## 5.4 On-click data information

This option offers the user the possibility of obtaining information about the values and reference (e.g. station name) for any of the points shown on available DELTA diagrams (e.g. scatter plots), interactively via the mouse.

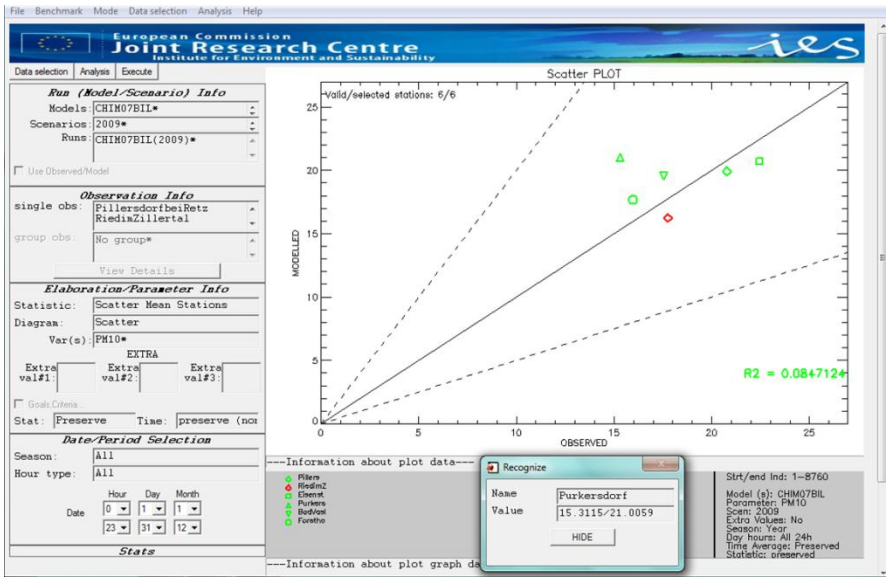



**Figure 11: Example of the mouse-recognition option in DELTA**

## 5.5 Category overview diagrams

This type of diagram provides an overview of the model performances for a given statistical indicator. Stations are classified according to the categories defined by the user in the configuration file (see **D1**). It allows the user to quickly identify in which of these regions the model performs better or worse. The diagram is now available for R, NMB and NMSD and a green zone indicates the criteria fulfillment area. It only works when observation criteria are provided (now available for PM10 daily averages, NO2 hourly values and O3 daily 8h max).

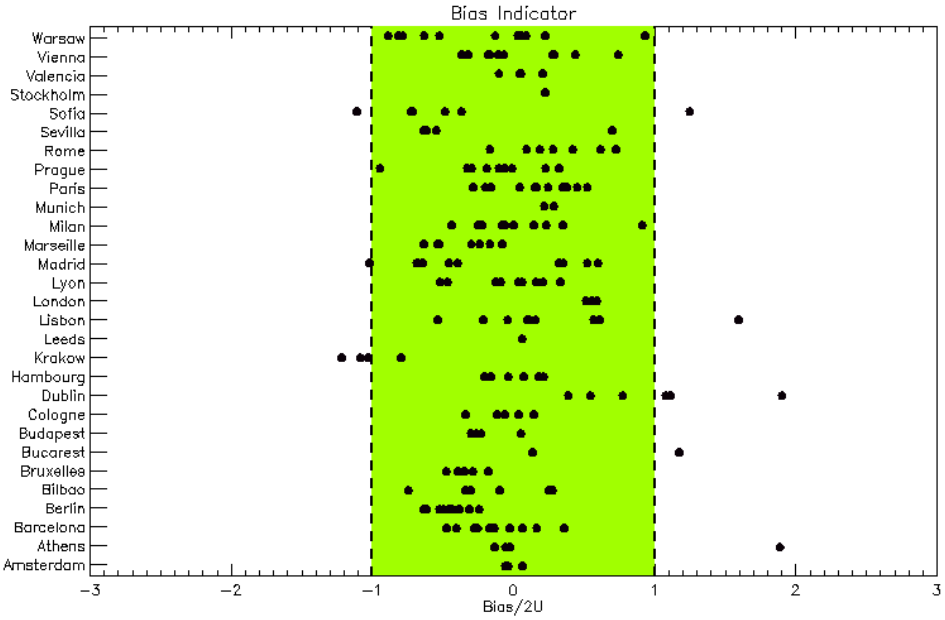



**Figure 12: Example of category diagram for the Mean Normalized Bias with all stations classified in terms of regions, EC4MACS dataset.**

## 5.6 Utilities

### 5.6.1 Pre-processor

The Delta pre-processor is an IDL-based tool for the extraction of time series at observational locations from meteorological or air quality model output for use in the DELTA Tool. Input to the pre-processor is the configuration file 'startup.ini' containing the variables (meteorological variables, and pollutants) to be treated, as well as geographical information about the observational stations. Model output should be in netCDF format with all the variables defined on longitude-latitude coordinates at ground level and hourly frequency. Three interpolation techniques are available for producing the modeled time series at the observational stations:

(i) **NN** (Nearest Neighbour) where the values at a station are taken from the nearest lon-lat grid point.

(ii) **BIL** (Bilinear) where a bilinear interpolation is performed on the grid cell in which the station is located; for this the gridcell is first transformed into a square using a bilinear mapping.

(iii) **DW** (Distance Weighted) where a weighted mean value is calculation in the station grid-cell. The weights are the inverse of the distance from the station to the 4 gridpoints.

Output of the pre-processor is written to a netCDF file.

During the pre-processing a number of checks are performed to guarantee the conformity with the DELTA Tool conventions.

The DeltaPreProcessor is available as an idl-executable (.sav file) and runs under the IDL-Virtual Michine in a Windows environment.

### 5.6.2 Input/Output checking program

Check_IO is an IDL-based tool which checks the consistency among the modeling results file (NetCdf), the observation files (.csv) and the main configuration file (startup.ini)
Inputs to Check_IO are:

i. the configuration file 'startup.ini' which contains the variables (meteorological variables , and pollutants) and metadata of observational stations,

ii. the observational data in *.csv format for each station, and the model time-series input in netCDF format for each station.

The 5 main steps in Check_IO are:

1) General initial check on the existence of the directories for input files to the DeltaTool.
2) “Startup.ini” related checks on the existence of the startup.ini file and its contents, as well as a first check on the variables in startup.ini and the variables in the observational data.
3) Observations related check on the compatibility of the station names in startup.ini and the observational data, including, correct number of variables, length of the observational time series, availability of observational data, and extreme values.
4) Model related checks on the existence of the modelling time series at the stations for all the variables requested in startup.ini, and a check on NaN, Inf and extreme values.
5) Observations & Model related checks on the availability and compatibility of model time series at the observational stations, with calculation of minValues, maxValues and meanValues for the observations and model time series

Check_IO produces a log report, as well as a summary report with details concerning the various checks.The Check_IO utility is available as an idl-executable (sav fle) and runs under the IDL-Virtual Michine in a Windows environment.

# 6 Generalization of the report template

## *6.1 Statistical models - Annual Averages*

The report presented in Figure 4 is designed for hourly or daily modeled values. A request made in SG4 was to adapt the performance reporting to include models producing annual values only. A proposal is made here. For annual values some statistical indicators and diagrams lose their meaning. This is the case of the Target diagram but also of all temporal indicators which were associated to the analysis of time series (e.g. correlation). In the case of annual averages, the performance criterion reduces to:

$$\text{Target} = \frac{RMSE}{2U} = \frac{|\text{BIAS}|}{2U} < 1 \qquad (12)$$

The interpretation made for hourly or daily values above remains valid, i.e. the Target indicator must remain below 1 with values larger than one indicating more uncertainty in the model results than in the observations. In order to graphically represent this quantity, we use the scatter plot (and do not consider the absolute value in (12))

The observation uncertainty (U) can be decomposed into a random and systematic components (ISO11222, 2002). While the systematic part remains unchanged regardless of the averaging time period this is not the case of the random part which tends to decrease with longer time averaging periods. As a result the value of the observation uncertainty U introduced in (12) for annual values is therefore significantly smaller than the one used in (2) for hourly or daily values.

Similarly to the Target diagram the percentage of available monitoring stations at which the model fulfills the performance criterion can be used as quality objective. Also for this definition the main input parameter in this formulation remains the observation uncertainty which needs here to be defined for annual averages.

For annual averages the summary statistics Table (Figure 4) is obviously simpler. The section dealing with observed mean and exceedance values drops because information on mean values is already available in the scatter plot. All statistics regarding time series disappear and only the spatial statistics are left (e.g. how well does the model perform in representing low/high values spatially) as well as some information on the quality objectives currently requested by the AQD.

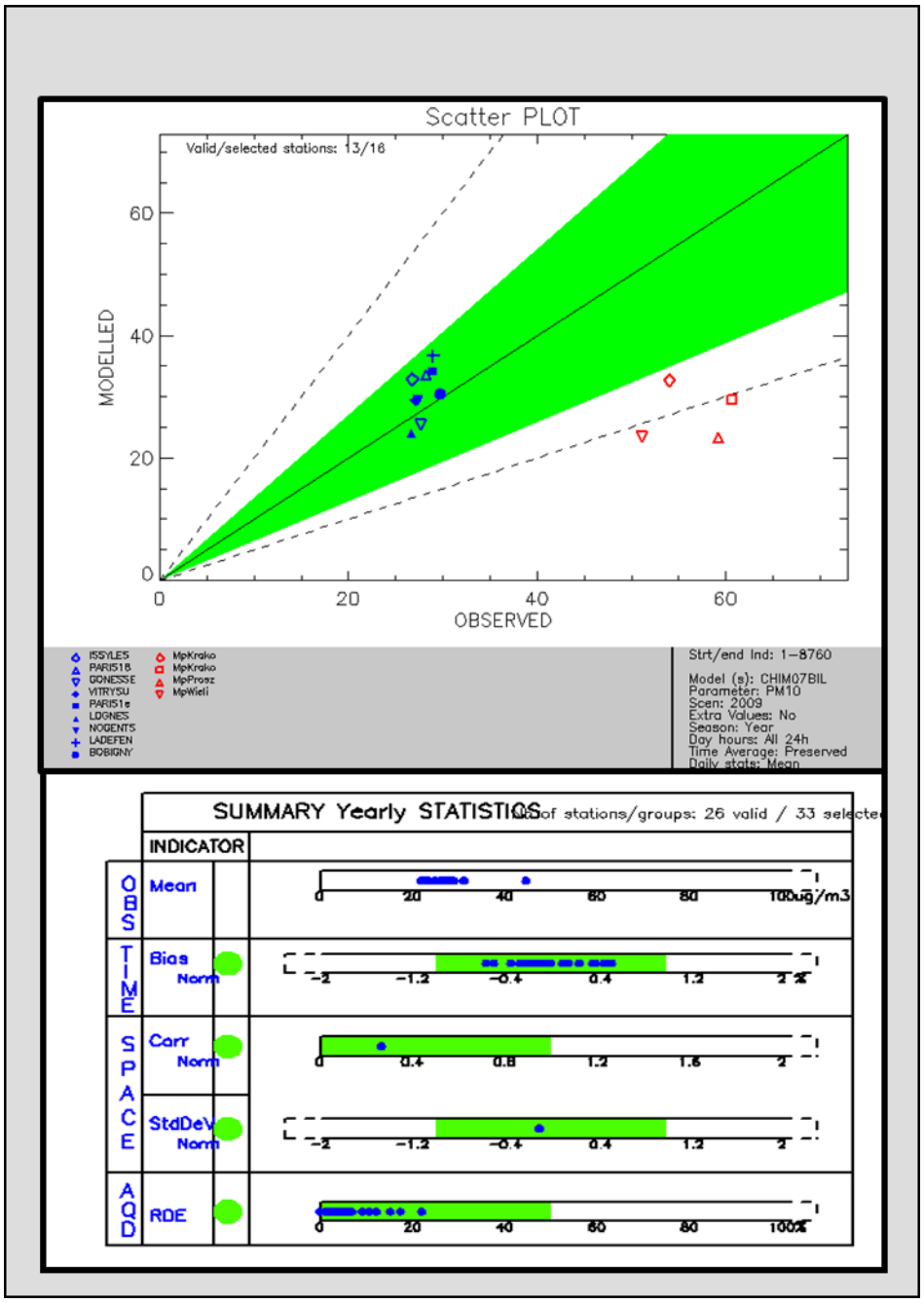



**Figure 13: Example of performance summary report for models reporting annual averages only (Paris & and Krakow data from the 2009 EC4MAX dataset). Performance criteria are adapted for long time averaging periods.**

## 6.2 Regional Scale

As mentioned in the previous section, the template for reporting model performances is structured similarly in terms of diagram and statistical indicator regardless of the spatial scale and pollutant considered. We propose in this section a performance template report adapted to regional scale applications. For this type of model application the number of stations becomes too large to represent stations individually. This is why group

functionality has been introduced in DELTA. This option enables the user to focus at one single statistics which summarizes performances for a given group of stations. Figure 14 refers to three groups of stations (e.g. Regional groups of countries or individual countries) for which the P90 statistics are calculated and plotted both in the Target diagram and summary statistics table. As an example, if for region X the correlation coefficients for the 10 available stations are (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0), we only plot the P90 (0.2) in the diagram and summary table. It is then easy to identify regions which show potential problems. Of course some further analysis will be required to understand how stations behave within one given group of stations. For the spatial statistics proposed in the summary table one value is obtained as the average of all stations belonging to a given group and represented directly as such. The red/orange/green flag present in Figure 4 disappears for regional applications as we directly know if the group of stations has a potential problem from the fact that it is within the acceptance area (performance criteria fulfilled, in green)

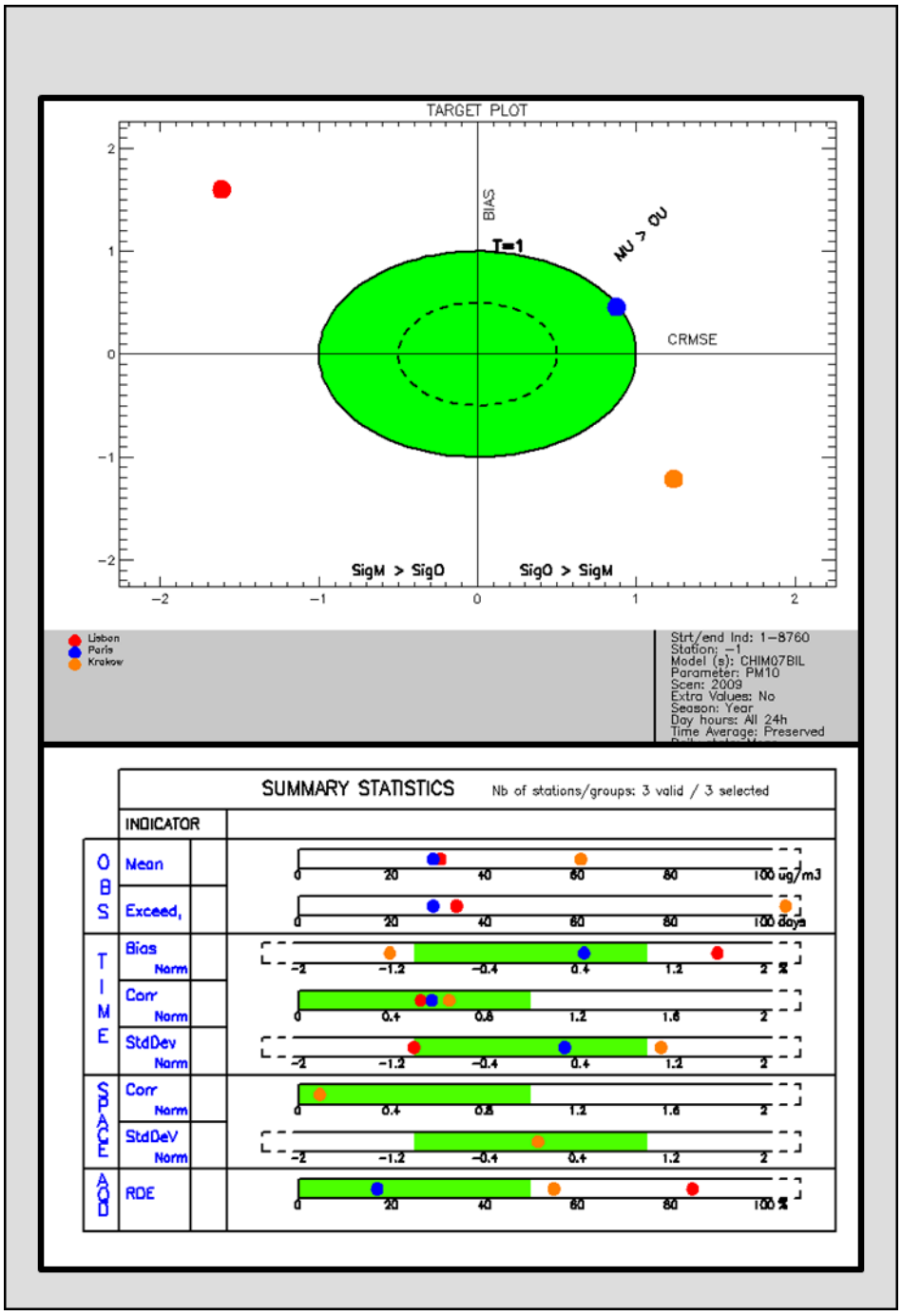



**Figure 14: Example of performance report for models operating at the regional scale**

# 7 Annexes

## *7.1 Forecasting application*

Although SG4 does not deal with forecasting applications, collaboration has been established with the Pasodoble (http://www.myair-eu.org/) and Atmosys (http://www.life-atmosys.be/EN/Home/Pages/default.aspx) EU projects which have their focus on forecasting. In these two projects, some benchmarking is planned which will require a template similar to the one developed in FAIRMODE to proceed with the evaluation of models in a systematic and consistent manner. In this section we propose a possible adaptation of the FAIRMODE report to make it more useful for forecasting applications.

As we are here more interested in the model capabilities to deliver accurate daily model forecasts, the Target indicators could be modified to consider these aspects. In place of using a normalization based on the standard deviation or observation uncertainty, another possibility is to normalize by a quantity representative of the day-to-day variations, i.e.:

$$\text{Target} = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right)^2}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_{i-1} - O_i\right)^2}}$$

where N is the length of the time series. In such case the Target indicator becomes one when the analyzed model forecast is as good as a *persistent model* (i.e. a model, which for day *i* forecasts day *i-1* value). Similarly to the original Target value, the positive and negative portions of the X axis could be adapted to provide additional information relative to forecast applications (e.g. the "odds ratio skill score").

The summary statistics table can be updated accordingly by selecting the performance indicators which are of key relevance for forecasting applications.

# 8 References


AQD, 2008. Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on ambient air quality and cleaner air for Europe ( No. 152), Official Journal.

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.

Boylan, J.W., Russell, A.G., 2006. PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models. Atmospheric Environment 40, 4946–4959.

Chemel, C., Sokhi, R.S., Yu, Y., Hayman, G.D., Vincent, K.J., Dore, A.J., Tang, Y.S., Prain, H.D., Fisher, B.E.A., 2010. Evaluation of a CMAQ simulation at high resolution over the UK for the calendar year 2003. Atmospheric Environment 44, 2927–2939.

Denby, B., 2010. Guidance on the use of models for the European Air Quality Directive. A working document of the Forum for Air Quality Modelling in Europe FAIRMODE (ETC/ACC, version 6.2).

ISO11222, 2002. Air quality - Determination of the uncertainty of the time average of air quality measurements. International Organization for Standardization.

Jolliff, J.K., Kindle, J.C., Shulman, I., Penta, B., Friedrichs, M.A.M., Helber, R., Arnone, R.A., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. Journal of Marine Systems 76, 64–82.

Stern, R., Flemming, J., 2004. Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU directives for air quality - examples using the chemical transport model REM-CALGRID. Freie Universitat Berlin, Institut for Meteorologie, Berlin.

Stow, C.A., Jolliff, J., McGillicuddy, D.J., Doney, S.C., Allen, J.I., Friedrichs, M.A.M., Rose, K.A., Wallhead, P., 2009. Skill assessment for coupled biological/physical models of marine systems. Journal of Marine Systems 76, 4–15.

Thunis, P., Emilia Georgieva, Anna Pederzoli, 2011. The DELTA tool and Benchmarking Report template. Concepts and User guide, version 2.

Thunis, P., Emilia Georgieva, Stefano Galmarini, 2010. A procedure for air quality models benchmarking (Version 1).