Atmospheric Environment 59 (2012) 476–482

Contents lists available at SciVerse ScienceDirect

Atmospheric Environment

journal homepage: www.elsevier.com/locate/atmosenv

Performance criteria to evaluate air quality modeling applications

P. Thunis [a,*], A. Pederzoli [b], D. Pernigotti [a]

[a] European Commission, JRC, Institute for Environment and Sustainability, Climate Change and Air Quality Unit, Via E. Fermi 2749, 21027 Ispra (VA), Italy
[b] University of Brescia, DII-Dipartimento di Ingegneria dell'Informazione, Via Branze 38, 25123 Brescia, Italy

HIGHLIGHTS

- New method to evaluate air quality models based on observation uncertainty.
- The same margin of tolerance is allowed for models as for observations.
- New criteria and diagrams to visualize fulfillment zones are described.
- Various dependencies (pollutant, concentrations level…) can be considered.
- Example diagrams and criteria values from measurements are provided.

ARTICLE INFO

Article history:
Received 14 March 2012
Received in revised form
22 May 2012
Accepted 25 May 2012

Keywords:
Statistical model evaluation
Performance criteria
Air quality modeling
Quality objectives
Observation uncertainty

ABSTRACT

A set of statistical indicators fit for air quality model evaluation is selected based on experience and literature: The Root Mean Square Error (RMSE), the bias, the Standard Deviation (SD) and the correlation coefficient (R). Among these the RMSE is proposed as the key one for the description of the model skill. 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 for model results a similar margin of tolerance (in terms of uncertainty) as for observations. U is pollutant, concentration level and station dependent, therefore the proposed MPC are normalized by U. Some composite diagrams are adapted or introduced to visualize model performance in terms of the proposed MPC and are illustrated in a real modeling application. The Target diagram, used to visualize the RMSE, is adapted with a new normalization on its axis, while complementary diagrams are proposed. In this first application the dependence of U on concentrations level is ignored, and an assumption on the pollutant dependent relative error is made. The advantages of this new approach are finally described.

1. Introduction

Air quality models are powerful tools for the assessment and forecast of pollutants concentration in the atmosphere. As models are increasingly used for policy support their evaluation is becoming an important issue which is addressed in several documents published by policy-making authorities, e.g. the Environmental Model Guidance document of the US-EPA (EPA, 2009), the UK-DEFRA report (Derwent et al., 2010), the Guidance on the use of models for the European Air Quality Directive 2008 (Denby, 2010) or the ASTM standard D6589 (2005).

Model evaluation is in general a complex procedure (Chang and Hanna, 2004) involving different steps (scientific evaluation, code verification, model validation, sensitivity analysis etc.). 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 (Dennis et al., 2010) 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. Although the comparison between modeled and observed concentrations cannot give a thorough insight into the properties of the model, it is seen as a good first step in the evaluation of model performance (Derwent et al., 2010; Irwin et al., 2008).

In order to perform a statistical performance analysis a wide variety of indexes can be found in literature (Chang and Hanna, 2004; Shluenzen and Sokhi, 2008; EPA, 2007, 2009; Boylan and Russell, 2006; Jolliff et al., 2009; Borrego et al., 2008; Denby, 2010) which have been proposed for different fields of application (meteorology, air quality, hydrology), different goals (forecast, study of specific episodes) or different types of application. And it is

* Corresponding author. Tel.: +39 0332785670; fax: +39 0332786574.
E-mail address: philippe.thunis@jrc.ec.europa.eu (P. Thunis).

1352-2310/$ – see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.atmosenv.2012.05.043

generally recommended to apply multiple performance indicators regardless of the model application since each one has its advantages and disadvantages. Based on a review of the literature and the scope of our analysis, indicators to describe the Root Mean Square Error, the correlation, the standard deviation and the bias have been selected in this work.

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 by Boylan and Russell (2006) as the minimum level of quality to be achieved by a model for policy use, need to be fixed. MPC have been proposed for various statistical indicators and for different types of applications. Boylan and Russell (2006) recommend the use of the Mean Fractional Bias (MFB) and Mean Fractional Error (MFE) indicators and provide a set of MPC for these indicators for Particulate Matter. Chemel et al. (2010) extended this approach to O_3 concentration. Derwent et al. (2010) also proposed MPC for the bias and factor of 2 (FAC2) statistical indicators. The Air Quality Directive (AQD, 2008) also suggests a quality objective that models should fulfill (Denby, 2010). This indicator, referred to as the Relative Directive Error (RDE) is based on the model-observed difference around the limit value defined by law, regardless of the timing of the events. An MPC of 50% is required in the AQD (2008). Flemming and Stern (2007) proposed an alternative interpretation to the 50% uncertainty referred in the AQD2008, based on percentiles (Relative Percentile Error, RPE). Finally Jolliff et al. (2009) propose to use the RMSE normalized by the standard deviation of the observations (σ_O) as a statistical indicator:

$$\frac{\mathrm{RMSE}}{\sigma_O} = \frac{\sqrt{\sum_{i=1}^{N}(O_i - M_i)^2}}{\sqrt{\sum_{i=1}^{N}(O_i - \overline{O})^2}},$$

where the letters O and M stand for observations and model results respectively, the subscript i indicates the time step, the overbars indicate the time average over N time intervals, while the symbol σ indicates the standard deviation. A "less or equal to one" value for this indicator 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 all M_i by $\overline{O}$ in the above formula.

Currently MPC are proposed for a limited set of indicators and their values are not always consistent with each other since they depend on the type of application, the spatial scale and/or the time period selected. In this work we propose an approach to derive a consistent set of MPC for a selection of key statistical indicators. This derivation makes use of the observation uncertainty (U) as the main element and requests that model results have a similar margin of tolerance (in terms of uncertainty) as observations. Existing composite diagrams are then adapted to visualize model performance in terms of the proposed MPC. The normalized (by U) indicators/MPC/diagrams are then applied on a real modeling application for which MPC values are presented. The main motivation of this work is to support the evaluation of model performance for regulatory applications but the proposed methodology can also provide support to model developers in getting insight in particular model performance aspects.

2. Model performance criteria based on observation uncertainty

Within the framework of FAIRMODE (Forum for Air Quality Modeling in Europe, http://fairmode.ew.eea.europa.eu/) the issue of the benchmarking of air quality models is currently discussed (Thunis et al., 2011). The objective is to propose a methodology to evaluate model performance for policy applications, especially those related to the 2008 AQD. One of the outcomes of those discussions is a model performance report which provides a summary of the model strengths and weaknesses for a particular application. For the statistical analysis of model performance, the report is based on a core set of statistical indicators: Root Mean Square Error (RMSE), correlation coefficient (R), Normalized Mean Bias (NMB), Normalized Mean Standard Deviation (NMSD) and Centered Root Mean Square Error (CRMSE) defined as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}; \quad \mathrm{NMB} = \frac{\mathrm{Bias}}{\overline{O}} = \frac{\overline{M} - \overline{O}}{\overline{O}};$$

$$\mathrm{NMSD} = \frac{(\sigma_M - \sigma_O)}{\sigma_O}; \quad \mathrm{CRMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}((O_i - \overline{O}) - (M_i - \overline{M}))^2}$$

for which the formalism was defined in the previous paragraph.

These indicators which have been selected to cover the main aspects of the model performance (in terms of amplitude, phase and bias) are close to the set proposed by Borrego et al. (2008).

MPC need to be defined and set for each statistical indicator in order to answer the question: "are model results good enough for my purpose?" Although the AQD (2008) defines model quality objectives for the main pollutants their practical application remains free to interpretation (Denby, 2010). In this work we propose to define performance criteria for RMSE, NMB and NMSD based on the observation uncertainty U with the simple principle of allowing a similar margin of tolerance to both model and observations.

We define the observation uncertainty U as follows:

$$U = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(U_r(O_i)*O_i)^2}$$

where $U_r(O_i)$ denotes the relative uncertainty for a given concentration level and a given species. Since the main objective is to present the methodology and not to focus on the actual values obtained for the MPC, U_r is assumed to be independent of the concentration level and is set according to the data quality objective (DQO) value of the AQD, i.e. 15%, 15% and 25% for O_3, NO_2 and particulate matter with diameter under 10 μm (PM10), respectively. Note that these values are meant as maximum allowed around the limit value for each pollutant. Recent field studies have shown that they quite accurately represent the current state-of-the-art in monitoring. A recent instrument field inter-comparison study performed in 15 European countries (Lagler et al., 2011) has shown that 7% of the PM10 measurements did not fulfill the 25% DQO, whereas this number raised to 24% for particulate matter with diameter under 2.5 μm (PM2.5). While for O_3 measurements generally fulfill the DQO, Gerboles et al. (2003) have shown that in situation characterized by high NO/NO_x ratios, the NO_2 monitoring uncertainty might exceed the 15% DQO.

Fig. 1 illustrates this principle of equal tolerance for model and measurement for a model (solid line) and an observed (dashed line) time series of PM10 at a given station. The relative observation uncertainty U_r is equal to 25% for all values. At each time step we request model result to have a maximum absolute uncertainty (represented by the light red area) equal to U_rO_i, i.e. equal to the absolute observation uncertainty (represented in cyan). In the following the concept of "true value" will be used to indicate the result of a perfect measurement of the given parameter. Three different situations can be identified:

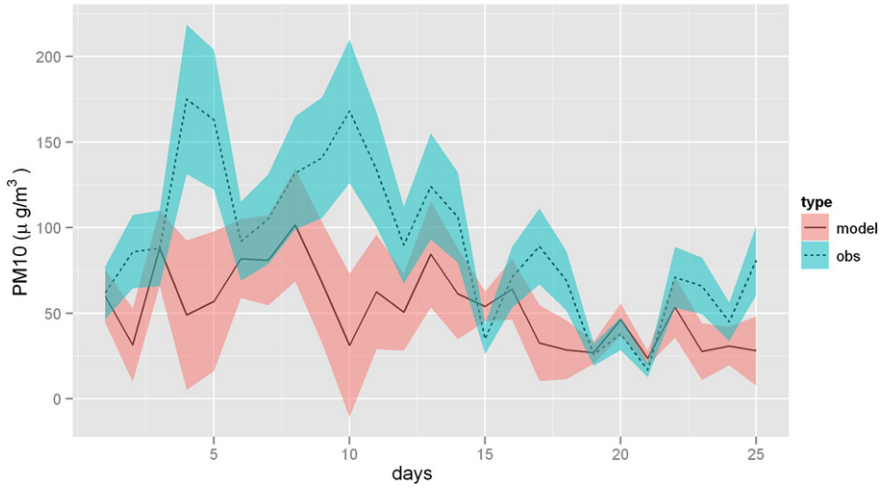

Fig. 1. Example PM10 time series (measured and modeled concentrations) for a single station, together with a colored area representative of the model and observed uncertainty ranges. We assumed here that the model uncertainty is equal to the absolute observation uncertainty, regardless of the modeled concentration level, i.e. $U_r = 25\%$.

a) Model results are within the range of observation uncertainty (e.g. days 3 and 6 in Fig. 1). Model and observations are then distant by less than U; it is then meaningless trying to further improve model performance;
b) Observations and model uncertainties ranges overlap (e.g. day 13): the model to observation distance is between U and $2U$ which means that the model might still be closer to the "true value" than the observation;
c) Observations and model uncertainties ranges do not overlap: model and observation are more than $2U$ apart. Observation is closer to the "true value" than the model result.

To account for the observation uncertainty in our statistical indicators and assign the MPC to unity we normalize the RMSE by $2U$ as follows:

$$\mathrm{RMSE_U} = \frac{\sqrt{\frac{1}{N}\sum(O_i - M_i)^2}}{2U}$$

With this formulation for the RMSE the error between observed and modeled values (numerator) is compared to the absolute uncertainty (denominator) and the three cases mentioned above translate into:

1) $\mathrm{RMSE_U} \leq 0.5$. This case corresponds to case (a) as the RMSE between observed and modeled values is less than U (as seen by substituting M_i by $O_i \pm U_j$ in the above equation). Model results are in average within the range of the observation uncertainty for that station and it is meaningless to further improve model performance.
2) $0.5 < \mathrm{RMSE_U} \leq 1$. This case corresponds to case (b) as the 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) $\mathrm{RMSE_U} > 1$. This case corresponds to the case (c) as model results are further away from the "true value" than observations.

As a general MPC we will therefore request that:

$$\mathrm{RMSE_U} = \frac{\mathrm{RMSE}}{2U} < 1 \qquad (1)$$

This approach is flexible as it allows introducing more detailed information on observation uncertainty as they become available. U might for example be made concentration level dependent (e.g. higher uncertainty at low concentration levels) or station dependent, allowing for a larger/smaller tolerance margin for model results under specific conditions. Other source of uncertainties than instrumental, e.g. the uncertainty resulting from the lack of representativeness of monitoring stations could be considered as well.

One of the drawbacks of the RMSE is that information about errors in either bias, σ_M and R is aggregated in a single number, as seen from formulas (2) and (3) in the next section. This is why a consistent set of MPC is derived in the next section for NMB, NMSD and R separately.

3. Performance criteria for complementary statistical indicators

With this formulation based on observation uncertainty, the criterion on $\mathrm{RMSE_U}$ is always unity regardless of the pollutant or scale considered (Eq. (1)).

As stated above, a set of additional statistical indicators (NMB, NMSD and R) has been chosen to detail various aspects of model performance. To derive MPC for these indicators the two basic equations which relate statistical indicators among themselves (Murphy, 1988; Jolliff et al., 2009) are used.

$$\mathrm{RMSE}^2 = \mathrm{CRMSE}^2 + \mathrm{Bias}^2 \qquad (2)$$

$$\mathrm{CRMSE}^2 = \sigma_O^2 + \sigma_M^2 - 2\sigma_O\sigma_M R \qquad (3)$$

Three cases are analyzed below to derive minimum performance criteria for NMB, R and NMSD, coherent with the criteria derived for RMSE. The cases presented are all hypothetical and represent extreme cases. The objective is to derive a set of minimum MPC to be fulfilled by a model, for a particular indicator, regardless of the particular application.

Case 1: $R = 1, \sigma_M = \sigma_O$: identification of an MPC for bias and NMB

In this case: CRMSE = 0 and from Eq. (2)

$$\frac{\mathrm{RMSE}^2}{(2U)^2} = \frac{\mathrm{Bias}^2}{(2U)^2} < 1 \Rightarrow |\mathrm{Bias}| < 2U \Rightarrow |\mathrm{NMB}| < \frac{2U}{\overline{O}} \qquad (4)$$

In the case of PM10 for which the observation relative uncertainty is required to be 25% around the limit value, the performance criteria for NMB would be slightly above 50%. This value is close to

the value of 60% proposed by Boylan and Russell (2006) for MFB. Similarly for O_3 or NO_2 concentrations the required uncertainty value of 15% would lead to an NMB slightly above 30%, close to the value proposed in literature (Chemel et al., 2010).

Case 2: bias = 0, $R = 1$: Identification of an MPC for NMSD

In this case RMSE = CRMSE and the criterion for NMSD is obtained from Eq. (3) with $R = 1$:

$$\frac{\mathrm{CRMSE}^2}{(2U)^2} = \frac{\sigma_O^2 + \sigma_M^2 - 2\sigma_O\sigma_M}{(2U)^2} < 1 \Rightarrow (\sigma_M - \sigma_O)^2 < (2U)^2 \Rightarrow |\mathrm{NMSD}| < \frac{2U}{\sigma_O} \quad (5)$$

As seen from this formula the MPC for NMSD depends on the ratio U/σ_O.

In this ratio the absolute observation uncertainty is compared to the standard deviation of the observations. The higher this ratio becomes (i.e. large U and/or low σ_O) the less stringent the performance criterion becomes.

Case 3: bias = 0, $\sigma_M = \sigma_O$: Identification of an MPC for R

In this case RMSE = CRMSE and the criteria for R is obtained from Eq. (2) with $\sigma_M = \sigma_O$. Eq. (3) then simplifies to:

$$\frac{\mathrm{CRMSE}^2}{(2U)^2} = \frac{2\sigma_O^2 - 2\sigma_O^2 R}{(2U)^2} < 1 \Rightarrow R > 1 - 2\left(\frac{U}{\sigma_O}\right)^2 \quad (6)$$

As for NMSD, the MPC for R is function of the ratio U/σ_O, but squared. Again the larger the uncertainty is the less stringent the performance criterion becomes. In the case of an absolute uncertainty being equivalent to the observation standard deviation, the performance criterion for R is equal to -1 (everything is then allowed for model results due to the fact that the observed variations cannot be discriminated on average from the observed uncertainty).

It is important to note that the MPC for NMB, NMSD and R represent necessary but not sufficient conditions to ensure that the main performance criterion based on RMSE is fulfilled. They are used here to indicate which aspects of the modeling application need to be improved. For example in the extreme case 1 (perfect R and perfect σ_M), an MPC on bias is obtained which must be interpreted as a threshold not be exceeded in any real case.

Since these performance criteria are station and time dependent (through σ_O), we also define normalized criteria from Eqs. (4)–(6) as follows:

$$\text{bias} \quad \frac{|\overline{M} - \overline{O}|}{2U} < 1 \quad (7)$$

$$\text{correlation} \quad \frac{(1-R)}{2}\left(\frac{\sigma_O}{U}\right)^2 < 1 \quad (8)$$

$$\text{standard deviation} \quad \frac{|\sigma_M - \sigma_O|}{2U} < 1 \quad (9)$$

One of the main advantages of this approach is to provide a selection of statistical indicators with a consistent set of performance criteria based on one single input: the observation uncertainty U. The main MPC based on the RMSE indicator provides a general overview of the model performance while the associated MPC for correlation, standard deviation and bias can be used to highlight which of the model performance aspects need to be improved. In the next section a summary of key diagrams, key indicators and associated performance criteria which can be used to visualize these MPC is provided.

4. MPC values as a function of pollutant, location, time and station type

Eqs. (7)–(9) show a dependency of the MPC for correlation, standard deviation and bias on the standard deviation and average value of the observation. These values are themselves depending on the compounds analyzed, the type of station selected and the geographical location. Through the normalization by the observation uncertainty the MPC will adapt to these different parameters and allow a larger tolerance for model results when the observed signal is not significant enough compared to the observation uncertainty. In this section we use the European Air quality database (AirBase, 1997) over the year 2009 to calculate average MPC values for daily average PM10, hourly NO_2 and daily 8h-max O_3. Data from more than 700 stations are classified in terms of station types (urban, rural, traffic) and by geographic locations (selection of areas around given cities in Europe).

For each pollutant, station type and geographic area, the average value of the MPC is shown in Table 1. Note that all stations types are considered when assessing the dependency on geographic locations whereas all geographic locations are considered when assessing the dependency on station types.

These MPC values are obtained from Eqs. (4)–(6) using the ratio $\overline{O}/\sigma_O$ of the observations' sample and assuming a constant U_r for each pollutant, set according to the AQD (2008) DQO value around the limit or target value for each pollutant: 25% for daily average PM10, 15% for 1-h NO_2 and daily maximum 8-h mean O_3 concentrations). The MPC for NMSD and R (Eqs. (5) and (6)) become more stringent with the decrease of the ratio $\overline{O}/\sigma_O$. As a consequence of the assumed constant U_r, NMB only slightly changes in relative terms but this will not remain the case once a more realistic concentration dependency for the uncertainty will be introduced in the formulation.

For daily averaged PM10 concentration the MPC show little variation with respect to the station type, while significant variations (in the ranges 87%–110% for NMSD) are visible for the different geographical areas. MPC are more restrictive over the Po valley than in the Paris area, as the observations show a larger standard deviation in the Po Valley.

		Daily PM10			8h-max O_3			Hourly NO_2		
		NMB (%)	R	NMSD (%)	NMB (%)	R	NMSD (%)	NMB (%)	R	NMSD (%)
type	Rural bckg.	58	0.48	101	32	0.54	93	38	0.87	50
	Urban bckg.	58	0.44	104	33	0.69	78	36	0.85	54
	Traffic	57	0.40	108	33	0.68	79	35	0.80	62
area	Po Valley	62	0.60	87	35	0.82	59	36	0.84	57
	Paris	60	0.56	94	33	0.68	79	36	0.84	57
	Krakow	56	0.62	110	33	0.74	72	36	0.83	57

For the daily maximum 8-h mean O_3 concentration, significant changes are visible both with station type and geographical area (in the ranges 0.54 to 0.82 for R and 59%–93% for NMSD). The more stringent criteria are found for urban and traffic stations, as they are characterized by larger σ_O.

For hourly NO_2 the MPC show limited variation with respect to the geographical area (in the ranges 0.83 to 0.87 for R and 50%–62% for NMSD) but are more stringent in all cases compared to the other pollutants, as a consequence of the larger observation standard deviation.

As expected, whenever the ratio $\overline{O}/\sigma_O$ becomes large enough, the described MPC become easy to satisfy. For example the comparison of the summer and winter observations average $\overline{O}$ and standard deviation σ_O for daily maximum 8-h mean O_3 (Fig. 2) shows a large variation in $\overline{O}$ whereas σ_O remains around 20 μg m^{-3} in both seasons. This means that U becomes larger in summer, comparable in magnitude to σ_O, leading to easier to satisfy MPC in summer than in winter. Note however that more stringent MPC for summer ozone will be obtained when the dependency of U_r on concentration level O_i will be included (U_r larger for smaller O_i).

For applications on other data-set the user can either use the normalized MPC (ranging between 0 and 1) referred by Eqs. (9)–(11), which remain valid regardless of the particular case studied, or the set of non-normalized MPC which varies depending on the case studied. These non-normalized MPC are $\overline{O}/\sigma_O$ dependent and can easily be calculated from the observations at the location of interest.

As mentioned above more realistic MPC values will be obtained when introducing a concentration dependent observation uncertainty. While it is sensible to use a maximum uncertainty as reference value combined with an MPC to unity for modeling applications related to policy-support, the use of a more stringent MPC condition (between 0.5 and 1) might be useful to model developers to help identifying the areas of potential improvements in a given model application.

5. Performance criteria values and diagrams for visualization

In the previous section, three statistical indicators have been related in a consistent way to the $RMSE_U$ indicator. In this section existing diagrams are adapted to include a visualization of the MPC in terms of fulfillment areas.

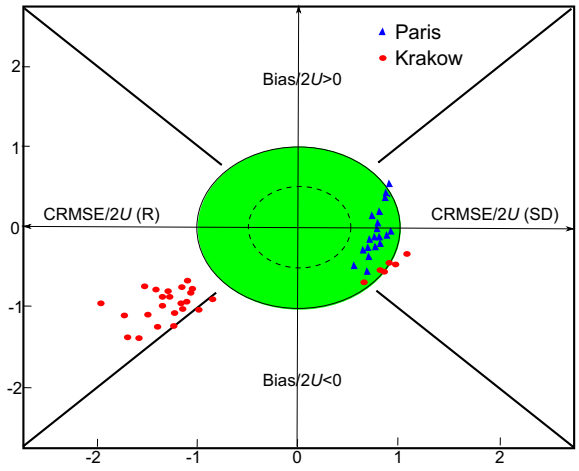

Fig. 3. Adaptation of the Target diagram to visualize the main aspects of model performance of EC4MACS-CHIMERE run at 7 km for Krakow (red) and Paris (blue) for the whole year 2009. The green area represents performance fulfilling the main criterion RMSE/2U < 1. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The graphical representations which have been retained in this approach are a normalized version of the Target diagram proposed by Jolliff et al. (2009) which summarizes the RMSE, bias and CRMSE statistics, the scatter plot for the bias and two new diagrams to represent the standard deviation (NMSD) and the correlation (R) performance. The proposed diagrams are tested on a real case in which AirBase observations are used to evaluate the CHIMERE model (Rouil et al., 2009; Honoré et al., 2008) run over Europe for the entire year 2009 with a spatial resolution of 7 km. This run is performed in the frame of the European Consortium for Modeling of Air Pollution and Climate Strategies (EC4MACS, 2007) with the aim of assessing the urban impact on daily exceedances of PM and NO_2 in European cities. All data reported in the example plots (Figs. 3–6) refer to Krakow and Paris for the year 2009, cities which have been selected for their different behavior in terms of MPC.

In the Target diagram (Jolliff et al., 2009) the X and Y axis (bias and CRMSE) are now normalized by the observation uncertainty

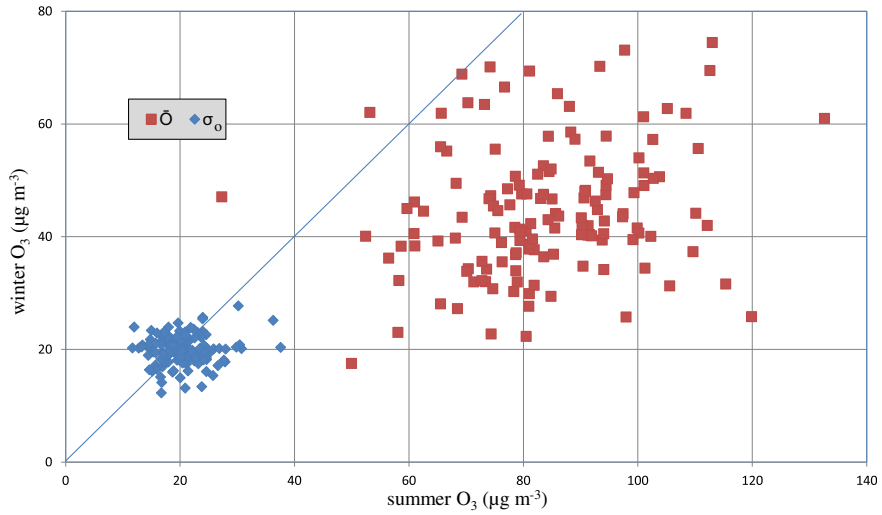

Fig. 2. $\overline{O}$ (blue dots) and σ_O (red dots) for daily maximum 8-h mean O_3 concentration at each site during a) winter and b) summer. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

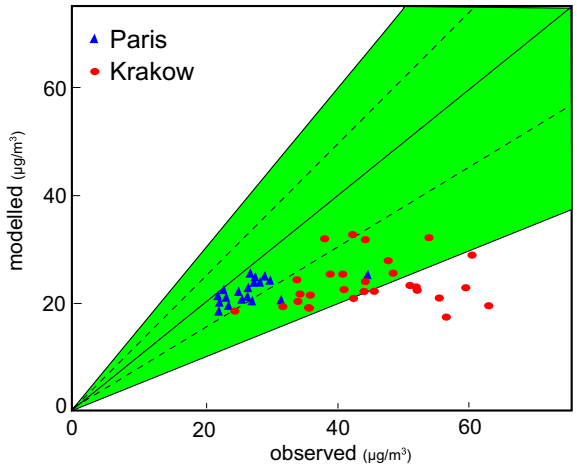

Fig. 4. Scatter plot with associated bias fulfillment zones (in green), with the assumption of U_r independent from concentration level O_i. The dataset is the same as for Fig. 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

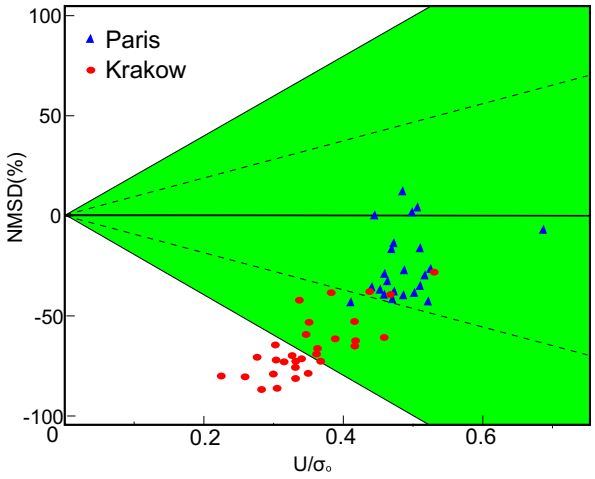

Fig. 6. R diagram with associated performance criteria and fulfillment zones (in green) for the same dataset of Fig. 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

(Fig. 3): for each point (representing one station) on the diagram the abscissa is bias/2U, the ordinate is CRMSE/2U and the radius is proportional to $RMSE_U$. The green area identifies the fulfillment of the RMSE criteria of Eq. (1).

Because CRMSE is always positive only the right part of the diagram would be needed in the Target plot, the negative X axis section can then be used to provide additional information (Jolliff et al., 2009). This information is obtained through relation (3) which is used to further investigate the CRMSE related error and see whether it is dominated by R or by SD. 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:

$$\frac{\mathrm{CRMSE}(R=1)}{\mathrm{CRMSE}(\sigma_M=\sigma_O)} = \frac{\mathrm{NMSD}}{\sqrt{2(1-R)}} \begin{cases} >1 \rightarrow \mathrm{SD\ dominates\ on\ } R \rightarrow \mathrm{right} \\ <1 \rightarrow R \mathrm{\ dominates\ on\ SD} \rightarrow \mathrm{left} \end{cases}$$

For ratios larger than 1 the SD error dominates and the station is represented on the right, whereas the reverse applies for smaller than 1 ratios. Four main zones are then identified in the diagram: the lower and top zones identify errors dominated by (negative and positive) bias errors whereas the left and right zones identify errors dominated by correlation or standard deviation, respectively.

The proposed normalization then allows visualizing in the same diagram model results for many different stations, despite their different U values.

Fig. 4 provides an example of a classical scatter plot, where the fulfillment area for the bias MPC ($|\mathrm{bias}| < 2U$ of Eq. (4)) is depicted in green.

The diagrams for NMSD (Fig. 5) and for R (Fig. 6) have U/σ_O in the abscissa. The green areas indicate the fulfillment area for their MPC as reported in Eqs. (5) and (6).

While MPC are fulfilled for all stations in the Paris area (Fig. 4) they are not fulfilled for 10 stations in Krakow. In this area observations range mostly between 30 and 60 µg m^{-3}, while the CHIMERE results remain between 20 and 30 µg m^{-3}. For the same area 15 stations do not fulfill the MPC for NMSD (Fig. 5) and 24 stations do not fulfill the MPC for R (Fig. 6). The poor CHIMERE performance in Krakow is therefore mostly related to CRMSE and in particular to a poor temporal correlation. The Target plot (Fig. 3) synthesizes the same information in a single diagram. Further analysis (not shown) indicates that this poor correlation is mostly due to a strong underestimation of concentrations in wintertime combined with a slight overestimation in summertime.

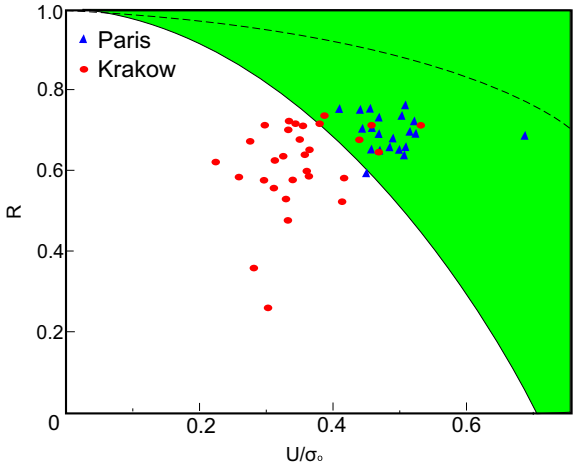

Fig. 5. NMSD diagram with associated performance criteria and fulfillment zones (in green) for the same dataset of Fig. 3. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

6. Conclusions

Models applied for regulatory air quality assessment are commonly evaluated on the basis of comparisons against observations. This comparison is generally achieved using a set of statistical indicators to analyze model performance. But these indicators are useful only if used together with a performance scale, in order to inform the user on the expected value an indicator should reach for a particular modeling application (e.g. what is the expected value correlation should reach for PM10 modeling at background sites?). MPC need therefore to be defined for each statistical indicator to ensure the user that a minimum level of quality is achieved for a given model application.

In this work a set of performance criteria is proposed for four statistical indicators (RMSE, NMB, NMSD and R) which summarize the model-observation errors in terms of phase, amplitude and

bias. These statistics have the peculiarity of being normalized by the observations uncertainty U, and performance criteria are derived based on the hypothesis that model results are allowed the same margin of uncertainty as for measurements. This normalization is chosen with the main objective to provide a common scale to evaluate model performance for policy support.

Based on AirBase data, MPC empirical values for NMB, NMSD and R for various compounds have been calculated with the hypothesis that U_r does not depend on the concentration level, showing important variations in terms of different station types and geographical locations. MPC fulfillment areas are depicted in green in the four diagrams proposed to help the user in the evaluation of its model application: an adapted Target plot, the scatter plot for the bias and two new diagrams for R and NMSD. The utility of those diagrams and in particular of the adapted Target plot have been emphasized for a real modeling application.

One of the main advantages of this approach is to allow the setting of a consistent set of MPC for different indicators based on one single input: the observation uncertainty U. As the MPC are based on a comparison of the data characteristics (mean concentration, standard deviation) with their uncertainty, the MPC depend on the station type, pollutant, geographic area… Normalized indicators have been proposed to provide an overview of model performance, while non-normalized MPC have been calculated here on a real case.

In this work the maximum uncertainty and an equal tolerance for both model and measurement has been used (i.e. MPC < 1) since the focus was on policy-support types of applications. But a more stringent condition (MPC between 0.5 and 1) might be useful to model developers to understand where (which stations) and how (which indicator) to improve model performance.

As more detailed information on U becomes available (e.g. dependency on concentration level, quantification of its systematic and random components, estimation of the uncertainty related to the lack of spatial representativeness), more realistic MPC values can be obtained.

Acknowledgments

We wish to thank our colleagues Claudio Belis, Michel Gerboles, Kees Cuvelier as well as Mario Miglietta for their useful comments and suggestions. The CHIMERE output data were provided by INERIS (E. Terrenoire and B. Bessagnet) in the frame of the EU LIFE EC4MACS project.

References

AirBase, 1997. European Air quality database (WWW Document). URL: http://acm.eionet.europa.eu/databases/airbase.

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.

ASTM standard D6589, 2005. Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (No. D6589). ASTM International, West Conshohocken, PA.

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.

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. Meteorology and Atmospheric Physics 87.

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 No. version 6.2).

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S.T., Scheffe, R., Schere, K., Steyn, D., Venkatram, A., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems. Environmental Fluid Mechanics 10, 471–489.

Derwent, D., Fraser, A., Abbott, J., Willis, P., Murrells, T., 2010. Evaluating the performance of air quality models (No. Issue 3). Department for Environment and Rural Affairs.

EC4MACS, 2007. European consortium for modelling of air pollution and climate strategies (WWW Document). URL: http://www.ec4macs.eu/home/index.html?sb=1.

EPA, 2007. Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, PM2.5, and Regional Haze. No. EPA-454/B-07-002. U.S. Environmental Protection Agency.

EPA, 2009. Guidance Document on the Development, Evaluation, and Application of Regulatory Environmental Models. No. EPA/100/K-09/003. U.S. Environmental Protection Agency.

Flemming, J., Stern, R., 2007. Testing model accuracy measures according to the EU directives—examples using the chemical transport model REM-CALGRID. Atmospheric Environment 41, 9206–9216.

Gerboles, M., Lagler, F., Rembges, D., Brun, C., 2003. Assessment of uncertainty of NO2 measurements by the chemiluminescence method and discussion of the quality objective of the NO2 European Directive. Journal of Environmental Monitoring 5, 529.

Honoré, C., Rouïl, L., Vautard, R., Beekmann, M., Bessagnet, B., Dufour, A., Elichegaray, C., Flaud, J.-M., Malherbe, L., Meleux, F., Menut, L., Martin, D., Peuch, A., Peuch, V.-H., Poisson, N., 2008. Predictability of European air quality: assessment of 3 years of operational forecasts and analyses by the PREV'AIR system. Journal of Geophysical Research 113.

Irwin, J.S., Civerolo, K., Hogrefe, C., Appel, W., Foley, K., Swall, J., 2008. A procedure for inter-comparing the skill of regional-scale air quality model simulations of daily maximum 8-h ozone concentrations. Atmospheric Environment 42, 5403–5412.

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.

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

Murphy, A.H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. Monthly Weather Review 116, 2417–2424.

Rouil, L., Honoré, C., Bessagnet, B., Malherbe, L., Meleux, F., Vautard, R., Beekmann, M., Flaud, J.-M., Dufour, A., Martin, D., Peuch, A., Peuch, V.-H., Elichegaray, C., Poisson, N., Menut, L., 2009. Prev'air: an operational forecasting and mapping system for air quality in Europe. Bulletin of the American Meteorological Society 90, 73–83.

Shluenzen, K.H., Sokhi, R.S., 2008. Overview of Tools and Methods for Meteorological and Air Pollution Meso-scale Model Evaluation and User Training. GAW report No.181, WMO/TD – No.1457 No. Joint Report of COST Action 728 and GURME.

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., Georgieva, Emilia, Pederzoli, Anna, 2011. The DELTA tool and Benchmarking Report template. Concepts and User Guide. Version 2.