A procedure for air quality models benchmarking

P. Thunis, E. Georgieva, S. Galmarini
Joint Research Centre, Ispra

Version 2

16 February 2011

1. Introduction

The aim of the proposed activities under the SubGroup4 of WG2 in FAIRMODE is to develop a procedure for the benchmarking of air quality models in order to evaluate their performances and indicate a way for improvements. The procedure is meant as a support to both model users and model developers.

In the literature on model's performance evaluation the term 'benchmark' is often related to a certain range of values for some statistical indicators valid for a specific type of application (Tesche et al., 2002). A broader meaning of 'benchmarking' based on the definition proposed by UNESCO (Vlãsceanu et al., 2004) is however better suited for our purposes: It is

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

The term *procedure* in the following of the document is understood as a sequence of operations, steps that a modeler is asked to perform using provided tools in order to acquire a quantitative assessment of his model results.

The benchmarking procedure is intended to support modelling groups in their application of AQ models in the frame of the Air Quality Directive, 2008 (AQD). The type of pollutants, period of interest and spatial scales will be determined by those required by the AQD.

The procedure will be based mainly on existing technical aids - the evaluation tools developed in the CityDelta (CD) and EuroDelta (ED) projects (URL1) and the *ENSEMBLE* systems (URL2). These tools will be properly adapted and renewed taking into account the experience gathered world wide on air quality model evaluations and available software. Some of the best known ones include: the Model Validation Kit of the Harmonization initiative (Olesen, 1995 and 2001) which contains the *BOOT software* (Chang and Hanna, 2004), the *ASTM Guidance* (ASTM, 2000), , the USA-EPA *AMET package* (Appel and Gilliam, 2008; URL3) and the *VERDI* tool (URL4). Recent developments in statistical model evaluation (e.g. Boylan and Russell, 2006, Jolliff et al. 2009) will be also taken in consideration. New auxiliary tools are set up as support to the main ones.

2. *Overview of the benchmarking procedure*

The procedure is explained here in its essence to give a global overview on the approach that is proposed.

2.1. Pre-requisites to the procedure

We assume:

- the existence of a community of modelling groups in the EU context.
- that each modelling group works periodically or on a regular bases on modelling the air quality for a specific country, sub-country region or geographical context
- that each member of the community is expected to report on a yearly basis on the air quality of its country to the Commission
- that the community can periodically come together to work on a common case study

2.2. Key elements for the procedure

The procedure is made of four key elements:

The single model evaluation tool here after referred to as DELTA

Each modelling group will be given what we define here as single model evaluation software. This software, DELTA, based on the ED and CD tools will constitute an EU wide distributed tool that will allow each group to evaluate the model results performing a series of specific tests (see section 4.2). The ED and CD tools have been already successfully used for model evaluation and intercomparison (Cuvelier et al. 2007; Thunis et al. 2007.) and the DELTA tool will include their main assets. The real advantage of using this tool is that every group will be using the same scale and a decent level of harmonization will be reached across EU in terms of at least a common evaluation standard. DELTA should be intended mainly as a tool to help in a rapid diagnostics of the model performance based on the so called *reduced set*. The latter is the set of variables of direct relevance for the AQ directive (see section 4.1 and *Annex C*). The tool will also allow multiple model analysis to help for example in the comparison of the result of different model versions.

The ENSEMBLE system

The ENSEMBLE system is a web based platform developed for multi purpose model application that allows on line model inter-comparison and evaluation (Galmarini, 2001, 2004, 2008, 2010, Potempski et al. 2008). It allows single and multi model analysis in the three spatial dimensions and time. The comparisons can be based as model versus monitoring data, model versus model, model versus group of models. ENSEMBLE will host all the model results and available measurement pertaining to campaigns and case studies. This time the full set of variables could be evaluated (including meteorology). The real scope of ENSEMBLE within the procedure is to provide a platform to thoroughly assess the results consistency and to verify whether the right results are obtained for the right reason. This is assumed to assist the modelling groups to acquire more confidence in the quality of their results. ENSEMBLE serves a series of modelling communities and enters in this procedure by providing a service to the AQ community beside the others and ad-hoc activities presently on going. It could be seen as an additional and redundant element in the procedure though it is supposed to act also as central repository of organized, and harmonized (format wise) model/monitoring data. The latter will be available (provided all confidentiality and copy right privileges are granted) to the research and policy support communities for repeated model evaluations, testing and development.

The benchmarking service

This is a service that will produce summary reports containing performance indicators related to a given model application in the frame of the AQ directive. The reports will be obtained through an automatic procedure and follow a pre-defined template structured around core indicators and diagrams (see *Annex B*). This service will also contain some bounds for specific indicators, called hereafter *goals and criteria* to help in the assessment of the model performance. More details are given in section 4.3. These goals and criteria will be regularly revised based on future joint modelling exercises.

The benchmarking service will be JRC based service for final model performance reports and for the updating of goals and criteria. A replica of the reporting service will be included in the DELTA tools to produce “working” reports to be used in the phase of model testing until a satisfactory level of model performance is achieved.

The reports will be the same for all models undergoing the procedure. The working report will be available only to the model user (i.e. the person undertaking the study) whereas final reports are intended for the whole community.

The data extraction facility

This is a web service that will facilitate model users in the extraction of monitoring data from existing centralized databases. A sort of one-stop-shop where, to start, air quality data could be collected for specific regions and periods of time.

2.3. Description of the procedure

The procedure proposed (see Figure 1) makes use of the above mentioned aids in the following three ways:

1- Individual-model/Member-State yearly reporting to the Commission:

In this case each country produces a yearly simulation on its territory or neighbouring countries.

a. The data extraction facility could be used to acquire additional information not available to the modelling group
b. The tool DELTA is supposed to be used until the modeller is satisfied with the quality of the results
c. The benchmarking working report can be produced within the DELTA tool on the single model performance. This will help model users to define the quality of current model result with respect to the defined standard
d. When satisfied with the quality of the results, the user submit the reduced dataset to JRC to produce the final report.
e. The final report and the data that generated them will be stored at JRC.

2- Periodical joint activities: Periodically the JRC organizes a common exercise where the MS modelling community is requested to take part. These exercises may focus on a specific case study, area or period of interest. Data on the case could be made available through the data extraction facility.

Each group is requested to:

a. Use the tool DELTA until the modeller is satisfied with the quality of the results, which includes producing working reports as in case (1)
b. Transfer of the data (*full set*) to the ENSEMBLE system at JRC for thorough testing and inter-comparison with other models. The transfer should include working performance report for consultation by other groups. This will represent a synthetic summary of each model performance defined on the same standard useful to all to know the quality of the results each one is comparing with.

When consensus will be reached on the acceptability of the full sets of every model, JRC will apply the benchmarking service to produce a working multi-model performance reports. When satisfied with the quality of the results:

a. The final performance reports will be produced and data archived
b. Goals and criteria will be updated and made available for future type 1 applications

This exercise will have as major scope that of collecting over time information on statistical behaviour of models that will be used together with measurements to set the standards of the benchmarking service and the reports. This is an important part of the procedure that allows the community to come together to discuss problems, issues or tackle specific problematic modelling areas.

3- It goes without say that should a country want to gather a group of neighbours typically in the analysis described at (1-) it may well do so. The DELTA tool could be used for fast diagnostics of model performance. The passage through ENSEMBLE could be relevant and could be considered as well for a more thorough analysis. However should this happen no contribution to the multi model benchmarking report will be considered out of this application.

The added value of the proposed procedure consists of:

1- All models used for regulatory purposes will be evaluated with the same single model tool
2- All models will have a common place where to compare, inter-compare, evaluate, test, experiment with each other on past or present case study without having to acquire data from other modelling groups but by using the online web facility.
3- A reporting tool is made available that will take into account the dynamical evolution of the state of the art and will record the historic evolution of model quality that would be relevant for policy decisions.
4- The data relevant to the various types of analysis will be available from a single source point.

In the next sections details on some of the aspects that characterize the procedure for model performance evaluation will be presented.

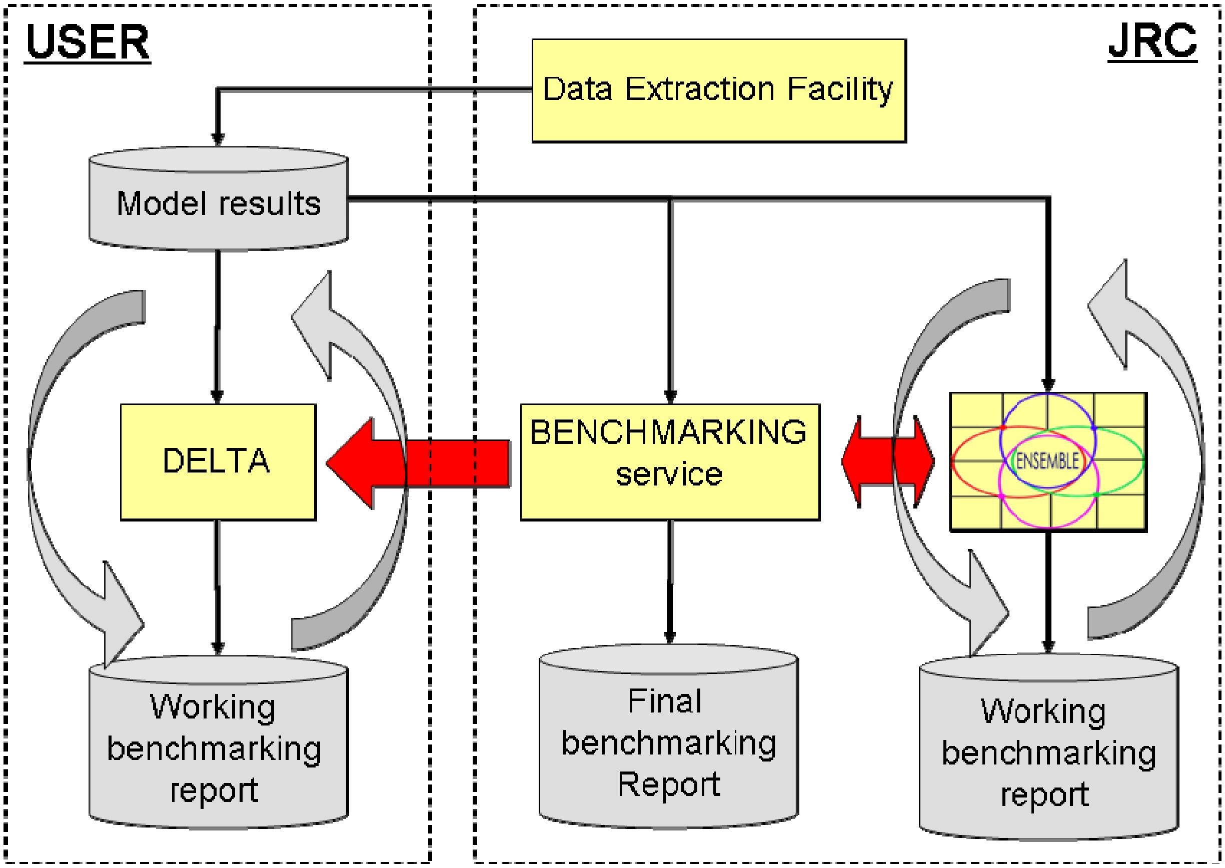

USER
JRC
Data Extraction Facility
Model results
DELTA
BENCHMARKING service
ENSEMBLE
Working benchmarking report
Final benchmarking Report
Working benchmarking report

Figure 1 Scheme of the procedure

3. *Procedure main concepts*

The procedure will take into account the following key concepts:

- The design and the structure shall address modelling for the AQD applications (AQ assessment and Planning). The procedure should be applicable for different models (excluding CFDs) at different spatial scales with the only requirement of providing data on a grid basis or at receptor locations, with adequate temporal and spatial resolution. At the initial stage the focus will be on O3, NOx and PM. The type of models, pollutants and spatial scales are briefly discussed in *Annex* A.

- AQM are complex systems that rely on information gathered from external source like emissions and meteorology. Emphasis will be put on the comparison of final model air quality concentrations with observational data. However, input consistency tests (emissions, meteorology, boundary conditions) shall also be designed in an attempt to identify possible causes for erroneous results and to understand differences in model results.

- Model performance will be assessed not only on the basis of comparison to monitored data (model-to-observations indicators, MOI) but also on the basis of model-to-model results comparison (model to model indicators, MMI); and on the basis of the comparison of model response to given emission perturbations (Model response indicators, MRI).

- A set of appropriately selected model performance indicators (statistical indicators) for MOI, MMI and MRI and relevant diagrams will be defined and included in the procedure. The set is referred hereafter as core set, for details see *Annex B*. Composite indicators or diagrams (e.g. index of agreement, Taylor diagram…) will be privileged and used for the reporting.

- Model performance indicators will be assessed (when possible) with respect to “criteria” and “goals”. Performance criteria are defined as the level of accuracy considered to be acceptable for regulatory applications whereas performance goals express a level of accuracy close to optimal that a model is expected to achieve for a given application. The definition of these two levels of performance used mainly for MOI will be based on literature, available AQ model results and/or expert judgment (see section 4.3 for details).

- The procedure is not meant to establish a ‘bright-line’ criterion on the acceptability of a given model for a given type of application. It rather aims at providing the model user with some information on the quality of the model results, including indications on expected model performances and pointing out the strengths and weaknesses in a specific application.

- The benchmarking service will contain an automatic reporting system for summarising model performances. The reports will be produced both in the single and multi-model mode and are intended to serve AQM users and developers, as well as regulatory agencies.

- Successful application of the procedure constitutes a necessary but not sufficient step in the model evaluation procedure.

4. The Benchmarking Service

4.1. Methodology

Reduced set

The reduced set of model results (see *Annex C*) is a subset of the full model results which are reduced temporally, spatially and in terms of pollutants. Full temporal resolution is kept only at selected grid points (corresponding to measurement stations or not (e.g. for multi model comparisons)) or for an array of selected points (e.g. predefined AQ zones).

The reduced data set will also include temporally averaged data (e.g. monthly) over the entire grid to provide 2-D overview of the whole domain under analysis.

Spatial and temporal sub-divisions

For modelling over large spatial domains or covering long time periods (e.g. yearly simulations) the evaluation will be broken down into individual segments such as geographic sub-regions (e.g. exceedance areas, urban vs. rural…) and/or months/seasons to allow for a more comprehensive assessment (see *Annex C*).

For each of these spatial / temporal segments, evaluation will cover different dimensions:

- Temporal: analysis of time series at given locations, correlations…
- Spatial: analysis of spatial patterns for given time periods.
- Distribution: comparison of cumulative distribution regardless of the pairing in time and/or space (only ranking is important here).

Statistical diagrams and statistics

A wide variety of indexes can be found in the literature, proposed for different fields of application (meteorology, air quality…), different scope (e.g. forecast, episode study) or different type of application (e.g. regulatory). Regardless of the model application, scope and type, it is recommended to apply multiple performance indicators since each one has its advantages and disadvantages.

The indexes proposed for the Benchmarking service have been selected based mainly on the review of model evaluation methodologies by Chang et Hanna (2004), the recent overview of tools and methods for meteorological and air pollution mesoscale model evaluation (Schluenzen and Sokhi, 2008), the EPA guidance documents (2007, 2009), the metric and statistic definitions and discussion in the project SEMIP (Smoke and Emissions Model Intercomparison Project, URL5**),** the recommendations of Boylan and Russel (2006) for setting model performance metrics, goals and criteria, the proposal for composite diagrams by Jolliff et al. (2009) and the recommendations from the EU-AIR4EU (Borrego et al, 2008) and the FAIRMODE Guidance document on the use of models for the European AQD (Denby, 2009). An overview of these statistical indicators and diagrams, named hereafter core statistical indicators is given in *Annex B.*

4.2. Testing levels

The purpose of the tests is to identify weaknesses and strengths of the AQ system and to flag areas of required improvements before AQ modelling results reach a certain degree of acceptability. We apply here the philosophy of model acceptability as a continuing process of non-rejections, i.e. without major or fatal flaws, as proposed by Morris and Tesche (2005). Note that this assessment is performed independently of whether the process descriptions in the model are accurate.

Series of tests with different level of complexity are used to elaborate a comprehensive (spatially and temporally) model performance evaluation. The tests in the benchmarking service will be carried out on the reduced set of models output focusing on the core performance indicators. However, both tools will allow exploratory analysis with more flexibility with respect to time intervals, species, sub-sets of data etc

1. *ICI* (Input Consistency Indicator): Specific input to the air quality models (i.e. emissions) are tested for consistency. A comparison with prescribed input data is performed (spatial distribution, temporal disaggregation…). Overall consistency / plausibility of the model results will be checked through the use of specific indicators (e.g. max-min values compared to references…). Inconsistencies will be flagged out to the user.

2. *MOI* (model-to-observation indicators): Air Quality model results and meteorological variables (those used as input to AQM) are compared to available observations using the core set of statistical indicators and diagrams. The Model Quality Objectives (MQO) as described and required in the AQD are also included in the analysis. MOI are evaluated against expected model performances (performance criteria and goals). Information on validation data uncertainty is provided whenever available.

3. *MMI* (model to model indicators): Comparison of model results with other models for both measured and non-measured compounds. In contrast to the MOI tests, observations or known input values are not used in the analysis. This implies that model results are also compared in areas without observations, allowing for a more complete analysis. Note that some species ratio (e.g. O3/NOy) which can provide some useful information on chemical regimes can also be compared in this test level. Model-to-model comparisons of some important meteorological parameters (or boundary conditions for AQM) for which no observations exist (e.g. diffusion coefficient, PBL height…) will be included at this test level.

4. *MRI* (Model response indicators): Comparison of model responses to given emission perturbations. The objective of these tests is to check that good MOI performances arise from a good representation of the atmospheric and chemical processes in the model. A range of expected responses (model variability) will be built-in to help modelers identifying the areas where the model might be improved.

4.3. Goals, Criteria and Uncertainty of Observations (UO)

In order to provide some help to assess quantitatively the performance of a model for a given application we propose here to follow some recommendations from Boylan and Russel (2006) which were included in the EPA methodology for model evaluation (EPA 2007 & 2009). This methodology makes use of criteria and goals in terms of prescribed values for some statistical indexes used in model validation. In addition to these criteria and goals we propose to include also some information related to the uncertainty characterizing the observations used for validation as proposed by Jolliff et al. (2009). A definition of these terms is given here below:

- Criteria is defined as a model acceptable performance for a given type of application.

- Goals is defined as the best performance a model should aim to reach given its current capabilities. It takes into account the fact that (1) models are limited in their physical and chemical representations of the reality, (2) volume averaged model results are not directly comparable with point measurements, (3) measurements are characterized by some uncertainty.

- Observation Uncertainty (OU): The last of the above item could be developed as an independent line providing some indication information on the level of accuracy of the measurements.

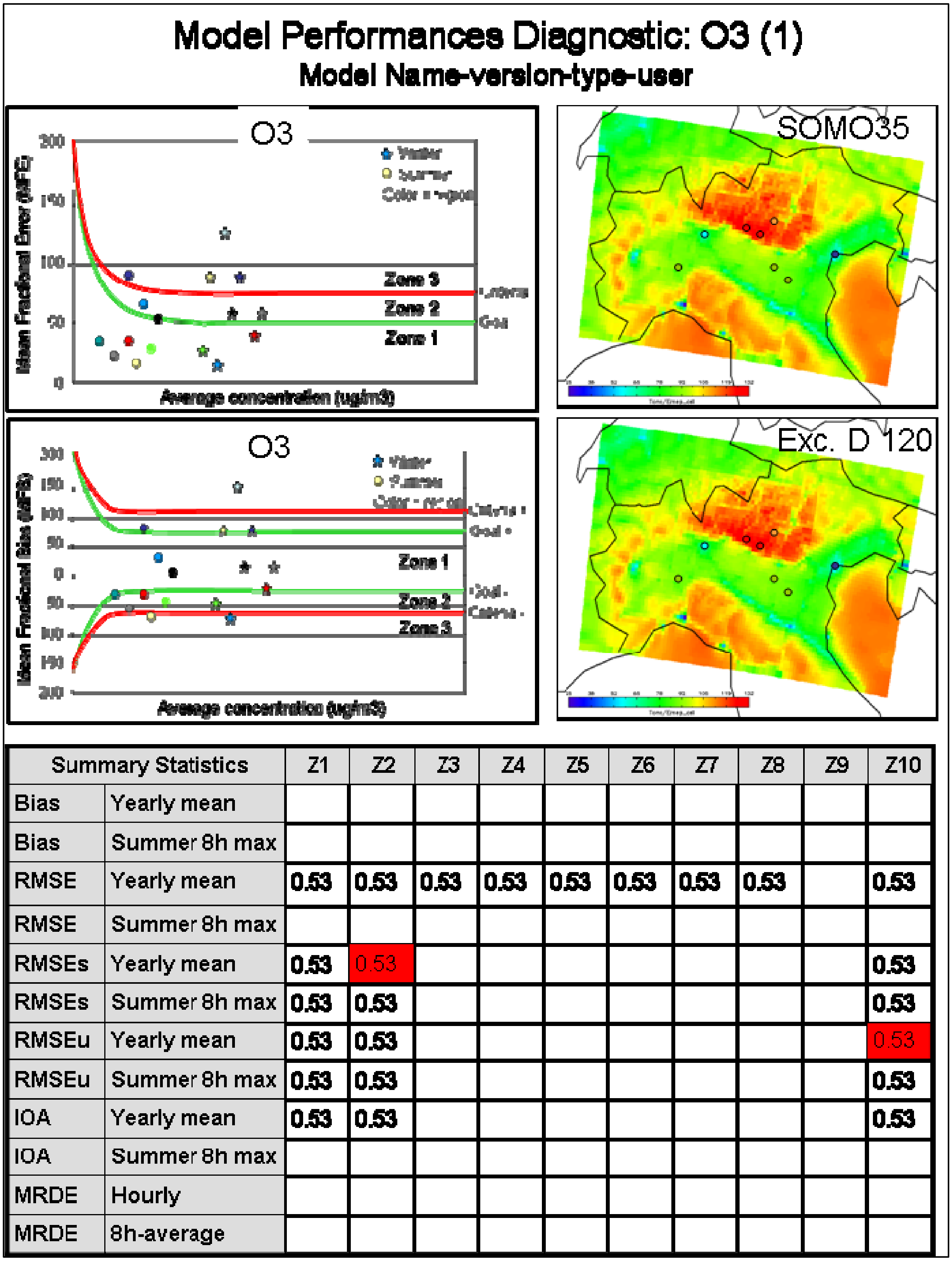
Based on a series of regional scale model evaluation studies carried out in the USA, performance goals and criteria have been proposed for some statistical metrics commonly used for air quality and meteorological model variables, (see Table 1 & Table2 in *Annex D*). These goals and criteria, based on USA- EPA studies, might be used as a first estimate in the benchmarking evaluation procedure and progressively tuned thanks to the planned joint inter-comparison activities to reach values that reflect better the model application area and the temporal period of simulations.

It has been recognized that criteria and goals could be different according to the geographical area, the season, or other factors. Based on an analysis of other model evaluation exercises in the USA, Boylan and Russell (2006) have proposed criteria and goals that vary according to the absolute value of the concentration. This methodology allows to include seasonal and spatial variations of the criteria which only becomes pollutant specific. For example PM summer concentrations, generally exhibiting lower values, would be allowed less stringent acceptability criteria.

The optimal level of performance (the goal) will reflect the fact that improvement beyond a certain limit becomes meaningless given the uncertainty in the input data, the uncertainty related to the validation data or the lack in knowledge in the physical and chemical model representations of the phenomena. The performance criteria and goals should remain flexible to include improved knowledge and to reflect different applications (purpose, time and spatial scale).

5. *Reporting*

The benchmarking service will produce automatically reports on model performances. The content of the reports will include both quantitative and qualitative information, based on the selected core indexes and summary diagrams applied on the reduced set. As a first step a single model performance report will be produced. For joint exercises both single and multi-model reports will be produced and archived in a database for consultation. An example of such report is provided below (Figure 2Figure 2.



Model Performances Diagnostic: O3 (1)

Model Name-version-type-user

Summary Statistics		Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9	Z10
Bias	Yearly mean										
Bias	Summer 8h max										
RMSE	Yearly mean	0.53	0.53	0.53	0.53	0.53	0.53	0.53	0.53		0.53
RMSE	Summer 8h max										
RMSEs	Yearly mean	0.53	0.53								0.53
RMSEs	Summer 8h max	0.53	0.53								0.53
RMSEu	Yearly mean	0.53	0.53								0.53
RMSEu	Summer 8h max	0.53	0.53								0.53
IOA	Yearly mean	0.53	0.53								0.53
IOA	Summer 8h max										
MRDE	Hourly										
MRDE	8h-average										

Figure 2 Example of model performance report for O3 – 1.part

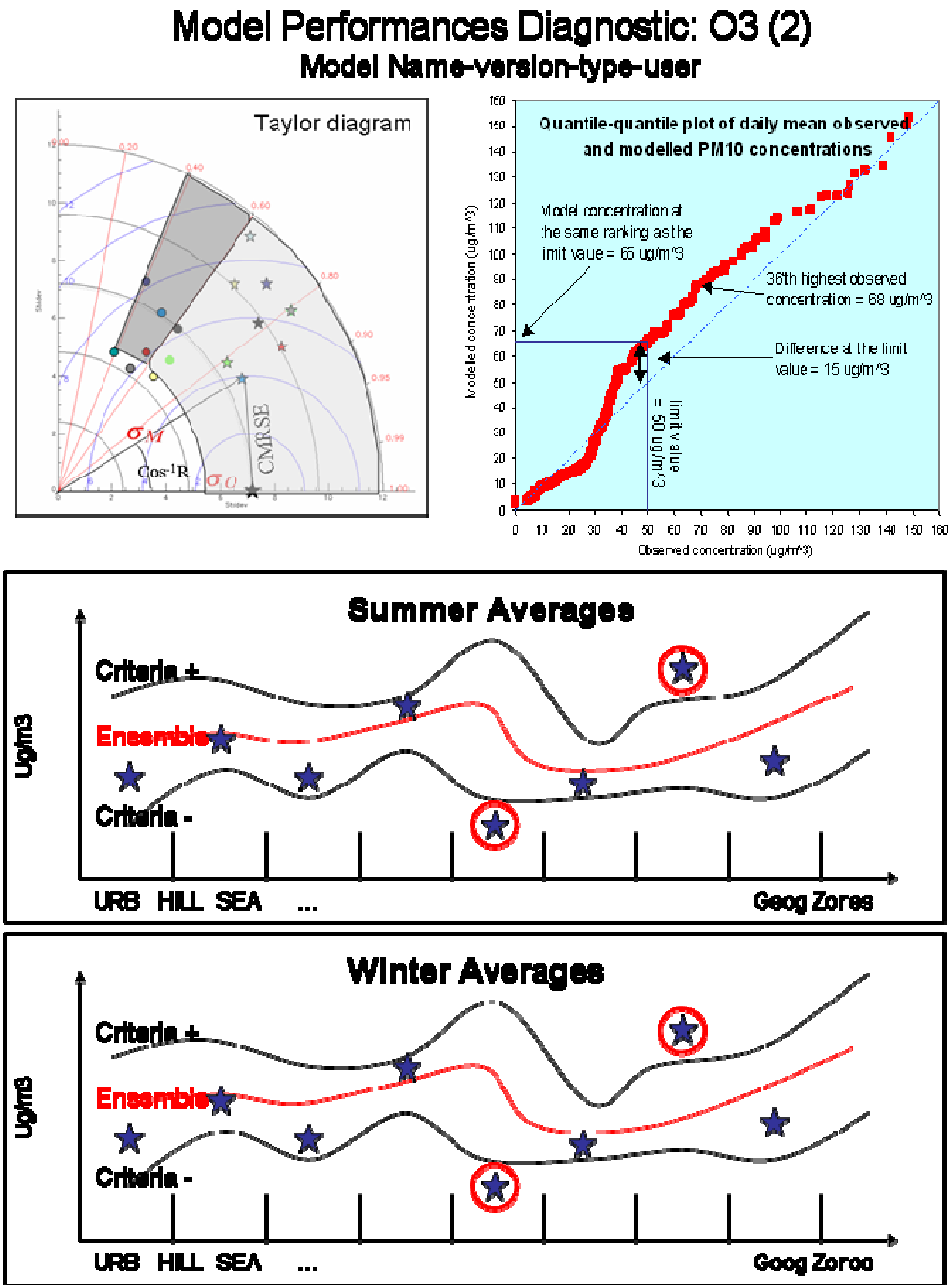


Figure 2. - continued. Example of model performance report for O3 – 2.part

6. *Proposed work plan*

- Release of the proposal to WG2 SG4 participants (01/05/2010)
- Discussion and consensus on overall methodology (01/07/2010)
- Development of the DELTA prototype and benchmarking service (31/12/2010)
- Testing of the DELTA prototype on existing datasets (2011)
- Development of the JRC on-line facilities (31/12/2011)
- Set-up of a joint exercise and testing of the whole system (2012)

7. *References*

Appel, K. W. and Gilliam, R. C.: Overview of the Atmospheric Model Evaluation Tool (AMET),2008. 7th Annual CMAS Conference, Chapel Hill, NC, 6–8 October 2008, http://www.cmascenter. org/conference/2008/agenda.cfm, 2008.

ASTM, 2000: Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance. D 6589-00. American Society for Testing and Materials, West Conshohocken, PA.

Borrego, C., Monteiro A., Ferreira J., Miranda A.I., Costa A.M., Carvalho A.C and Lopes M., 2008: Procedures for estimation of modeling uncertainty in air quality assessment, Environment International, 34, 613-620

Boylan J and Russel A. 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. and Hanna, S. R., 2004. Air quality model performance evaluation. Meteorol. Atmos. Phys. 87: 167-196.

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

Cuvelier C., P. Thunis, R. Vautard, M. Amann, B. Bessagnet, M. Bedogni, R. Berkowicz, J. Brandt, F. Brocheton, P. Builtjes, C. Carnavale, A. Coppalle, B. Denby, J. Douros, A. Graf, O. Hellmuth, A. Hodzic, C. Honoré, J. Jonson, A. Kerschbaumer, et al., 2007: CityDelta: A model intercomparison study to explore the impact of emission reductions in European cities in 2010 Atmospheric Environment, Volume 41, Issue 1, Pages 189-207

Denby B. (ed), 2009. Guidance on the use of models for the European Air Quality Directive (A FAIRMODE working Document), ETC/ACC report, ver 5.1. http://fairmode.ew.eea.europa.eu/

Emery, C., E. Tai, and G. Yarwood, 2001: “Enhanced Meteorological Modeling and Performance Evaluation for Two Texas Ozone Episodes”, report to the Texas Natural Resources Conservation Commission, prepared by ENVIRON, International Corp, Novato, CA.

EPA (U.S. Environmental Protection Agency), 2007. Guidance on the Use of Models and Other Analyses for Demonstrating Attainment of Air Quality Goals for Ozone, PM2.5, and Regional Haze, EPA-454/B-07-002, http://www.epa.gov/scram001/guidance/guide/final-03-pm-rh-guidance.pdf

EPA (U.S. Environmental Protection Agency), 2009. Guidance Document on the Development, Evaluation, and Application of Regulatory Environmental Models, EPA-HQ-ORD-2009, http://www.epa.gov/crem/library/cred_guidance_0309.pdf

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

Galmarini S. et al. ,2001: Forecasting the consequences of accidental releases of radionuclides in the atmosphere from ensemble dispersion modelling. Journal of Environmental Radioactivity, 57, 3, 203-219.

Galmarini S. et al. , 2004: Ensemble dispersion forecasting, Part II: application and evaluation Atmospheric Environment, 38, 28, 4619-4632.

Galmarini S., R. Bianconi, G. de Vries, R. Bellasio (2008) Real-time monitoring data for real-time multi-model validation: coupling ENSEMBLE and EURDEP, Journal of Environmental Radioactivity, 99, 1233-1241

Galmarini S., F. Bonnardot, A. Jones, S. Potempski, L. Robertson, (2010) Multi-model versus EPS-based ensemble atmospheric dispersion predictions: a quantitative assessment on the ETEX-1 tracer experiment case, Atmospheric Environment, under review.

Jolliff J. et al., 2009. Summary diagrams for coupled hydrodynamic-ecosystem model skill assessment. Journal of Marine Systems, 76, 64-82.

Morris R. and Tesche Th., 2005. St. Luis Ozone and PM2.5 Modeling Study:Modelling protocol. http://www.carallc.net/pdfs/13%20St%20Louis%20Ozone%20and%20PM2.5%20Modeling%20Protocol.pdf

Olesen, H.R., 1995, Data sets and protocol for model validation. Workshop on Operational Short-range Atmospheric Dispersion Models for Environmental Impact Assessment in Europe, Mol, Belgium, Nov. 1994, Int. J. Environment and Pollution, Vol. 5, Nos. 4-6, 693-701.

Olesen H.R. , 2001. A Platform for Model evaluation, Proceed. 7th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Belgirate, 28-31 May, 2001, Italy, pp.42-46. available at http://aqm.jrc.it/harmo7/

Potempski S. et.al. (2008) Multi-model ensemble analysis of the ETEX-2 experiment, Atmos. Environ., 42, 7250-7265

Potempski S. . Galmarini. A. Riccio, G. Giunta,(2010) On the optimal combination of multi-model ensemble prediction, JGR, submitted.

Righi S., P.Lucialli and E.Pollini, 2009. Statistical and diagnostic evaluation of teh ADMS-Urban model compared with an urban air quality monitoring network, Atm.Environment,43,3850-3857.

Schluenzen K.H. and R.S. Sokhi (eds), 2008 Overview of tools and methods for meteorological and air pollution meso-scale model evaluation and user training, Joint Report of COST Action 728 and GURME, WMO/TD-No.1457, ISBN 978-1-905313-59-4.

Sokhi R.S., H.Mao, S.T.G.Srimath, S.Fan,N.Kitwiroon, L.Luhana, J.Kukkonen,M.Haakana, A.Karppinen, K.D. van den Hout, P.Boulter, I.S.McCrae, S.Larssen, K.I. Gjerstad, R.S. Jose, J.Bartzis, P.Neofytou, P. van den Breemer, S. Neville, A. Kousa, B.M. Cortes and I.Myrtvert, 2008. An integrated multi-model approach for air quality assessment: Development and evaluation of teh OSCAR air quality assessment system, Environmental Modelling&Software, 23, 268-281.

Taylor, K.E., 2001. Summarizing multiple aspects of model performance in a single diagram. Journal of Geophysical Research 106, 7183–7192

Tesche T . W., D.E. McNally, and C. Tremback, 2002, Operational evaluation of the MM5 meteorological model over the continental United States: Protocol for Annual and Episodic Evaluation, Prepared for US EPA by Alpine Geophysics, LLC, Ft. Wright, KY, and ATMET, Inc., Boulder, CO.
http://www.epa.gov/scram001/reports/tesche_2002_evaluation_protocol.pdf

Thunis P., L. Rouil, C. Cuvelier, R. Stern, A. Kerschbaumer, B. Bessagnet, M. Schaap, P. Builtjes, L. Tarrason, J. Douros, N. Moussiopoulos, G. Pirovano, M. Bedogni, 2007, Analysis of model responses to emission-reduction scenarios within the CityDelta project, Atmospheric Environment, Volume 41, Issue 1, January 2007, Pages 208-220

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

Willmott C.J., 1982. Some comments on evaluation of model performance, Bull. Am. Meteorol. Soc. 63, pp. 1309–1313

Willmott, C.J., S.G. Ackleson, R.E. Davis, J.J. Feddema, K.M. Klink, D.R. Legates, J. ’Donnell, and C.M. Rowe. 1985. Statistics for the evaluation and comparison of models. J. Geophys. Res. 90:8995–9005.

URL1 : EURODELTA http://aqm.jrc.it/eurodelta/index.html
URL2: ENSEMBLE http://ensemble.jrc.ec.europa.eu/Ensemble_files/background.htm
URL3: AMET
http://www.epa.gov/scram001/9thmodconf/modeling%20conference%20presentation%20-%20appel%202008.pdf
URL4: VERDI http://www.verdi-tool.org/index.cfm
URL5: SEMIP: Smoke and Emissions Model Intercomparison Project http://www.semip.org/analysis/protocols/analysis-protocols#section-1

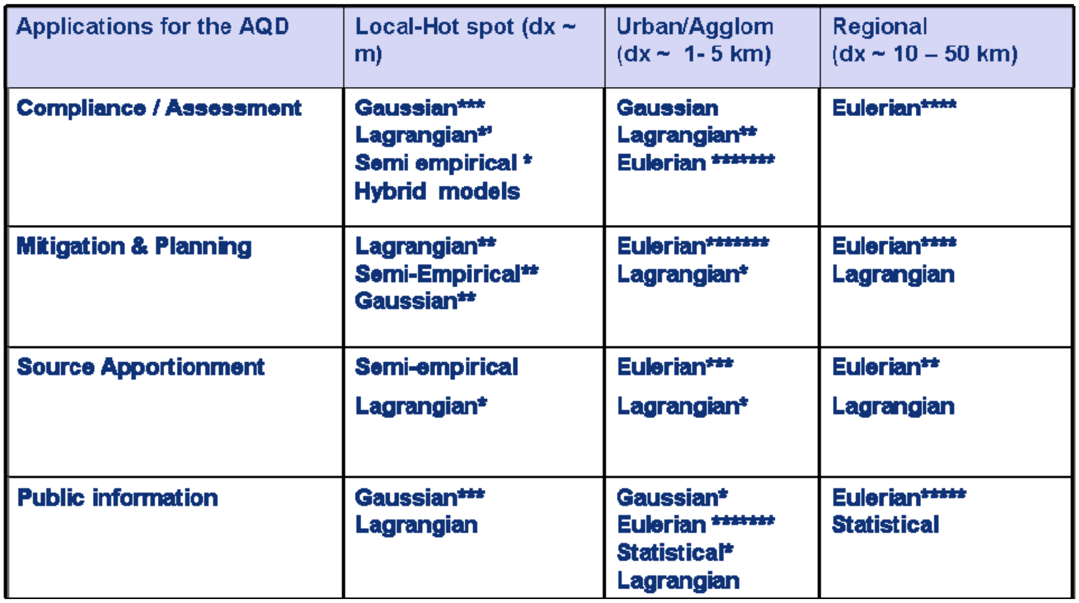
Annex A: Applications, pollutants and scales considered

The design and the structure of the tool will take into consideration modelling for AQD applications. Table 1, assembled during the FAIRMODE WG2 meeting in Ispra in November 2009, reflects the use of different models for different type of applications, as reported by the participants during the meeting. Following the discussions in SG4 at the meeting, the tool will address applications for AQ assessment and Planning. Within these type of applications the models considered will include:

- At the regional and urban scales: All model types are accepted with the only requirements that they can provide data with adequate temporal and spatial resolutions. Only gridded results are considered in the procedure.
- At the local/hot-spot scales. Only models which can provide results with sufficient spatial and temporal resolutions will be accepted. Models should also be capable to deliver results for a sufficiently large area (significant part of a city, e.g. Righi et al. 2009) and treat all emission sources (not restricted to traffic). According to the level classification of Sokhi et al.(2008), this implies that only models of at least level III (Gaussian type) are accepted and that simpler approaches (semi-empirical or screening type) are not. This separation between simpler and more complex models at the local scale is done to enable the application of a consistent methodology across the spatial scales (from local to regional).

The pollutants in consideration (at the initial stage) will be O3, NOx and PM (including speciation), see Table 2. The time scale will be set according to the requirements of the AQD, i.e. results from one year model simulations will be the basis for the analysis

Table 1: Type of models used in Member States for AQD implementation (based on participants claim, noted by the asterisk)



Applications for the AQD	Local-Hot spot (dx ~ m)	Urban/Agglom (dx ~ 1- 5 km)	Regional (dx ~ 10 – 50 km)
Compliance / Assessment	Gaussian*** Lagrangian*' Semi empirical * Hybrid models	Gaussian Lagrangian** Eulerian ******	Eulerian****
Mitigation & Planning	Lagrangian** Semi-Empirical** Gaussian**	Eulerian****** Lagrangian*	Eulerian**** Lagrangian
Source Apportionment	Semi-empirical Lagrangian*	Eulerian*** Lagrangian*	Eulerian** Lagrangian
Public information	Gaussian*** Lagrangian	Gaussian* Eulerian ****** Statistical* Lagrangian	Eulerian***** Statistical

Table 2: Pollutants to be considered (at the initial phase of the prototype) and relative AQ monitoring sites to be used in the validation

	Local/hot-pot	**Urban/agglomerate**	**Regional**
Compounds of interest	NOx, PM	O3, NOx, PM (incl. speciation)	O3, PM (incl. speciation)
AQ Monitoring data for validation	Urban traffic & industrial sites	Regional and urban background sites	Regional background sites

Annex B: Core statistical indicators and summary diagrams

1. Model-to-Observations Indicators (MOI)

MOI will express the quality of the air quality model results with respect to observed values, based on a selection of statistical indexes and diagrams.

1.1 Statistical indexes

The following statistical indicators have been identified as most appropriate for the core set:

- *The Pearson Correlation Coefficient (**R**)*:

$$R = \frac{\sum_{i=1}^{N}\left(M_i - \overline{M}\right)\bullet\left(O_i - \overline{O}\right)}{\left[\sqrt{\sum_{i=1}^{N}\left(M_i - \overline{M}\right)^2}\right]\left[\sqrt{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}\right]} \qquad (1)$$

M denotes modelled value, O denotes observed value, N is the number of paired values considered, $\overline{M}$ and $\overline{O}$ denote, respectively, the mean of modelled and observed values. ***R*** ranges from -1 to +1 and indicates the strength of a linear relationship between the two datasets. A value of +1, the so-called ”complete positive correlation” corresponds to all the pairs lying on a straight line with positive slope in the scatter diagram. A value of ***R*** near to zero indicates the absence of linear correlation between the variables. Willmott (1982) discourages the use of this statistical indicator pointing out to its sensitivity of to extreme pairs (outliers).

- *The Root Mean Square Error (**RMSE**):*

Calculated as the square root of the mean squared difference in model-observation pairings with N valid data within a given analysis region and for a given time period.

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

The ideal value is 0. The ***RMSE*** is a good overall measure of model performance. However, since large errors are weighted heavily (due to squaring), few large errors (for example in a small sub-region) may produce a large RMSE even though the errors may be small and quite acceptable elsewhere.

- *The Systematic Root Mean Square Error (**RMSE$_S$**):*

$$RMSE_S = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{M}_i - O_i\right)^2} \qquad (3)$$

where $\hat{M}_i = a + bO_i$ are the regressed model values, estimated from a least square fit to observations. ***RMSE$_S$*** is determined by the distance between the linear regression best-fit line and the 1:1 line, and thus describes the linear bias between model and observations (Willmott et al. 1985)

- *The Unsystematic Root Mean Square Error (**RMSE$_U$**):*

$$RMSE_U = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - \hat{M}_i\right)^2} \qquad (4)$$

RMSE$_U$ is determined by the distance between the data points and the linear regression best-fit line. Thus, ***RMSE$_U$*** is a measure of the scatter about the regression line and can be interpreted as a measure of the potential model accuracy.. Note that $RMSE^2 = RMSE_S^2 + RMSE_U^2$. A good model is considered to have a very low ***RMSE$_S$*** and ***RMSE$_U$*** close to the ***RMSE.***

- *The Index of Agreement (**IOA**)*

$$IOA = 1 - \frac{N \bullet RMSE^2}{\sum_{i=1}^{N}\left(\left|M_i - \overline{O}\right| + \left|O_i - \overline{O}\right|\right)^2} \qquad (5)$$

The perfect value of ***IOA*** is 1. ***IOA*** determines the extent to which magnitudes of $\overline{O}$ (mean) are related to the predicted deviations about $\overline{O}$, and allows for sensitivity toward differences in O and M .

- *The Mean Fractional Bias (**MFB**)*:

$$MFB = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i - O_i}{\left[\left(M_i + O_i\right)/2\right]} \qquad (6)$$

The mean normalized bias ($MNB = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{M_i - O_i}{O_i}\right)$) can become very large when a minimum threshold is not used for the observed/measured data. The mean fractional bias is used as a substitute. The fractional bias for cases with factors of 2 under- and over-prediction are -67 and +67 percent, respectively (as opposed to -50 and +100 percent, when using normalized bias). ***MFB*** is a useful indicator because it has the advantage of equally weighting positive and negative bias estimates. It has also the advantage of not considering observations as the true value. It can be especially useful for assessing performance in PM modelling where species might exhibit close to zero values.
MFB ranges from -200% to +200%.

- *The Mean Fractional Error (**MFE**)*:

$$MFE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|M_i - O_i\right|}{\left[\left(M_i + O\right)/2\right]} \qquad (7)$$

Similarly to the ***MFB***, the mean fractional error ***MFE*** gives equal weight to under- and over-prediction, is not sensitive to a threshold in measured values and does not assume that observations are the truth (i.e. the denominator is the sum of observed and predicted). ***MFE*** ranges from 0% to +200%.

- *The Relative Directive Error (**RDE**)*:

The ***RDE*** has been defined in relation to the AQD, 2008 in order to give a mathematical expression of the "model uncertainty" term in the AQD. Following Denby, 2009 at a single station it is calculated as:

$$RDE = \frac{\left|O_{LV} - M_{LV}\right|}{LV} \qquad (8)$$

where O_{LV} is the closest observed concentration to the limit value concentration (LV) and M_{LV} is the correspondingly ranked modeled concentrations.

- *The Relative Percentile Error (**RPE**)*

RPE, proposed by Flemming and Stern, 2007 is an alternative model error measure for the purposes of the AQD. It is defined as the concentration difference at the percentile p corresponding to the allowed number of exceedances of the limit value normalized by the observation:

$$RPE = \frac{\left|O_p - M_p\right|}{O_p} \qquad (9)$$

1.2 Summary diagrams

Many different types of graphics (scatter plot, quantile-quantile…) have been used for model evaluation, each focusing on particular aspects (concentration distribution, correlation…). In general terms the graphical evaluation used in the benchmarking service will follow these principles:

- Graphical representations which enable the visualization of more than one statistical indicator at a time will be favored (e.g. Taylor diagram, Bugle plots…)

- As proposed by Boylan and Russell, 2006 the performance criteria and goals will be added on the graphics to indicate the minimum expected level of performance for regulatory applications (criteria) and the best performance we can expect from model applications (goal). At the initial phase in the development of the tool these criteria and goals can be set at values based on model applications in the USA and allowing their modification with evolving of the tool in time and with the experience gained in the EU benchmarking exercises.

- Information on the uncertainty of validation data will be provided whenever available to put model results and associated performances in perspective.

- The selection of graphical representations is made on the experience gained in previous model evaluation and inter-comparison exercises. The choice also responds to the need of finding a compromise between the complexity of the evaluation, and the various possibilities of exploring the data in multiple dimensions on one hand and the need of providing to the model users a simple and synthesized model performance diagnostic on the other hand.

In the following of the section we provide a brief overview of some of the summary diagrams which will be used in the benchmarking service. The first five types of diagrams (Taylor, Target, Quantile-quantile, Bugle and Soccer) will be used when observations or known data with which model results can be compared, are available.

- *Taylor diagram*

The Taylor diagram (2001) allows representing different statistical indicators in a single plot – the correlation coefficient ***R***, the centered root-mean-square error ***CRMSE*** (computed by pairing observations and predictions in time and space), and the standard deviation of model values ***SDM*** defined as follows:

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(M_i - \overline{M}\right) - \left(O_i - \overline{O}\right)\right]^2} \qquad (8)$$

$$SDM = \sigma_M = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(M_i - \overline{M}\right)\right]^2} \qquad (9)$$

Figure 3 is a sample Taylor diagram which shows the performance of a single model for a chosen evaluated pollutant (e.g. ozone daily mean concentrations) in different geographical sub-regions (spatial evaluation) and for two seasons (temporal evaluation). A shaded area is used to indicate the expected model performances.

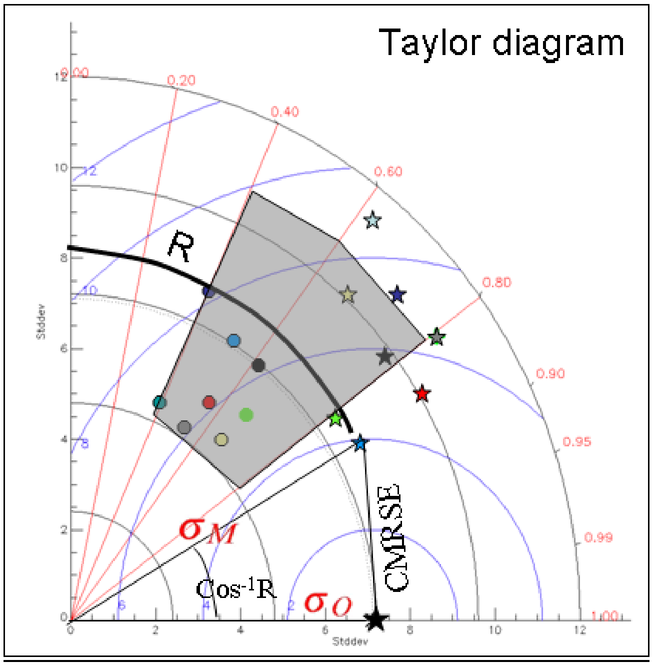


Figure 3 : Sample Taylor diagram displaying one single model performance for two different seasons (different symbols) and for eight different geographical sub-regions (different symbol colors)

Variations of the Taylor diagram use normalized statistics, dividing both the ***CRMSE*** and ***SDM*** by the standard deviation of the observations (σ_o). This makes it possible to plot statistics for different fields (with different units) on the same plot.

- *Target diagram*

The target diagram (proposed very recently by Jolliff et al., 2009) is an evolution to the Taylor diagram that provides information also on the ***BIAS***:

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i) \qquad (9)$$

Figure 4 shows a sample of a Target diagram. In addition to ***SD, BIAS , RMSE (CRMSE) and*** R it can display another evaluation parameter – the model efficiency (***MEF***) score, defined as:

$$MEF = 1 - RMSE^2 \qquad (10)$$

The target diagram displays the model to observation field bias (Y axis) and the model to observation unbiased RMSE (CRMSE) (X-axis). Both axes are normalized by the standard deviation of the observations σ_O. The distance between any point and the origin is then the value of the total RMSE. The outermost marker ($CRMSE/\sigma_O$) establishes also that all points between it and the origin represent positively correlated model and observations, and also have a better than average MEF score. A second marker may be added to indicate another positive R value such as R=0.7, for which all points between it and the origin are greater than R. Finally a dashed line indicates the estimate of average observational uncertainty and further model to data agreement for points between this marker and the origin may not be meaningful. If the CRMSE is multiplied by the sign of the standard deviation observation-model difference then the target diagram provides also information about whether the model standard deviation is larger (X>0) or smaller(X<0) than the observations standard deviation.

The target diagram was proposed for marine ecosystem models, as a novel diagram it has to be checked if it is appropriate also for air quality model evaluation.

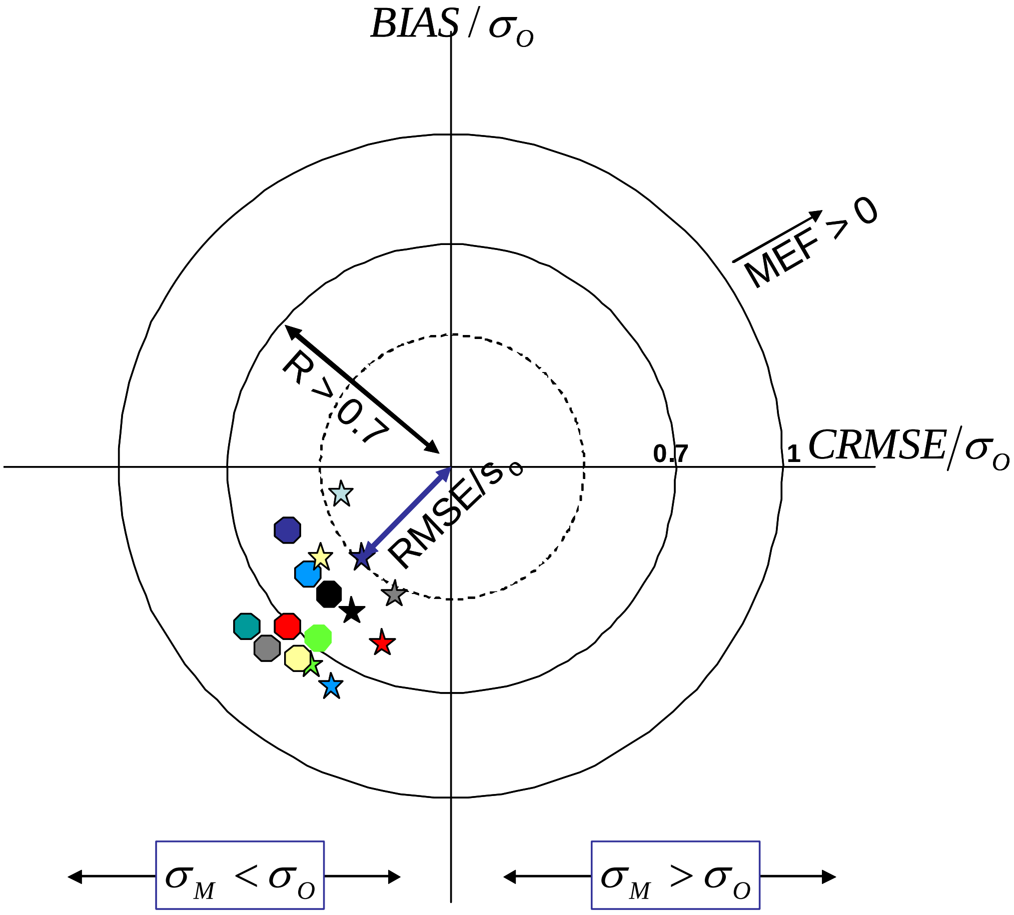


Figure 4: Sample Target diagram. Symbol and colors as in Fig.3, the dashed line represents observation uncertainty

- *Bugle plots*

The bugle plot displays the MFB and the MFE versus "average concentration" along with the corresponding model performance goal and criterion. The goals and criterion are plotted as an exponential curve in function of concentration, as proposed by Boylan and Russell, 2006 in order to take into consideration that goal and criteria values should depend on different factors (e.g. geographical regions, seasons of the year).
The "average concentration"(x-axes in these plots) is the average of the mean modeled and mean observed concentrations.

Figure 5 is an example of bugle plot for PM and shows the MFE and MFB by a single model for two seasons – winter (stars) and summer (circles) in 8 different sub-regions (different symbol color).

Model performance can be identified with the help of the three zones in Figure 5.

Zone 1: the area below the goal line for MFE and between the negative and positive goal lines for MFB corresponds to good model performance;
Zone 2: the area between the goal and criteria lined corresponds to average model performance and would require additional model improvement (e.g. investigating the influence of specific model processes on the model results)
Zone3: the area outside the criteria line(s) corresponds to poor model performance. Extended model evaluation would be necessary (e.g. sensitivity analysis)

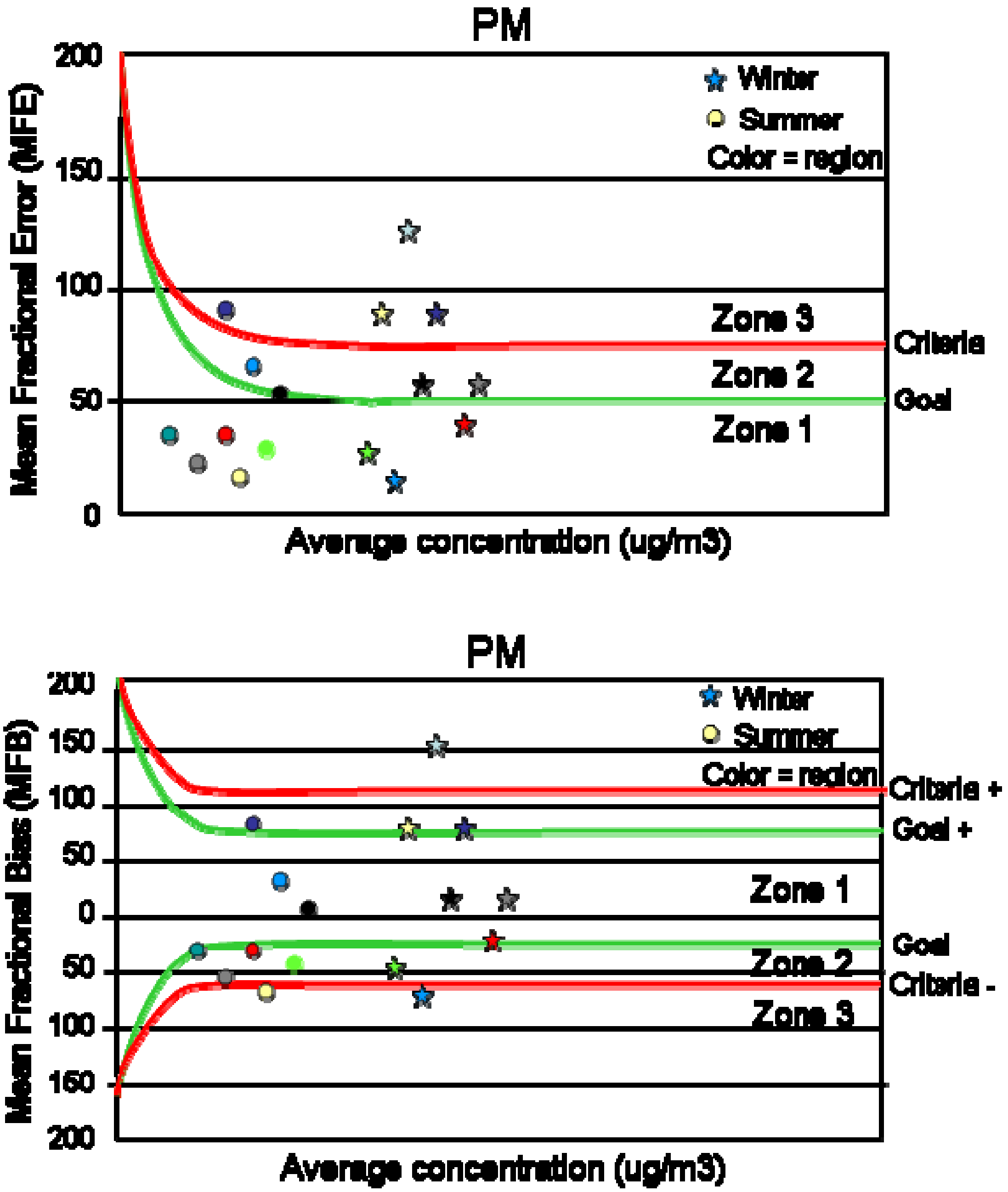


Figure 5: Example of bugle plot for PM concentrations simulated by a single model Symbols (shape and color) are as in Fig.5 Colored lines denote performance goal and criteria values as function of the average concentrations (the mean between modeled and observed values)

o *Quantile-quantile plot*

The Quantile-Quantile (QQ) plot (Figure 6) is used to demonstrate similarity between the distribution of modeled and observed concentrations. The data are unpaired in time and space. Both datasets are sorted from lowest value to highest value and the new pairs are plotted. If the datasets have a similar distribution the plotted values will fall along a 1:1

line. Over the 1:1 line indicates general model over-simulation and under the 1:1 line indicates general model under-simulation. The overall model tendencies are displayed in the QQ plots and the model's general capability to simulate low, average, or high values becomes apparent.

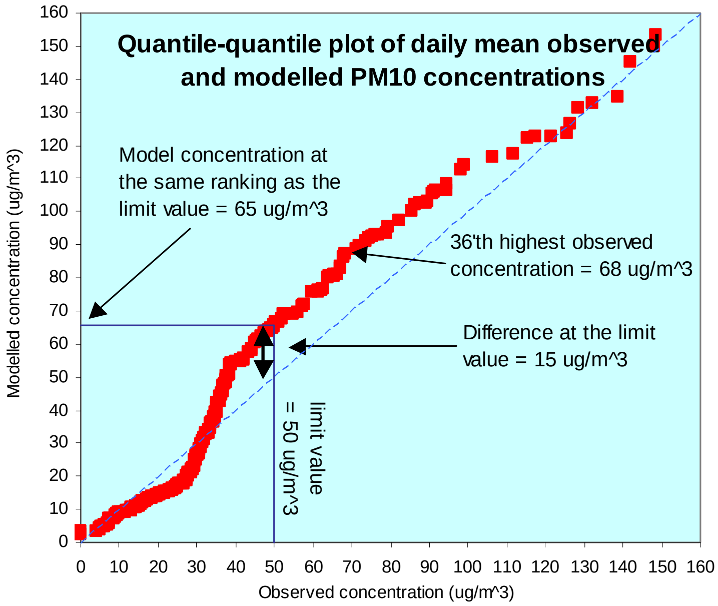


Figure 6: Quantile-Quantile plot of daily mean observed and modeled PM10 concentrations (from Denby et al., 2009)

2. Input Consistency Indicators (ICI)

2.1 Summary indexes

The two following indexes will be used for assessing the meteorological input data.

- *The* ***BIAS***

$$BIAS = \frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)$$

- *The* Gross Error ***(GE)***

$$GE = \frac{1}{N}\sum_{i=1}^{N}|M_i - O_i|$$

- RMSE and IOA , as defined by expressions (2) and (5)

2.2 Summary diagrams

For meteo soccer plots will be used whereas for emissions and boundary conditions the Benchmark-percentile diagram will be used.

o *Soccer plots*

Soccer plots are interesting options as they allow visualizing in a single plot the mean bias and meaning error (fractional, normalized or absolute) along with the respective performance goals and criteria. Example of a soccer plot is illustrated in Figure 7.
Points within the smallest box are thought to be representative of exceptional model performance. Points that are outside the greater box indicate areas in which model performance may need improvement.

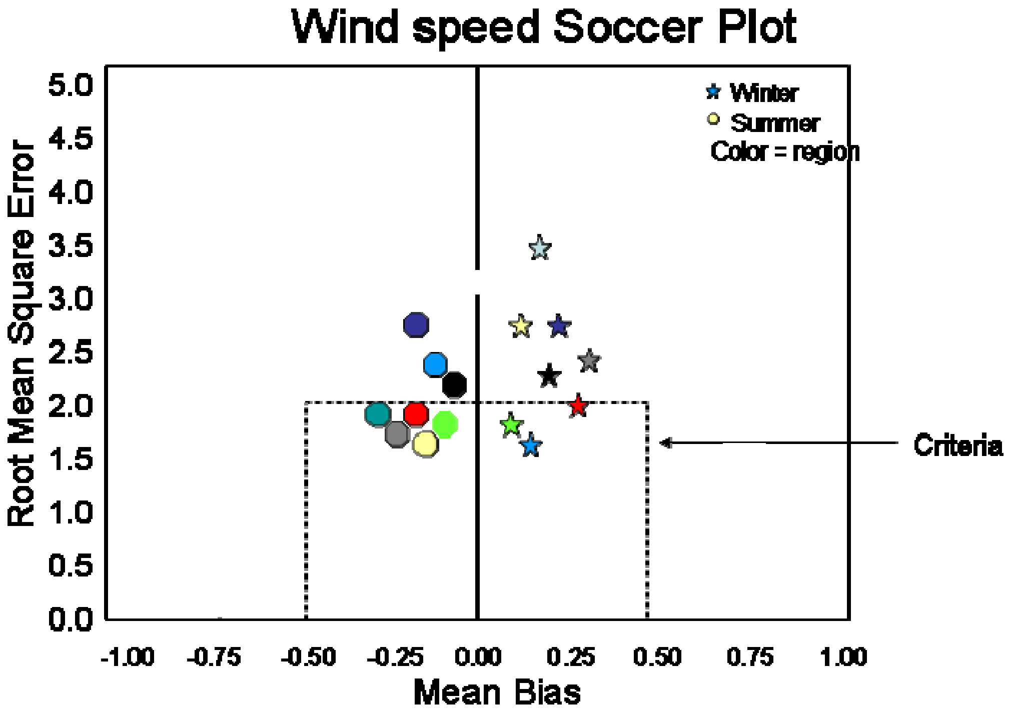


Figure 7: Sample soccer plot illustrating single model performance for wind speed – winter mean (stars), summer mean (circles) for 8 geographical sub-regions (different symbol colors). Dashed line indicate model performance criteria from Table 3

o *"Benchmark-Percentile" plot*

This diagram is useful when no known data or observations are available for comparison with the model results. The diagram provides an overview of the model results averaged over the various geographical sub-domains (or subsets) and temporal periods. Criteria lines are provided to inform the model user on how well his model performs with respect to other results, Figure 8. These criteria lines could be based on percentiles of available model results to avoid giving too much weight to extreme model results.

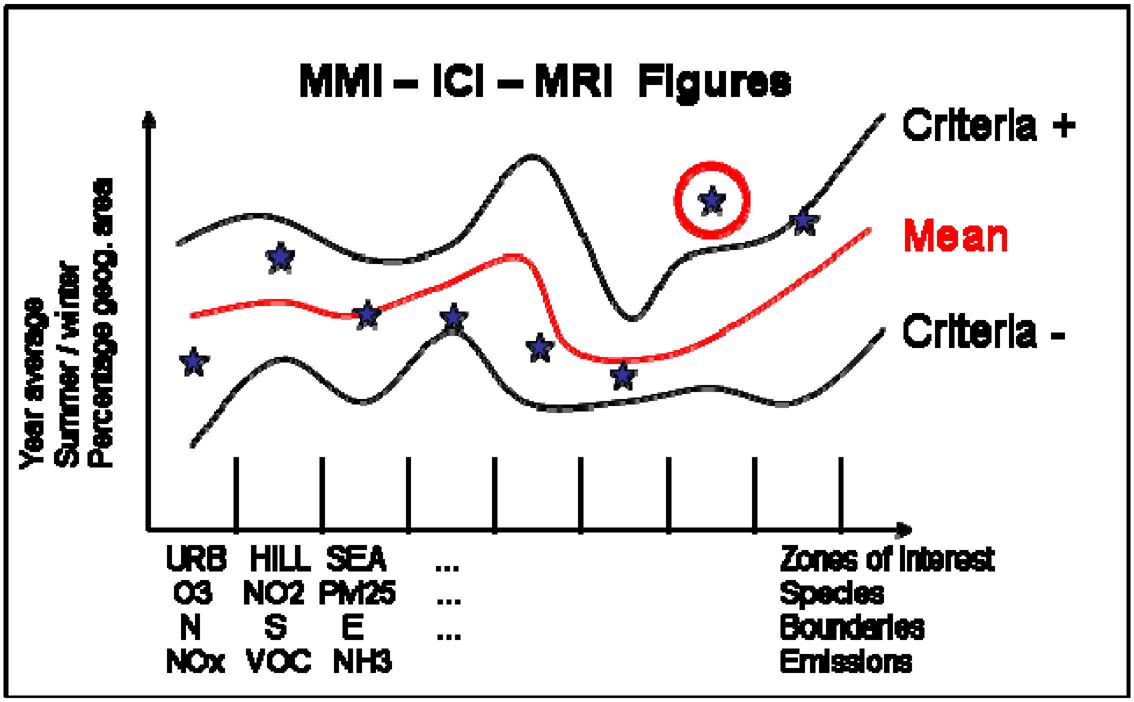


Figure 8: Example of Benchmark-Percentile diagram: Mean model responses or results are plotted against regions or species.

3. Model-Model Indicators and Model-Response Indicators

Since these indicators do not use validation data, no statistical indexes will be used. Benchmark-Ensemble diagrams will be used to provide an overview of the model uncertainty for all species over the entirety of the domain. Criteria and goals will be indicative (since truth is not known!) and it is proposed to fix them according to percentiles of model results, to avoid outlying model results to make these comparison meaningless. The Ensemble-Benchmark plot will be used for the MMI and MRI tests levels where the comparison with other model results is the only benchmark.

Annex C: Reduced set

The following Tables provides a example of how selection of the species, time averaging, and geographical regions could be made in order to compose the reduced set of model output from the full model results.

Model performance indicators		**Time resolution for output submission**	**Local**	**Urban**	**Regional**
MOI		hourly	PM2.5, PM10, NO2	O3, NOx	
MOI		daily		PM10, PM2.5, SO4, NH4, NO3, OC, EC, DST	
MMI		hourly		O3, NO2, NOy, PAN, SO2, CO, NH3, HNO3, H2O2	
		daily		PM10, PM2.5, SO4, NH4, NO3, OC, EC, DST	
		monthly		?	SO4-NO3-NH4 depositions
MRI		monthly		O3, SIA, species ratios	
Variables ICI	Meteo	monthly	WS, WD, MO length, Ustar, min turbu.	WS, WD, T, Q, Kz, PBLheight, Precipitation	
	Emissions		NO2, PPM	NOx, AVOC, BVOC, PPM	

Sub-domains	**Local**	**Urban**	**Regional**
Air quality /Meteo land-use, topography, administrative or environmental targeted areas	**1.** Station 1 **2.** Station 2 **3.** …	1. Urban 2. Rural 3. suburban 4. Po-plain 5. hills 6. Lombardy 7. PM exceed. Area 8. …	1. Po Valley 2. Benelux-Ruhr 3. Medit. sea 4. North – Baltic sea 5. Mountains 6. Iberian Peninsula 7. UK-Ireland 8. France
Boundary conditions	**Same as AQ**	1. East 2. West 3. North 4. South 5. Top	1. East 2. West 3. North 4. South 1. Top
Emissions	**All domain**	1. Urban 2. Rural 3. Lombardy 4. User choice	1. Benelux 2. EU6 3. EU9 4. EU15

Annex D: Performance Goals and Criteria

Table 1 : Performance criteria for meteorological model results (from Emery et al., 2001)

Parameter	Metric	Criteria
Wind speed	RMSE Bias IOA	≤ 2 m/s ≤ ± 0.5 m/s ≥ 0.6
Wind direction	Gross error Bias	≤ 30 deg ≤ ± 10 deg
Temperature	Gross error Bias IOA	≤ 2K ≤ ± 0.5 K ≥ 0.8
Humidity	Gross error Bias IOA	≤ 2 g/kg ≤ ± 1 g/kg ≥ 0.6

Table 2 : Performance criteria and goals for gas- and aerosol phase species (from EPA, 2007 , Boylan and Russell ,2006 and Chemel et al. 2010)

Species	Metric	Criteria	Goal
Main PM constituents (> 30% total mass), PM2.5	MFE MFB	75% ±60%	50% ±30%
Minor PM constituents (< 30% total mass)		Exp variations to reach 100%/200% at 0 concentrations	
Ozone	MFE MFB	45% ±30%	30% ±15%