Strengths and weaknesses of the FAIRMODE benchmarking methodology for the evaluation of air quality models

Monteiro, A.[1], Durka, P.[2], Flandorfer, C.[3], Georgieva, E.[4], Guerreiro, C.[5], Kushta, J.[6], Malherbe, L.[7], Maiheu, B.[8]. Miranda, A. I.[1], Santos, G.[5], Stocker, J.[9], Trimpeneers, E.[10], Tognet, F.[7], Stortini, M.[11], Wesseling, J.[12], Janssen, S.[8], Thunis, P.[13]

[1*]CESAM, Department of Environment and Planning, University of Aveiro, 3810-193 Aveiro, Portugal.

[2]Institute of Environmental Protection – National Research Institute, Poland

[3]Zentralanstalt für Meteorologie und Geodynamik (ZAMG), Section Environmental Meteorology, Vienna, Austria

[4]National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, Sofia, Bulgaria

[5]Norwegian Institute for Air Research (NILU), Kjeller 2027, Norway

[6]The Cyprus Institute, Energy, Environment and Water Research Centre, Nicosia, Cyprus

[7]INERIS, Parc Technologique ALATA, BP2, Verneuil en Halatte 60550, France

[8]VITO, Boeretang 200, 2400 Mol, Belgium

[9]Cambridge Environmental Research Consultants (CERC), United Kingdom

[10]Belgian Interregional Environment Agency (IRCEL), Belgium

[11]Regional Agency for Prevention, Environment and Energy (ARPAE), Emilia-Romagna, Italy

[12]National Institute for Public Health and the Environment, Centre for Environmental Quality, The Netherlands

[13]European Commission, Joint Research Centre (JRC), Directorate for Energy, Transport and Climate, Air and Climate Unit, Via E. Fermi 2749, I-21027, Ispra, VA, Italy

*Corresponding author: alexandra.monteiro@ua.pt, Tel: +351 234370220, Fax: +351 234 370309

Abstract

The Forum of Air Quality Modelling in Europe (FAIRMODE) was launched in 2007 to bring together air quality modellers and users in order to promote and support the harmonised use of models by EU Member States, with emphasis on model application under the European Air Quality Directive. In this context a methodology for evaluating air quality model applications has been developed. This paper presents an analysis of the strengths and weaknesses of the FAIRMODE benchmarking approach, based on users' feedback. European wide, regional and urban scale model applications, developed by different research groups over Europe, have been taken into account. The analysis is focused on the main pollutants under the Air Quality Directive, namely: PM10, NO_2 and O_3. The different case studies are described and analysed with respect to the methodologies applied for model evaluation and quality assurance. This model evaluation intercomparison demonstrates the potential of a harmonised evaluation and benchmarking methodology. A SWOT analysis of the FAIRMODE benchmarking approach is performed based on feedback from users of the tool. This analysis helps to identify the main advantages and value of this model evaluation benchmarking approach compared with other methodologies, in addition to highlighting requirements for future development.

This is a post-peer-review, pre-copyedit version of an article published in Air Quality, Atmosphere and Health. The final authenticated version is available online at: http://dx.doi.org/10.1007/s11869-018-0554-8.

KEYWORDS: air quality modelling; model evaluation; DELTA Tool; benchmarking; FAIRMODE (MQO).

1. INTRODUCTION

Air quality models can be particular relevant tools for the assessment and forecasting of the distribution of pollutants in the atmosphere. As models are increasingly used for policy support, their evaluation becomes an important issue (Solomon 2012). Several documents published by policy-making authorities address this issue trying to develop good practices in terms of model assessment and critical review, e.g. the Standard Guide for Statistical Evaluation of Atmospheric Dispersion Model Performance (ASTM 2005), the US EPA Environmental Model Guidance document (2009), the Guidance on the use of models for the European Air Quality Directive (2008) (Denby 2010) and also the UK government (Defra) report (Derwent et al. 2010).

Model evaluation is, however, a complex procedure involving different steps (scientific evaluation, code verification, model validation, sensitivity analysis etc.), which has been identified already in several scientific studies (e.g. Jakeman et al. 2006; Borrego et al. 2008; Alexandrov et al. 2011). Models applied for regulatory air quality assessment are commonly evaluated on the basis of comparison of modelled results with observations (model validation). This element of the model evaluation process is also known as operational model evaluation (Dennis et al. 2010) with a procedure usually based on statistical performance analysis, using statistical indicators and graphical analysis to determine the skill of an air quality model to reproduce the measured concentrations. Although the comparison between modelled and observed concentrations cannot give a complete insight in the quality and adequacy of the model, it is seen as a good first screening in the model evaluation process (Irwin et al. 2008; Derwent et al. 2010; Carnevale et al. 2015).

FAIRMODE is the Forum for Air Quality Modelling in Europe (http://fairmode.jrc.ec.europa.eu/), organized around four main working groups (WGs), following 4 themes: assessment (including uncertainty analysis), emissions, source apportionment and planning. In the WG1 (Assessment) a methodology to benchmark model performances according to a common scale and common template has been the focus for several years. In this context, modelling quality objectives (MQO) based on measurement uncertainty have been discussed and the methodology is consolidated in the so-called DELTA Tool. This methodology has been extensively tested by the FAIRMODE community.

In this framework a procedure for the benchmarking of air quality models was suggested and discussed (Thunis et al. 2012a, 2012b; Pernigotti et al. 2013; Thunis et al. 2013). It aims at harmonizing the diagnostics and reporting of air quality model performances, focusing on the pollutants mentioned in the EU Air Quality Directive (AQD) (2008) and addressing all relevant spatial scales (from local to regional). This procedure provides information about the quality of the model results, indicating expected model performances and highlighting the strengths and weaknesses of a specific model application. This is particularly important in order to assess whether or not a model is of sufficient quality for policy support. In this context, Thunis et al. (2012a) proposed a 'Modelling Quality Objective' (MQO) based on an indicator defined as the ratio of the root mean square error (RMSE) of measured and modelled concentrations to the measurement uncertainty. This objective was further revised and elaborated in order to assign complementary 'Modelling Performance Criteria' (MPC) (Thunis et al. 2013). In addition, this procedure was discussed extensively during FAIRMODE meetings, and the associated software (DELTA Tool) was applied by air quality model and environmental experts from a wide range of EU countries, providing thus sufficient basis for critically assessing the proposed methodology and its application.

The motivation for the work presented here is primarily to provide a critical review of the FAIRMODE evaluation methodology by a broad user community. To this end, applications of the benchmarking methodology by a number of air quality model users were gathered and analysed, highlighting both the main advantages of, and any issues with, the proposed methodology. The user feedback was compiled using a SWOT analysis. Information from this user feedback and the SWOT analysis will allow the methodology to be extended and refined with the aim of standardising the use of this model evaluation approach in the context of the European AQD.

The structure of the paper is as follows: the benchmarking methodology and the performance report are detailed in Section 2. The description and analysis of the gathered modelling applications are included in Section 3. The SWOT analysis is presented in Section 4 and remaining open issues are summarised in Section 5.

2. THE BENCHMARKING METHODOLOGY

2.1 Modelling Quality Objective (MQO)

The FAIRMODE benchmarking methodology is aimed at evaluating the performance of an air quality model application through comparison between modelled and measured data. It is primarily based on the calculation of the Modelling Quality Indicator (MQI), taking the measurement uncertainty into account. Further insight into modelling performance is provided by supplementary Modelling Performance Indicators (MPI). The methodology has been incorporated into a software package (DELTA Tool) that facilitates results visualization.

The Modelling Quality Indicator (MQI) is defined as a statistical indicator calculated on the basis of measurements and modelling results in order to describe the discrepancy between the observations and model predictions. The Modelling Quality Objective (MQO) is the criterion for the value of the MQI; specifically, the MQO is said to be fulfilled if the MQI is less than or equal to unity.

In addition to the MQI, several Modelling Performance Indicators (MPI) are defined. The MPI describe various aspects of the discrepancy between measurement and modelling results: correlation, bias and normalised standard deviation. Furthermore, MPI are also defined to assess model performance in terms of spatial variation. Similarly to the MQI and MQO described above, the Modelling Performance Criteria (MPC) are the criteria that the MPI are expected to fulfil. Fulfilment of the MPC is a necessary, but not sufficient condition to ensure that the model is fit for purpose. For this, both the MPC and the MQO need to be fulfilled simultaneously.

The main elements of the derivation of the MQI are summarised below and described in detail in Thunis et al. (2012b). The MQI is defined as the ratio of the model (Mi) - measured (Oi) bias to a quantity proportional to the measurement uncertainty. It is calculated as:

$$\mathrm{MQI} = \frac{|\mathrm{O_i} - \mathrm{M_i}|}{\beta \mathrm{U_{95}(O_i)}} \tag{1}$$

Where index i denotes a given time (hour or day), $U_{95}(O_i)$ is the 95th percentile highest value of the measurement uncertainty and β is a coefficient of proportionality linked to the MQO stringency. β is arbitrarily set to 2, thus allowing the deviation between modelled and measured concentrations to be twice the measurement uncertainty in the current formulation.

The MQO requires MQI to be less than or equal to 1 MQO:MQI≤1.

Equation (1) can then be used to generalise the MQI to a time series:

$$MQI = \frac{\mathrm{RMSE}}{\beta RMS_U} \quad \text{and MQO: MQI} \leq 1 \tag{2}$$

Figure 1 illustrates the concept of model and measurement uncertainty on the basis of modelled and observed concentrations for a selected time period. In Figure 1, the MQO is fulfilled, for instance, on days 3 to 10 whereas it is not fulfilled on days 1, 2 and 11. This condition $|\mathrm{O_i} - \mathrm{M_i}| \leq \mathrm{U_{95}(O_i)}$ indicates also when model-observed differences are within the measurement uncertainty (e.g. days 5 and 12 in Figure 1).

Figure 1. Example for a PM10 time series: measured (bold black) and modelled (bold red) concentrations are represented for a single station. The grey shaded area indicates the measurement uncertainty and the dashed black lines represent the MQI limits (proportional to the measurement uncertainty). Modelled data fulfilling the MQO must be within the dashed lines.

With this MQO formulation, the RMSE between observed and modelled values (numerator) is compared to a value ($\mathrm{RMS_U}$) representative of the maximum allowed measurement uncertainty (denominator). The value of β determines the stringency of the MQO.

Thunis et al. (2013) showed that the root mean square of the measurement uncertainty, RMS_U, can be expressed as:

$$RMS_U = U_{95r}^{RV}\sqrt{(1-\alpha^2)(\bar{O}^2 + \sigma_O^2) + \alpha^2.RV^2} \tag{3}$$

in which $\bar{O}$ and σ_0 are the mean and the standard deviation of the measured time series, respectively, U_{95r}^{RV} is the standard measurement uncertainty around the reference value (RV) for a reference time interval (e.g. the daily/hourly limit value) and α is the non-proportional fraction (between 0 and 1) of the measurement uncertainty around that reference value (see Pernigotti et al. 2013 for more details).

For air quality models that provide yearly averaged pollutant concentrations, the MQI is modified so that the mean bias between modelled and measured concentrations is normalised by the expanded uncertainty of the mean measured concentration at the 95th percentile:

$$MQI = \frac{|\bar{O} - \bar{M}|}{\beta U_{95}(\bar{\mathrm{O}})} \quad \text{and } MQO\colon MQI \leq 1 \tag{4}$$

For this case, Pernigotti et al. (2013) derived the following expression for the uncertainty of the yearly averaged observation:

$$U(\overline{O}) = U_{95r}^{RV}\sqrt{\frac{(1-\alpha^2)}{N_p}\overline{O}^2 + \frac{\alpha^2 . RV^2}{N_{np}}} \tag{5}$$

where N_p and N_{np} are two coefficients that are used only for annual averages and that account for the compensation of errors (and therefore a smaller uncertainty) due to random noise and other factors like periodic re-calibration of the instruments. Details on the derivation of (5) and in particular the parameters N_p and N_{np} are provided in Pernigotti et al. (2013).

Table 1 summarises values currently used in the MQI expression.

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

As the AQD requirements have been followed when defining all statistical indicators, the MQO must be fulfilled for at least 90% of available stations. The practical implementation of this approach results in the calculation of the MQI associated with each station, followed by the ranking of the stations in ascending order to infer the 90th percentile value according to the following linear interpolation (for 'nstat' station):

$$MQI_{90th} = MQI(stat_{90}) + [MQI(stat_{90}+1) - MQI(stat_{90})] * dist \tag{6}$$

where stat90 = integer(nstat*0.9) and dist= $[nstat * 0.9 - integer(nstat * 0.9)]$. If only one station is used in the benchmarking, $MQI_{90th} = MQI(station) * 0.9$. A similar approach is used to calculate the corresponding model uncertainty (Thunis et al., 2013); the MQO is then expressed as:

$$MQO: \quad MQI_{90th} \leq 1 \tag{7}$$

2.2. Reporting model performance

The presented methodology was embedded into an IDL software package – the DELTA Tool (Thunis et al. 2012a). The tool takes as input pairs of measurement and modelled data at a given location. It allows the user to perform two types of analysis: exploratory, looking at various statistical parameters, diagrams, pollutants and time intervals and benchmarking, when preselected model performance indicators for some regulated pollutants are compared to modelling quality objective and model performance criteria.

Benchmarking reports are currently produced for the hourly NO_2, the 8h daily maximum O_3 and daily PM10 and PM2.5. These benchmarking reports are different for hourly (or daily) model values and for yearly average model results. Details of these two types of reports are presented below.

2.2.1. Reporting for hourly/daily model results

The benchmarking report consists of a Target diagram followed by a summary table (see Figure 2). The MQO as described by Eq (2) is used as the main indicator. The main graphical view for the MQO is the Target diagram constructed with statistical indicators normalised by the measurement uncertainty. In this diagram, the MQI represents the distance between the origin and a given station point. The MQO for the target indicator is set to unity (green circle) regardless of spatial scale and pollutant and it is expected to be fulfilled by at least 90% of the available stations. Additional details on the interpretation of the diagram can be found in Thunis et al. (2012a).

The MQI associated with the 90th percentile worst station is calculated (Eq 6) and indicated in the upper left corner; this value is used as the main indicator in the benchmarking procedure and should be less than or equal to one. The uncertainty parameters used to produce the diagram are listed on the top right-hand side, with the resulting model uncertainty also being displayed on the right (in blue font). The value of the MQI obtained, if data averaged over a year, is given as 'Y'.

A summary statistics table provides a complementary source of information to the MQO in order to identify model strengths and weaknesses (Figure 2). The first two rows provide information about the observed annual means calculated from the hourly values and the number of exceedances for the selected stations. The following three rows provide an overview of the temporal statistics for bias (row 3), correlation (row 4) and standard deviation (row 5) in addition to information relating to the ability of the model to capture the highest range of concentration values (row 6). Stations where the model performance criterion is fulfilled lie within the green and the orange shaded areas. If a point falls within the orange shaded area, the error associated with the particular statistical indicator is dominant. The next two rows provide an overview of spatial statistics for correlation and standard deviation. For all indicators, the second column with the coloured circle provides information on the number of stations fulfilling the performance criteria: in line with the AQD, the circle is coloured green if more than 90% of the stations fulfil the criterion and red if the number of stations is lower than 90%.

Figure 2. Example of benchmarking report for hourly model results over one year. The following symbols are used: R (correlation), SO (standard deviation), CRMSE (Centered root mean square error), Exceed (number of exceedances above a given threshold (50 $\mu g.m^{-3}$)), Corr Norm (normalised correlation), Std dev norm (normalised standard deviation)

2.2.2. Reporting for yearly averaged model results

For the evaluation and reporting of yearly averaged model results, a Scatter diagram is used to represent the MQI instead of the Target plot. The report then consists in a Scatter diagram followed by the Summary Statistics (Figure 3).

The MQI (Eq 4) for yearly averaged results (i.e. based on the bias) is used as main indicator. In the Scatter plot, it is used to represent the distance from the 1:1 line. The summary statistics table includes the observed means for the selected stations (first row), information on the fulfilment of the bias-based MPI for each selected stations (second row) and an overview of spatial statistics for correlation and standard deviation (third and fourth rows).

Figure 3. Example of Benchmarking report based on yearly averaged model results. The following symbols are used: OBS (Observations), MOD (model results), Corr Norm (normalised correlation), Std dev norm (normalised standard deviation)

3. COLLECTION OF USERS' EXPERIENCE

Within the FAIRMODE community, a questionnaire was circulated in order to collate users' feedback in relation to their experiences in terms of model evaluation, both *before* and *after* the development of the FAIRMODE common model evaluation methodology. A total of 11 case studies were compiled, with applications varying in purpose (beyond the assessment for AQD), model type and range of pollutants. Table 2 summarises the 12 cases with a brief description, which is then further analysed, in terms of results and users experience/feedback.

Table 2. Description of the case studies using the FAIRMODE model evaluation.

The case studies correspond to 11 different European countries (UK, France, Portugal, Bulgaria, Norway, Poland, Italy, The Netherlands, Belgium, Cyprus and Austria), and to the application of nine different models, mainly configured by research modelling groups (with their own meteorological and emission input data) and applied to different years. The purpose of the model evaluation case studies includes model validation exercise for air quality assessment/forecast and/or research projects, with a few particular cases that focus on air quality plans. In 9 of the cases (80%) the models used are mesoscale/regional models applied over large areas or over the entire country with high resolutions ($\leq$ 6x6 km^2). The other three cases, namely the ADMS-Urban (London), OPS+SRM (RIVM) and EPISODE (Olso) models, are applied to urban areas. With the exception of the OPS (The Netherlands) all models produce hourly data. Regarding the pollutants, NO_2 is the focus of all case studies, followed by PM_{10} and O_3 in 80% of the cases. Besides that, PM2.5, and SO_2 are also included in 3 of the cases. Only two case studies use data assimilation approaches, with a different method being used for each.

In order to evaluate the differences between this methodology and the previous evaluation practices, Table 3 describes how users performed model evaluation before adoption of the FAIRMODE evaluation framework.

Table 3. Model evaluation procedure before the FAIRMODE evaluation framework

The comparison in Table 3 shows that the majority of the case studies are applications of mesoscale/regional models and only consider background stations for the model evaluation procedure. The three case studies with urban scale models include all the stations in the analysis i.e. roadside and kerbside. Further, three statistical parameters are consistently used for model evaluation: BIAS (Fb), RMSE (NMSE) and R; these are all included in the FAIRMODE model evaluation procedure. No threshold values for statistical indicators have been applied for none of the case studies, which suggests that the MQO procedure and the associated MPC can bring an added-value to these previous model evaluation practices.

Regarding the use of plots, the Scatter diagram is mentioned by all groups; in addition, others plots are used such as the Taylor diagram, contour plots and Quantile-Quantile (QQ) plots.

4. SWOT analysis

A SWOT analysis was set up based on the 12 case studies that applied the FAIRMODE framework (Table 3) in order to identify the main Strengths (characteristics of the approach that give it an advantage over others), Weaknesses (characteristics that place the approach at a disadvantage relative to others), Opportunities (elements that the approach could exploit to its advantage) and Threats (elements that could cause trouble for the approach) of this model evaluation scheme. This SWOT analysis is presented below:

4.1. Strengths (S)

4.1.1 A deep insight into the performance of a model application, combining innovative and traditional indicators

- The MQO is based on a comprehensive statistic (MQI) that accounts both for model performance and measurement uncertainty, which is an improvement on previous assessment methods that usually neglect uncertainty. Taking into account uncertainties (modelling as well as measurement) in this methodology is evidently a realistic approach to evaluating model performance. The variety of quality and performance indicators provides information on different aspects of the modelling.
- The MQI integrates several indicators in one (RMSE, BIAS & R). The Target plot is well visualized, clear and summarizes all of the individually used indicators into one graph (in contrast to comparing RMSE, BIAS & R separately), which facilitates understanding for all, not only specialists in air quality field. The synthetic way of comparing modelling performance between different stations or different modelling outputs is an additional asset. Identifying stations where a model is underperforming (MQI>1) is a straightforward process and the diagram immediately indicates if this is due to issues related to correlation, bias or standard deviation.
- The methodology provides Model Performance Criteria (MPC) that set limits for acceptable values for RMSE, BIAS and R (i.e. MPI) taking into account the measurement uncertainty.
- The methodology applies the 90th percentile concept for the MQI and MPI. By using the 90th percentile concept, the methodology is consistent with the EU Directive 2008/50 allowance for noncompliance of the MQO for one out of 10 monitoring stations. By re-

working this rule as a percentile, the restriction may be applied even for cases where the number of stations differs from n x 10

- The summary statistics table provides additional useful information that is not accounted for in the MQI, for example, the model's ability to predict high percentile concentrations.

4.1.2 A common EU methodological framework

- This new evaluation methodology allows use of a standard methodology for the evaluation of air quality modelling results in the frame of the EU Directive 2008/50, which is accepted throughout Europe. The methodology is open and publically available, proposes common plots and indicators for the analysis, therefore providing useful and ready-to-use tools that facilitate the task of smaller modelling groups when evaluating their modelling exercises. It also triggers a concerted discussion with other modelling groups.
- The methodology is well documented, easy to apply and works with data from any model, without taking into consideration differences such as domain size, output resolution, model output format etc.
- The methodology is useful for a wide range of target groups: policy makers at all levels, as well as for people other than experts. It also allows air quality modellers to dig further into statistical indicators and point out where their air quality model can be improved.
- A common methodology triggers discussions among groups from all over Europe (modelling communities), leading to a better general acceptance of the need for a MQO and thus can support the refinement of the methodology and the possibility to make recommendations for the revision of the AQD. It is a solid example of the EU consensus model: the proposed methodology is the result of numerous discussions and iterations within the European air quality modelling community.

4.2. Weaknesses (W)

4.2.1 Statistical issues

- The methodology still suffers from inconsistencies between the annual and hourly/daily mean indicators. The MQO for hourly/daily mean values is often attained whereas it is not the case for the annual values. This can be hard to explain when one has to convince policymakers to use models.
- The MQO accounting for measurement uncertainty is a novelty, but more research evidence is necessary to check sensitivity to uncertainty parameters (Carnevale et al.

2014). Not all of the parameters used to construct the MQI are well defined (e.g. a value for measurement uncertainty of PM2.5 has been arbitrarily modified; the N_p and N_{np} values were chosen to be the same as for PM10 because of the lack of available measurements). The methodology assumes symmetric confidence intervals around the observations (Oi +/- U) which, for lognormal distributions of observations, is probably less correct at lower concentrations. The representativeness error is not included in the measurement uncertainty.

- The MPC for high percentiles currently does not consider the timing of the extreme events. Therefore, the MPI_{perc} might be <= 1 for the wrong reason.

4.2.2 Current limitations

- By default the MQI does not include parameters for NO_x as it is not included in the AQD, but it is an important indicator of dispersion model performance and accuracy of the underlying emissions.
- The station representativeness for the scale of the model is often based on expert opinion (the choice of the stations can influence conclusions on modelling quality). No (consensus) methodology yet exists to determine which measurements should be used to evaluate model performance.
- A standardised way of dealing with data assimilated assessments is still missing in the methodology. Indeed the MQI methodology treats air quality assessments with and without data assimilation fusion equally, which is not always desirable when comparing results from different models.

4.3. Opportunities (O)

4.3.1 Increasing and improving the use of air quality models

- The target plot is an easy-to-use assessment of models that can promote the use of models for different applications (local to European level). It can provide guidance for Member States who have yet to choose assessment models. It has the potential to increase the application, quality and harmonisation of models throughout Europe. With this methodology, authorities can easily make it a requirement to meet the MQO when requesting modelling support for AQD applications.
- The model results can easily be compared. The approach helps defining the highest performing model for each pollutant. If the same model has been used to model air quality in different regions, the MQO template is a useful way to assess model

performance and may help to highlight inconsistencies in model inputs or configurations.

- The methodology has all the elements to elaborate reports tailored to different target groups.

4.3.2 Extension to other pollutants or modelling applications

- The methodology should be extended to all AQD regulated pollutants (for instance CO, SO_2, benzene …)
- A section for AQ assessment prepared to work with all AQD thresholds should be considered;
- This MQO methodology could be extended to support the evaluation of models when used to assess the impacts of of air quality plans (i.e. for the evaluation of model emission reduction scenarios). Other types of indicators need then to be defined. Thunis et al. (2015) have proposed to use indicators such as "potency" and "potential" for this purpose.
- The approach to consider forecasting applications with specific model skill/scores should be generalised (this is currently in preparation).

4.3.3 Extension to other communities

- The FAIRMODE community can be used as an example of joint cooperation on common subject for other environmental fields. There is an opportunity to export this unique EU-consensus methodology outside of the EU or to use a similar approach in other environmental fields.

4.4. Threats (T)

4.4.1 Doubts on the robustness of the methodology

- The MQO should not be too relaxed because in this case there is no added value from the use of such a tool; conversely, it needs to reflect a realistic attainable model quality. It is important and challenging to obtain a correct level that allows characterisation by a single MQI and MQO.
- The definitions of the annual and hourly MQI values are similar, but assessing the results of a model that calculates hourly values using both the annual and hourly MQI

approaches gives different results. Diverging conclusions about MQO attainment could be difficult to interpret and communicate.

4.4.2 Barriers to using the methodology

- There is a risk that the methodology is not applied if the community cannot force this work through EU legislation.
- The methodology is still evolving. There is therefore a risk of comparing performance templates obtained with different versions of the MQO.
- This methodology should be used with caution when a limited number of stations exist (since the MQO must be fulfilled for at least 90% of available stations). This is often the case for urban models with few measurement stations available.
- Habits are hard to change, many users probably already have a set of indicators (namely BIAS, correlation factor and RMSE) that they use regularly and are accustomed to.

Regarding strengths, the user community states that this methodology is by now widely used and with promising results and added-values, namely: recognition of a standard methodology for evaluation of modelling results in the frame of the EU Directive, integration of the most essential quality indicators (and a comprehensive MQO and MPC taking into account uncertainties); the performance report is easy to interpret for both policy makers and model experts; continuous updates and revisions. Nevertheless, several problems were recognised, mainly: inconsistency of the annual/daily mean MQO; the mismatch between the spatial representativeness of the station and the model grid resolution; definition of arbitrary parameters (no clear definition and use of measurement uncertainty); and the need of updated guidance documents.

Opportunities and threats were also identified. Some of them are already being considered along the next and future developments planned. Others are recognised as open issues and need further research, analysis and testing before a proper solution can be put forward. In the next section these open issues - and how they will be handled - are detailed.

5. OPEN ISSUES & STRATEGIES

The section below discusses the topics that are identified as opportunities or threats in the SWOT analysis. Some of them do not currently have a consensus but merit further consideration, namely: the use of data assimilation; the possible lack of spatial representativeness of the monitoring station (or the inadequacy between the spatial representativeness of the measurement and the grid resolution of the model); changes in measurement uncertainty; performance criteria for high percentiles; data availability and also the application of the procedure to other parameters.

- Data assimilation:

The AQD suggests the integrated use of modelling techniques and measurements to provide suitable information about the spatial and temporal distribution of pollutant concentrations. However, when validating these integrated data sets, different approaches can be found in the literature. All of them are based on dividing the set of measurement data into two groups, one for the data assimilation or data fusion (also called the "assimilation set") and one for the evaluation of the integrated fields (the "validation set"). The challenge is to select, in a harmonised way, the set of validation stations. FAIRMODE is currently investigating which of the methodologies is most robust and applicable in operational contexts.

- Station representativeness:

In the current approach, only the uncertainty related to the measurement device is accounted for. However, as described in Janssen et al. (2012) (and also Kracht, 2018 and Martin et al., 2014) another source of divergence between model results and measurements is linked to the lack of spatial representativeness of a given measurement station (or to the mismatch between the model grid resolution and the station representativeness). The formulation proposed for the MQO and MPC may be extended to account for the lack of spatial representativeness when quantitative information on the effect of a station (type) representativeness on measurement uncertainty becomes available.

- Performance criteria for high percentile values:

The model quality objective described above provides insight on the quality of the model average performances but does not provide information on the model capability to reproduce extreme events (e.g. exceedances). For this purpose, a specific MQO indicator is proposed but further testing and fine-tuning is required. It is also under debate whether the timing of the exceedance has to be taken into account, as the AQD states that the timing of events can be ignored.

- Inconsistency between the hourly and annual approach:

FAIRMODE's evaluation framework is designed for models that produce hourly output as well as for model that only produce annual averages. However, the analysis made clear that the MQO for the hourly approach is less strict than the annual one. Discussions are currently taking place to assess the need for models producing hourly/daily results to fulfil both MQO (annual and hourly/daily). These hourly/daily models can indeed be aggregated to produce yearly average assessments that would need to fulfil the yearly MQO.

- Data availability:

Currently Data Quality Objectives are defined in the AQD with a minimum data capture percentage depending on the pollutant (to guarantee a sufficient number of stations), the time period/coverage and type of station, with additional rules for including calibration and maintenance of the instrumentation. Nevertheless, other criteria can be found in the European Environment Agency reports. Harmonisation should be done in order to use the most adequate requirements.

- Application of the procedure to other parameters:

Currently only particulate matter (PM10 and PM2.5), O_3 and NO_2 have been considered but the methodology could be extended to other pollutants such as heavy metals and polyaromatic hydrocarbons which are considered in the Ambient Air Quality Directive 2004/107/EC. Besides that, the procedure can off course be extended to other variables including meteorological data as proposed in Pernigotti et al. (2013).

5. CONCLUSIONS

The FAIRMODE benchmarking approach for air quality models evaluation was developed over the last years and has been applied and tested by several Member States, regarding European, regional and urban scale model applications. This paper presents the experiences of the different modelling teams and evaluates the benchmarking approach based on the user feedback. The analysis was focused on the main pollutants under the Air Quality Directive, namely: PM10, NO_2 and O_3. A SWOT analysis was built in order to identify the main advantages and value of this model evaluation benchmarking approach compared with other methodologies, in addition to highlighting requirements for future development. The main strengths recognise the success on promoting harmonised reporting relevant to AQ model applications under AQD and the integration of the most essential quality indicators. The weaknesses identified are mainly related to inconsistency of the annual/daily mean MQO and no clear definition and use of measurement uncertainty. Finally, some strategies are elaborated regarding the main open issues and threats identified.

ACKNOWLEDGEMENTS

Thanks are due for the financial support to CESAM (UID/AMB/50017 - POCI-01-0145-FEDER-007638), to FCT/MCTES through national funds (PIDDAC), and the co-funding by the FEDER, within the PT2020 Partnership Agreement and Compete 2020. This work was partly performed within FAIRMODE (http://fairmode.ew.eea.europa.eu/), the community members are acknowledged for their contribution.

REFERENCES

Adriaenssens S, Trimpeneers E, (2015) Transnational model intercomparison and validation exercise in North-West Europe. Interregional Environment Agency Belgium (IRCEL). Final report of the Joaquin EU-Interreg IVB project.

Alexandrov GA, Ames D, Bellocchi G, Bruen M, Crout N, Erechtchoukova M, Hildebrandt A, Hoffman F, Jackisch C, Khaiter P, Mannina G, Mathunaga T, Purucker ST, Rivington M, Samaniego L (2011) Technical assessment and evaluation of environmental models and software: letter to the Editor. Environ Modell Softw 26 (3): 328-336.

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. Environ Int 34: 613-620.

Carnevale C, Finzi G, Pederzoli A, Pisoni E, Thunis P, Turrini E, Volta M (2014) Applying the delta tool to support the Air Quality Directive: evaluation of the TCAM chemical transport model. Air Qual Atmos Hlth 7 (3): 335–346.

Carnevale C, Finzi G, Pederzoli A, Pisoni E, Thunis P, Turrini E, Volta M (2015) A methodology for the evaluation of re-analyzed PM10 concentration fields: a case study over the PO Valley. Air Qual Atmos Hlth 8 (6): 533–544.

Denby B (2010) Guidance on the Use of Models for the European Air Quality Directive (ETC/ACC No. version 6.2). In: A Working Document of the Forum for Air Quality Modelling in Europe FAIRMODE.

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. Environ Fluid Mech 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.

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. Atmos Environ 42: 5403-5412.

Jakeman A J, Letcher R A, Norton J P (2006) Ten iterative steps in development and evaluation of environmental models. Environ Modell Softw 21(5): 602-614.

Janssen S, Dumont G, Fierens F, Deutsch F, Maiheu B, Celis D, Trimpeneers E, Mensink C, (2012) Land use to characterize spatial representativeness of air quality monitoring stations and its relevance for model validation. Atmos Environ 59: 492-500.

Kracht O. (2018) Spatial Representativeness of Air Quality Monitoring Sites - Outcomes of the FAIRMODE / AQUILA Intercomparison Exercise, JRC Technical report (in press)

Georgieva E., Syrakov D., Prodanova M., Etropolska I, Slavov K. (2015) Evaluating the performance of WRF-CMAQ air quality modelling system in Bulgaria by means of the DELTA tool International Journal of Environment and Pollution, 57 (3/4), 272-284.

Martin F. Fileni L., Palomino I., Vivanco M.G., Garrido J.L. (2014) Analysis of the spatial representativeness of rural background monitoring stations in Spain. Atmospheric Pollution Research 5, 779-788.

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

Ribeiro I., Monteiro A., Miranda A.I., Fernandes A.P., Monteiro A.C., Lopes M., Borrego C. (2014). Air quality modelling as a supplementary assessment method in the frame of the European air quality directive. International Journal of Environmental Pollution 54, Nos. 2/3/4, 262 – 270.

Solomon P A (2012) Introduction: Addressing air pollution and health science questions to inform science and policy. Air Qual Atmos Hlth 5 (2): 149–150

Stidworthy A., Jackson M., Johnson K., Carruthers D., Stocker J. (2017) Evaluation of Local and Regional Air Quality Forecasts for London. In Proc. 18th Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes, Bologna, 9-12 October 2017.

Stortini M., Agostini C., Maccaferri S., Amorati R. (2017) RIAT+ and pair2020: an integrated assessment tool useful for air quality planning. In Proc. 18th International conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes. Bologna, Italy, October 9-12. Submitted to the IJEP Special Issue

Thunis P, Georgieva E, Pederzoli A (2012a) A tool to evaluate air quality model performances in regulatory applications. Environ Modell Softw 38: 220-230.

Thunis P, Pederzoli A, Pernigotti D (2012b) Performance criteria to evaluate air quality modeling applications. Atmos Environ 59: 476-482.

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

Thunis P, Pisoni E, Degraeuwe B, Kranenburg R, Schaap M, Clappier A (2015) Dynamic evaluation of air quality models over European regions. Atmos Environ 111: 185-194.

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

Veldeman N., Maiheu B., Lefebvre W. et al. (2016) Activity report for 2015 reference task on air quality modelling in Flanders. VITO Report nr. 2016/RMA/R/0582 (in Dutch).

	β	$U_{95,r}^{RV}$	RV	α	N_p	N_{np}
NO_2	2.00	0.24	200 µg.m^{-3}	0.20	5.2	5.5
O_3	2.00	0.18	120 µg.m^{-3}	0.79	11	3
PM10	2.00	0.28	50 µg.m^{-3}	0.13	30	0.25
PM2.5	2.00	0.36	25 µg.m^{-3}	0.30	30	0.25

Model	Country & affiliation	Context of using MQO	Test-case		
			Model resolution	Pollutants	Data assimilation
ADMS-Urban http://pandora.meng.auth.gr/mds/showshort.php?id=18	UK (CERC)	Evaluating the *air*TEXT* forecasting modelling system for Greater London (Stidworthy et al., 2017)	Greater London (60 x 50 km^2) Variable resolution (10 m roadside, 50 m for background), Hourly data	NO_2, O_3, PM10	No
CHIMERE-FR http://pandora.meng.auth.gr/mds/showlong.php?id=144	France (INERIS)	Operational air quality modelling system on national scale (PREV'AIR)	France 0.1 x 0.15° Hourly data	O_3, PM10	No
CHIMERE-PT http://pandora.meng.auth.gr/mds/showlong.php?id=144	Portugal (UA)	Air quality assessment and forecasting over Portugal (http://previsao-qar.web.ua.pt/) (Ribeiro et al., 2014)	Portugal 5 x 5 km^2 Hourly data	NO_2, O_3, PM10	No
CMAQ 3.6	Bulgaria (NIMH)	Research project on evaluation of the Bulgarian chemical weather forecasting and information system (Georgieva et al., 2017)	Bulgaria 9 x 9 km^2 Hourly data	NO_2, O_3, PM10	No
EPISODE v7.4.3 http://pandora.meng.auth.gr/mds/showlong.php?id=127	Norway (NILU)	Air Quality Plans	Urban area of Oslo/Bærum 1 x 1 km^2; Hourly data (http://info.meteo.bg/cw2.2)	NO_2, NOx, PM10, PM2.5	No
GEM-AQ	Poland (IEP)	Daily forecasts for Poland Research projects (AQ assessment and episodes study)	Poland and neighbor countries 5 x 5 km^2 Hourly data	NO_2, O_3	No
NINFA (COSMO-I7 + CHIMERE)	Italy (ARPAE, Emilia Romagna)	Air quality system for both operational forecast and regional assessment (Stortini et al., 2017)	Emilia romagna and neighbor region 5x5km^2 Hourly data	NO_2, O_3, PM10, PM2.5	No
OPS + SRM http://pandora.meng.auth.gr/mds/showlong.php?id=73	The Netherlands (RIVM)	Checking the model quality of the official Dutch Standard Calculation Methods (for Air Quality)	The Netherlands Variable resolution Annual data	NO_2, NOx, PM10 and PM2.5	No
RIO/AURORA http://pandora.meng.auth.gr/mds/showlong.php?id=167	Belgium (VITO)	Research project (assessment of best model, temporal and spatial validation) (Veldeman et al., 2016)	Belgium domain 4 x 4 km^2 Hourly data	PM10, PM2.5, NO_2, O_3	Yes (residual kriging)
RIO/ RIO-IFDM http://pandora.meng.auth.gr/mds/showshort.php?id=50	Belgium (IRCEL)	Air quality assessment to public information. Model evaluation and model intercomparison (Adriaenssens and Trimpeneers, 2015)	Belgium, Receptors interpolated up to 10x10 m Hourly data	NO_2	Yes (residual kriging)
WRF/Chem v3.6 http://ruc.noaa.gov/wrf/wrf-chem/	Cyprus (EEWRC, Cyprus Institute)	Air quality assessment and forecasting purposes for Cyprus	Cyprus domain 50, 10 and 2.5 km^2; Hourly data	O_3, NOx, PM10, PM2.5, CO, SO_2	No
WRF-Chem v3.4 http://ruc.noaa.gov/wrf/wrf-chem/	Austria (ZAMG)	Evaluating daily air quality forecasts; and also for research topics (dust events, volcanic eruptions, pollens)	2 domains: Europe (12x12km^2); Alpine region (4x4km^2) Hourly data	PM10, Dust, O_3, NO_2, NO, SO_2, Ash	Yes

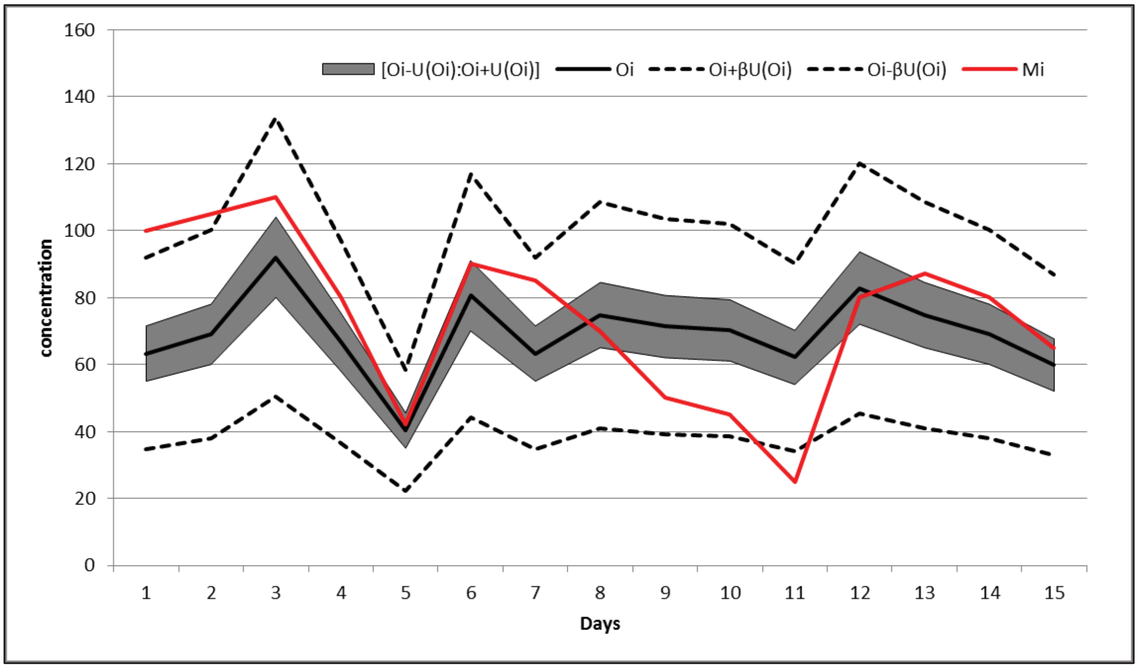
*airTEXT is a free service for the public providing air quality alerts by SMS text message, email and voicemail and 3-day forecasts of air quality, pollen, UV and temperature.

This is a post-peer-review, pre-copyedit version of an article published in Air Quality, Atmosphere and Health.
The final authenticated version is available online at: http://dx.doi.org/10.1007/s11869-018-0554-8.

Model	Model scale	Selection of stations	Statistical indicators*	Threshold values	Diagrams used
ADMS-Urban	Urban	All available monitoring stations	Fb, NMSE, R, FAC2	No values defined/used	Scatter diagram; Quantile-Quantile (QQ); Bar charts
CHIMERE-FR	Regional	Only background stations were selected	BIAS, RMSE, R, FAC2	No values defined/used	Maps of scores; Time series
CHIMERE-PT	Regional	Only background stations selected	BIAS, RMSE, R	No values defined/used	Scatter diagram; Time series;
CMAQ 3.6	Regional	Only background stations selected	NMB, RMSE, R	No values defined/used	Scatter, Box and Whisker plot, Bar plots, Time series
EPISODE v7.4.3	Urban	All available monitoring stations	BIAS, RMSE, R, FAC2	No values defined/used	Scatter diagram; QQ plot; Time series
GEM-AQ	Regional	All available monitoring stations	BIAS, NMSE, RMSE, R	No values defined/used	Scatter diagram; Time series; Taylor plot.
NINFA	Regional	Only background stations selected	BIAS, RMSE, R	No values defined/used	Scatter diagram; Boxplot; Time series
OPS (Operational Priority Substances)	Urban/regional	All available monitoring stations	Fb, NMSE, R	No values defined/used	Scatter diagram; Bland Altman plots and QQ plots.
RIO/AURORA	Regional	Traffic stations were omitted from the analysis	BIAS, ME, RMSE, R, fraction false alerts	No values defined/used	Scatter diagram; QQ plots; boxplots of statistical indicators
RIO/ RIO-IFDM	Regional	Passive sampling points (field campaign)	BIAS, RMSE, R	No values defined/used	Scatter diagram; QQ plots
WRF-Chem v3.6	Regional	1 site from measuring campaign and 3 background stations of the national network	BIAS, RMSE, R, NMB	No values defined/used	Scatter diagram; Time series and Taylor diagrams
WRF-Chem v3.4	Regional	All available national monitoring stations	Fb, NMSE	No values defined/used	Scatter diagram; Time series; Contour plot

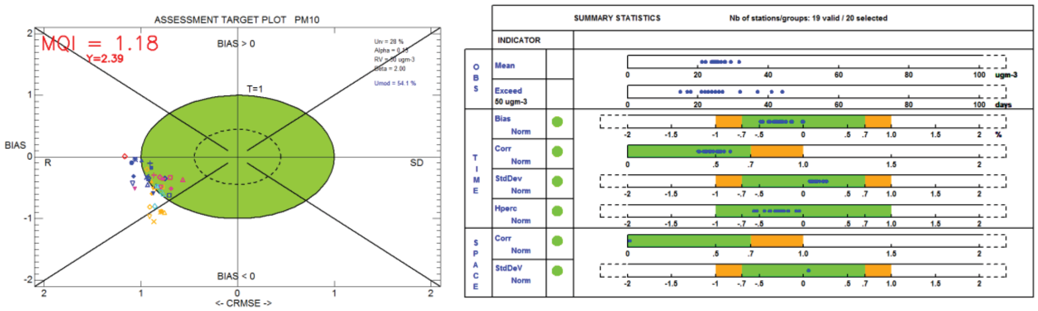
*Fb- Fractional Bias; NMSE – Normalized Mean Square Error, R- correlation factor; FAC2 - ; BIAS – systematic error; RMSE – Root Mean Square Error; NMB – Normalized Mean Bias; ME – Mean Error

Figure1



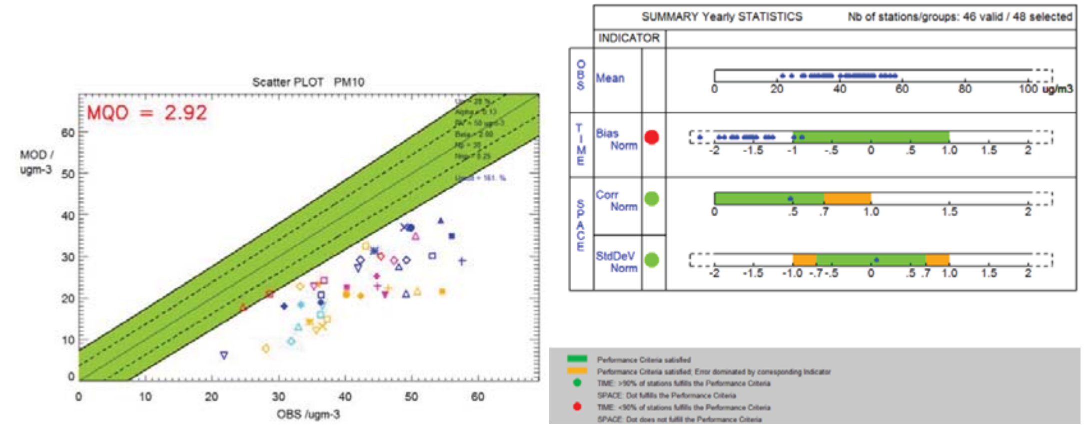
This is a post-peer-review, pre-copyedit version of an article published in Air Quality, Atmosphere and Health.
The final authenticated version is available online at: http://dx.doi.org/10.1007/s11869-018-0554-8.

Figure2



This is a post-peer-review, pre-copyedit version of an article published in Air Quality, Atmosphere and Health.
The final authenticated version is available online at: http://dx.doi.org/10.1007/s11869-018-0554-8.

Figure3



This is a post-peer-review, pre-copyedit version of an article published in Air Quality, Atmosphere and Health.
The final authenticated version is available online at: http://dx.doi.org/10.1007/s11869-018-0554-8.