European Commission

JRC TECHNICAL REPORT

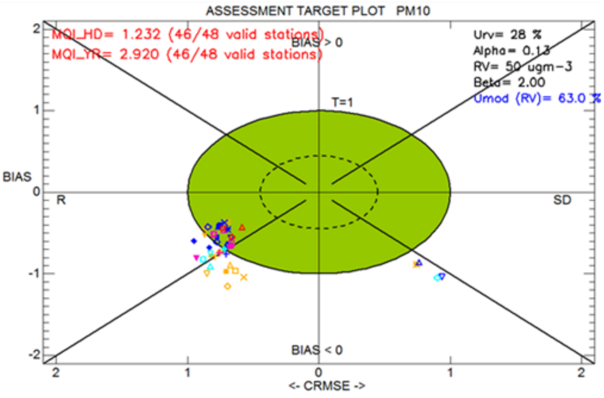
# FAIRMODE Guidance Document on Modelling Quality Objectives and Benchmarking

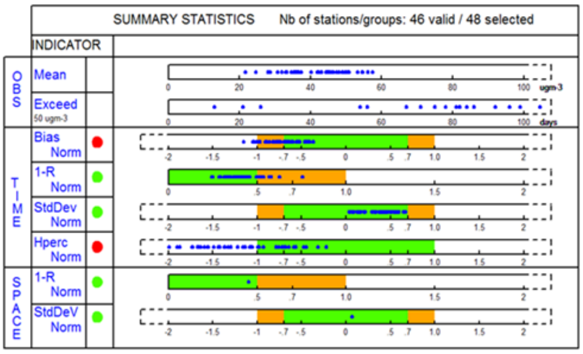
*Version 3.3*

Janssen, S., Thunis, P.

With contributions of: Adani, M., Piersanti, A., Carnevale, C., Cuvelier, C., Durka, P., Georgieva, E., Guerreiro, C., Malherbe, L., Maiheu, B., Meleux, F., Monteiro, A., Miranda, A., Olesen, H., Pfäfflin, F., Stocker, J., Sousa Santos, G., Stidworthy, A., Stortini, M., Trimpeneers, E., Viaene, P., Vitali, L., Vincent, K., Wesseling, J.

2022





Joint Research Centre

EUR 31068 EN

This publication is a Technical report by the Joint Research Centre (JRC), the European Commission's science and knowledge service. It aims to provide evidence-based scientific support to the European policymaking process. The scientific output expressed does not imply a policy position of the European Commission. Neither the European Commission nor any person acting on behalf of the Commission is responsible for the use that might be made of this publication. For information on the methodology and quality underlying the data used in this publication for which the source is neither Eurostat nor other Commission services, users should contact the referenced source. The designations employed and the presentation of material on the maps do not imply the expression of any opinion whatsoever on the part of the European Union concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries.

**Contact information**
Name: P. Thunis
Address: European Commission, Joint Research Center (JRC), via Fermi 2749, 21027 Ispra (VA), Italy
Email: philippe.thunis@ec.europa.eu
Tel.: +39 0332 785670

**EU Science Hub**
https://ec.europa.eu/jrc

JRC129254

EUR 31068 EN

| | | | |
|---|---|---|---|
| PDF | ISBN 978-92-76-52425-0 | ISSN 1831-9424 | doi:10.2760/41988 |

Luxembourg: Publications Office of the European Union, 2022

How to cite this report: Janssen, S., Thunis, P., *FAIRMODE Guidance Document on Modelling Quality Objectives and Benchmarking* (version 3.3*)*, EUR 31068 EN, Publications Office of the European Union, Luxembourg, 2022, ISBN 978-92-76-52425-0, doi:10.2760/41988, JRC129254

## Contents

Acknowledgements 1
Abstract 2
Version history 3
1 INTRODUCTION 4
1.1 Why Modelling Quality Objectives? 4
1.2 What are the purposes of this Document? 4
1.3 Who is the target audience of this Document? 4
1.4 What are the main components of this Document? 4
2 BENCHMARKING: A WORD OF CAUTION 6
3 DEFINITIONS 7
4 MAIN ASSUMPTIONS 8
5 MODELLING INDICATORS AND CRITERIA 9
5.1 Statistical indicators 9
5.2 Modelling quality indicator (MQI) and modelling quality objective (MQO) 10
5.2.1 MQI and MQO for hourly/daily/maximum daily 8-hour mean concentration data 10
5.2.2 MQI and MQO for yearly average model results 11
5.2.3 The 90% principle 11
5.3 Fulfilment of the hourly/daily vs annual MQO 12
5.4 Calculation of the associated model uncertainty 12
5.5 Comparison to values in the AQD 13
5.6 Modelling performance indicators (MPI) and criteria (MPC) 13
5.6.1 Temporal MPI and MPC 13
5.6.2 Spatial MPI and MPC 14
5.7 Data requirements for application of the MQI and MPI 15
5.7.1 Measurement data quality 15
5.7.2 Minimum number of stations for application of the MQO 16
5.7.3 Data assimilation / data fusion 16
5.7.4 Exceptional cases 16
6 REPORTING MODEL PERFORMANCE 18
6.1 Hourly data 18
6.1.1 Target Diagram 18
6.1.2 Summary Report 19
6.2 Yearly average data 20
6.2.1 Scatter Diagram 20
6.2.2 Summary Report 20
7 OPEN ISSUES 22
7.1 Performance criteria for high percentile values 22

7.2 Application of the procedure to other parameters ... 22
8 FORECASTING & EXCEEDANCES INDICATORS ... 23
8.1 The forecast Modelling Quality Indicator ... 23
8.2 The forecast Modelling Performance Indicators ... 24
8.3 Threshold exceedances indicators ... 25
8.4 Diagrams in the DELTA tool ... 26
8.4.1 Forecast Target Plot ... 26
8.4.2 Forecast MPI Plot ... 27
8.4.3 Forecast Threshold Performance Plot ... 28
8.4.4 Diagrams for exceedance indicators ... 29
8.4.5 Summary report ... 30
8.4.6 Forecast diagrams using Air Quality Indices ... 32
9 CONCLUSIONS ... 35
10 REFERENCES ... 36
10.1 Peer reviewed articles ... 36
10.2 Reports/ working documents / user manuals ... 36
List of abbreviations and definitions ... 38
List of figures ... 39
List of Tables ... 40
Annexes ... 41
ANNEX 1. MEASUREMENT UNCERTAINTY ... 41
ANNEX 2. OVERVIEW OF EXISTING LITERATURE ... 43
ANNEX 3. RELATED TOOLS ... 45

## Acknowledgements

The authors thank the Fairmode community for their contributions as well as for the fruitful discussions that took place around the particular topic of MQO and benchmarking. The authors wish to acknowledge in particular the contributions of:

Mario Adani, ENEA, Italy

Piersanti Antonio, ENEA, Italy

Claudio Carnevale, University Brescia, Italy

Cornelis Cuvelier, ex-JRC, Ispra, Italy

Pawel Durka, IOS, Poland

Emilia Georgieva, National Institute of Meteorology and Hydrology, Bulgaria

Cristina Guerreiro, NILU, Norway

Laure Malherbe, INERIS, France

Bino Maiheu, VITO, Belgium

Frederic Meleux, INERIS, France

Alexandra Monteiro, University Aveiro, Portugal

Ana Miranda, University Aveiro, Portugal

Helge Olesen, Aarhus University, Denmark

Florian Pfäfflin, IVU Umwelt GmbH, Germany

Jenny Stocker, CERC, UK

Gabriela Sousa Santos, NILU, Norway

Amy Stidworthy, CERC, UK

Michele Stortini, ARPA Emilia Romagna, Italy

Elke Trimpeneers, IRCEL, Belgium

Peter Viaene, VITO, Belgium

Keith Vincent, Ricardo AEA, UK

Lina Vitali, ENEA, Italy

Joost Wesseling, RIVM The Netherlands

**Authors**

Stijn Janssen, VITO, Belgium

Philippe Thunis, JRC, Ispra, Italy

# Abstract

The development of the procedure for air quality model benchmarking in the context of the Air Quality Directive 2008/50/EC (AQD) has been an on-going activity in the context of the FAIRMODE[1] community, chaired by the JRC. A central part of the studies was the definition of proper modelling quality indicators and criteria to be fulfilled in order to allow sufficient level of quality for a given model application under the AQD. The focus initially on applications related to air quality assessment has gradually been expanded to other applications, such as forecasting and planning. The main purpose of this Guidance Document is to explain and summarise the current concepts of the modelling quality objective methodology, elaborated in various papers and documents in the FAIRMODE community, addressing model applications for air quality assessment and forecast. Other goals of the Document are linked to presentation and explanation of templates for harmonised reporting of modelling results. Giving an overview of still open issues in the implementation of the presented methodology, the document aims at triggering further research and discussions.

A core set of statistical indicators is defined using pairs of measurement-modelled data. The core set is the basis for the definition of a modelling quality indicator (MQI) and additional modelling performance indicators (MPI), which take into account the measurement uncertainty. The MQI describes the discrepancy between measurements and modelling results (linked to RMSE), normalised by measurement uncertainty and a scaling factor. The modelling quality objective (MQO) requires MQI to be less than or equal to 1. With an arbitrary selection of the scaling factor of 2, the fulfilment of the MQO means that the allowed deviation between modelled and measured concentrations is twice the measurement uncertainty. Expressions for the MQI calculation based on time series and yearly data are introduced. MPI refer to aspects of correlation, bias and standard deviation, applied to both the spatial and temporal dimensions. Similarly to the MQO for the MQI, modelling performance criteria (MPC) are defined for the MPI; they are necessary, but not sufficient criteria to determine whether the MQO is fulfilled. The MQO is required to be fulfilled at 90% of the stations, a criterion which is implicitly taken into account in the derivation of the MQI. The associated modelling uncertainty is formulated, showing that in case of MQO fulfilment the modelling uncertainty must not exceed 1.75 times the measurement one (with the scaling factor fixed to 2).

A reporting template is presented and explained for hourly and yearly average data. In both cases there is a diagram and a table with summary statistics. In a separate section open issues are discussed and an overview of related publications and tools is provided. Finally, a chapter on modelling quality objectives for forecast models is introduced. In Annex 1, we discuss the measurement uncertainty which is expressed in terms of concentration and its associated uncertainty. The methodology for estimating the measurement uncertainty is overviewed and the parameters for its calculation for PM, NO2 and O3 are provided. An expression for the associated modelling uncertainty is also given.

This aim of this document is to support modelling groups, local, regional and national authorities in their modelling application, in the context of air quality policy.

[1] The Forum for Air quality Modelling (FAIRMODE) is an initiative 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 Directives.

## Version history

| Version | Release date | Modifications | DELTA version |
|---|---|---|---|
| **1.0** | 6/02/2015 | First version | |
| **1.1** | 1/04/2015 | NILU application added to Examples of good practice.<br>Small textual corrections. | |
| **2.0** | 26/05/2016 | Update of Section on definition of Modelling Quality Objective<br>Update of Section on performance reporting<br>Open Issue list updated according to ongoing discussions within WG1<br>Update of CERC, NILU and IRCEL contributions in the Best Practice section | |
| **2.1** | 13/02/2017 | Improvement of the overall readability of the text.<br>Introduction of a chapter on “Definitions” & “Related Tools”<br>Update of Section on Modelling uncertainty<br>Introduction of a chapter on MQO for forecast models<br>Removal of Section “Examples of Best Practice”. The user feedback will be reused in a publication (Monteiro et al, 2018) | |
| **2.2** | 20/07/2018 | Modified section of MQO for forecast, including the introduction of a probabilistic threshold exceedance indicator | 5.6 |
| **3.1** | 30/10/2019 | Restructuring of the text with the description of the measurement uncertainty being moved to the Annex.<br>Change of notation (simplification) for some parameters used in the derivation of the measurement uncertainty.<br>Update of the forecast section based on the outcome of the September 2019 Hackathon.<br>Review of the open issues with some of them being moved to the main text. | 5.6 |
| **3.2** | 01/04/2020 | Update of the forecast section based on the outcome of the Plenary Meeting in Berlin and preceding discussions. | 6.0 |
| **3.3** | 15/02/2022 | Update of the forecast section | 7.0 |

# 1 INTRODUCTION

## 1.1 Why Modelling Quality Objectives?

In general, the quality of models is understood in terms of their 'fitness for purpose'. The modelling experience indicates that there are no 'good' or 'bad' models. Evidence is rather based on the question of whether a model is suitable for the intended application and specified objectives. As such, the quality of a model is always relative and is measured against the quality objectives for any particular model application. Statistical performance indicators, which provide insight on model performance, are generally used to assess the performance of a model against measurements for a given application. They do not however tell whether model results have reached a sufficient level of quality for a given application. This is the reason for which modelling quality objectives (MQO), defined as the minimum level of quality to be achieved by a model for policy use, need to be set.

Modelling quality objectives are described in Annex I of the Air Quality Directive 2008/50/EC (AQD) along with the monitoring quality objectives. They are expressed as a relative uncertainty (%) which is then further defined in the AQ Directive. However, as mentioned in the FAIRMODE technical guidance document2 the wording of the AQD text needs further clarification in order to become operational. It is important to note that these modelling quality objectives apply only to assessment of the current air quality when reporting exceedances, and do not refer to other model applications, such as planning or forecasting. However, there is clearly an expectation when using models for these other applications that they have been verified and validated in an appropriate, albeit unspecified, way.

## 1.2 What are the purposes of this Document?

The main objectives of this Guidance Document are to:

- Explain the MQO concepts and methodology developed within FAIRMODE;
- Provide recommendations and guidance for assessing model performance related to a given air quality model application in the frame of the AQD, based on the experience and elaborations in the FAIRMODE community. In a first step PM10, PM2.5, NO2 and O3 are prioritised but ultimately the methodology should also cover other pollutants such as heavy metals and polycyclic aromatic hydrocarbons. The focus of this document is mainly on the use of air quality models for the assessment of air quality, however hints for forecast applications are also provided;
- Promote consistency in model evaluation for policy applications related to the AQD;
- Promote harmonised reporting of modelling performance in the EU Member States;
- Promote further discussions around remaining open issues.

## 1.3 Who is the target audience of this Document?

This Guidance Document is intended primarily for environmental experts using air quality models in the context of the EU AQD. Some of these experts apply tools (software), developed around the concepts in this document (as DELTA, ATMOSYS or MyAir, see Annexes) and thus, the current text provides additional support to the respective User Guides. Developers of the mentioned tools might also benefit from the notes in this document.

A wider target audience consists of air quality modellers, who are interested in methods and criteria for evaluating model performance and follow recent developments in various modelling communities.

## 1.4 What are the main components of this Document?

This Document is built upon the following major components:

- The concept and methodology for modelling quality objectives and modelling performance criteria developed within the FAIRMODE community (Chapter 5)
- The techniques for reporting of model performance in harmonised way (Chapter 6)

[2] https://www.eea.europa.eu/publications/fairmode

- Opens issues to the above components, which merit consideration and further development within the FAIRMODE community (Chapter 7)
- A benchmarking methodology for forecast models (Chapter 8)
- Summary of additional resources (publications, tools), (Annexes)

This Guidance Document is periodically revised to ensure that new FAIRMODE developments or expanded regulatory requirements are incorporated, as well as to account for User's feedback.

# 2 BENCHMARKING: A WORD OF CAUTION

Based on the UNESCO[3] definition, adapted to the context of air quality modelling, benchmarking can be defined as follows:

- a standardised 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 inter-comparison 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 model quality and promote good practices.

When we talk about benchmarking, it is normally implicitly assumed that the best model is one, which produces results the closest to measured values. In many cases, this is a reasonable assumption. However, it is important to recognise that this is not always the case, so one should proceed with caution when interpreting benchmarking results. Here are three examples in which blind faith in benchmarking statistics would be misplaced:

- Emission inventories are seldom perfect. If not all emission sources are included in the inventory used by the model then a perfect model should not match the observations, but have a bias. In that case, seemingly good results would be the result of compensating errors;
- If the geographical pattern of concentrations is very patchy – such as in urban hot spots – monitoring stations are only representative of a very limited area. It can be a major challenge – and possibly an unreasonable challenge – for a model to be asked to reproduce such monitoring results;
- Measurement data are not error free and a model should not always be in close agreement with monitored values.

In general, in the EU member states there are different situations, which pose different challenges to modelling including among others the availability of input data, emission patterns and the complexity of atmospheric flows due to topography.

The implication of all the above remarks is that if one wishes to avoid drawing unwarranted conclusions from benchmarking results, then it is not sufficient to inspect benchmarking results. Background information should be acquired on the underlying data to consider the challenges they represent.

Good benchmarking results are therefore not a guarantee that everything is perfect. Poor benchmarking results should be followed by a closer analysis of their causes. This should include examination of the underlying data and some exploratory data analysis.

Benchmarking in the context of FAIRMODE strategy is intended as the compilation of different approaches and the subsequent development and testing of a standardised evaluation/inter-comparison methodology for collecting and reporting model inputs/outputs in a way that enables relevant comparisons. The aim is to identify good practices and propose ways to diagnose problems in performance.

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

# 3 DEFINITIONS

**Modelling Quality Indicator (MQI)**

Statistical indicator calculated on the basis of measurements and modelling results. It is used to determine whether the Modelling Quality Objectives are fulfilled. It describes the discrepancy between measurements and modelling results, normalised by the measurement uncertainty and a scaling factor. The MQI might be regarded as an MPI. However, it has a special status and is assigned its own name because it determines whether the MQO is fulfilled.

**Modelling Quality Objective (MQO)**

Criterion for the value of the Modelling Quality Indicator (MQI). The MQO is said to be fulfilled if MQI is less than or equal to unity.

**Modelling Performance Indicator (MPI)**

Statistical indicator calculated on the basis of measurements and modelling results. In the context of the present Guidance document, several Modelling Performance Indicators are defined. Each of them describes a certain aspect of the discrepancy between measurement and modelling results. There are Modelling Performance Indicators referring to the three aspects of correlation, bias and normalised mean square deviation. Model Performance Indicators are also developed to assess the capability of the model to reproduce spatial variation. See section 5.6 for definitions of the MPI's. Specific MPI's, based on mean fractional error, are also developed for the assessment of models in forecast mode. See section 8.2 for details.

**Modelling Performance Criterion**

Criterion that a Model Performance Indicator is expected to fulfil.

**Measurement uncertainty**

Uncertainty related to the measurement of ambient concentrations. FAIRMODE relies on the expertise of the AQUILA network to define those quantities.

**Combined uncertainty**

Uncertainty taking into account the individual uncertainties associated with the input quantities in a measurement/model

**Expanded uncertainty**

Product of a combined uncertainty and a factor larger than the number one; the factor depends upon the type of probability distribution of the output quantity in a measurement/model and on the selected coverage probability.

**Model evaluation**

Sum of processes that need to be followed in order to determine and quantify a model's performance capabilities, weaknesses and advantages in relation to the range of applications for which it has been designed.
Note: The present Guidance document does not prescribe a procedure for model evaluation.
[SOURCE: EEA Technical Reference Guide No. 10, 2011]

**Modelling validation**

Comparison of modelled predictions with observations, using a range of modelling quality indicators.
[SOURCE: EEA Technical Reference Guide No. 10, 2011]

# 4 MAIN ASSUMPTIONS

The focus of this Guidance Document is on presenting the modelling quality objective (MQO) and associated modelling performance criteria (MPC) for different statistical indicators related to a given air quality model application for air quality assessment in the frame of the AQD. These statistical indicators are produced by comparing air quality model results and measurements at monitoring sites. This has the following consequences:

### 1. Species and time frame considered

The modelling quality objective (MQO) and modelling performance criteria (MPC) are in this document defined only for pollutants and temporal scales that are relevant to the AQD. Currently only O3, NO2, PM10 and PM2.5 data covering an entire calendar year are considered.

### 2. Fulfilment criteria

According to the Data Quality Objectives in Annex I of the AQD the uncertainty for modelling is defined as the maximum deviation of the measured and calculated concentration levels for 90 % of individual monitoring points over the period considered near the limit value (or target value in the case of ozone) and this without taking into account the timing of the events. While the MQO and MPC proposed in this document do consider the timing of the events, we also need to select a minimum value for the number of stations in which the model performance criterion has to be fulfilled and propose to also set this number to 90 %. This means that the model performance criteria must be fulfilled for at least 90% of the available stations. This is further detailed in Section 5.2.3.

### 3. Measurement uncertainty[4]

A novelty in the concept for defining MQO and MPC is the introduction of the measurement uncertainty in the respective statistical parameters. The measurement uncertainty is expressed as dependent on the concentration. Methods for estimating parameters for the key species treated are further explained in Annex 1.

**Main points:**

- Pollutants covered: O3, NO2, PM10 and PM2.5
- Paired data series of model results and observations at fixed locations
- Fulfilment criteria fixed at 90% of available individual locations
- Measurement uncertainty included in the modelling quality indicator (MQI) and in the modelling performance indicators (MPI).

[4] In previous versions of the Guidance Document, this term was often interchanged with “observation uncertainty”. Further on, we will use only the term “measurement uncertainty”.

# 5 MODELLING INDICATORS AND CRITERIA

## 5.1 Statistical indicators

Models applied for regulatory air quality assessment are commonly evaluated based on comparisons against measurements. This element of the model evaluation process is also known as operational model evaluation or statistical performance analysis, since statistical indicators and graphical analysis are used to determine the capability of an air quality model to reproduce measured concentrations. It is generally recommended to apply multiple statistical indicators regardless of the model application since each one has its advantages and disadvantages.

To cover all aspects of the model performance in terms of amplitude, phase and bias the following core set of statistical indicators has been proposed within FAIRMODE for the statistical analysis of model performance with Mi and Oi respectively the modelled and observed values where i is a number (rank) between 1 and N and N the total number of modelled or observed values:

**Table 1:** Core set of statistical indicators

| Indicator | Formula | |
|---|---|---|
| Root Mean Square Error ($RMSE$) | $RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}$ | (1) |
| Correlation coefficient (R) | $R = \frac{\sum_{i=1}^{N}(M_i - \bar{M})(O_i - \bar{O})}{\sqrt{\sum_{i=1}^{N}(M_i - \bar{M})^2}\sqrt{\sum_{i=1}^{N}(O_i - \bar{O})^2}}$<br>with $\bar{O} = \frac{\sum_{i=1}^{N} O_i}{N}$ the average observed value<br>and $\bar{M} = \frac{\sum_{i=1}^{N} M_i}{N}$ the average modelled value. | (2) |
| Normalised Mean Bias (NMB) | $NMB = \frac{BIAS}{\bar{O}}$<br>where $BIAS = \bar{M} - \bar{O}$ | (3) |
| Normalised Mean Standard Deviation (NMSD) | $NMSD = \frac{(\sigma_M - \sigma_O)}{\sigma_O}$<br>with $\sigma_O = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - \bar{O})^2}$ the standard deviation of the observed values and $\sigma_M = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(M_i - \bar{M})^2}$ the standard deviation of the modelled values. | (4) |

Source: JRC

Although statistical performance indicators provide insight on model performance, they do not tell whether model results have reached a sufficient level of quality for a given application, e.g. for policy support. In the literature, different recommended values can be found for some of the statistical indicators for assessment of modelling performance.

In the FAIRMODE community, a main statistical indicator has been introduced based on measurement and modelling results - the modelling quality indicator (MQI). Then a criteria for this MQI is defined, the modelling quality objective (MQO), representing the minimum level of quality to be achieved by a model for policy use. A specific feature of the MQI is its link to the measurement uncertainty. In the following, the formulation for MQI and MQO is first presented, while a simple expression for measurement uncertainty is discussed in Annex 1.

**Core set of statistical indicators**

- Includes: Root Mean Square Error, Correlation, Normalised Mean Bias, Normalised Mean Standard Deviation
- Serves as a basis for the definition of the main model quality indicator MQI (linked to RMSE) and additional Model Performance Indicators (MPI), linked to the remaining core statistical indicators.

## 5.2 Modelling quality indicator (MQI) and modelling quality objective (MQO)

### 5.2.1 MQI and MQO for hourly/daily/maximum daily 8-hour mean concentration data

The Modelling Quality Indicator (MQI) is a statistical indicator calculated based on measurements and modelling results. It is defined as the ratio between the model-measured bias at a fixed time (i) and a quantity proportional to the measurement uncertainty as:

$$\mathrm{MQI(i)} = \frac{|\mathrm{O_i} - \mathrm{M_i}|}{\beta U(\mathrm{O_i})} \qquad (5)$$

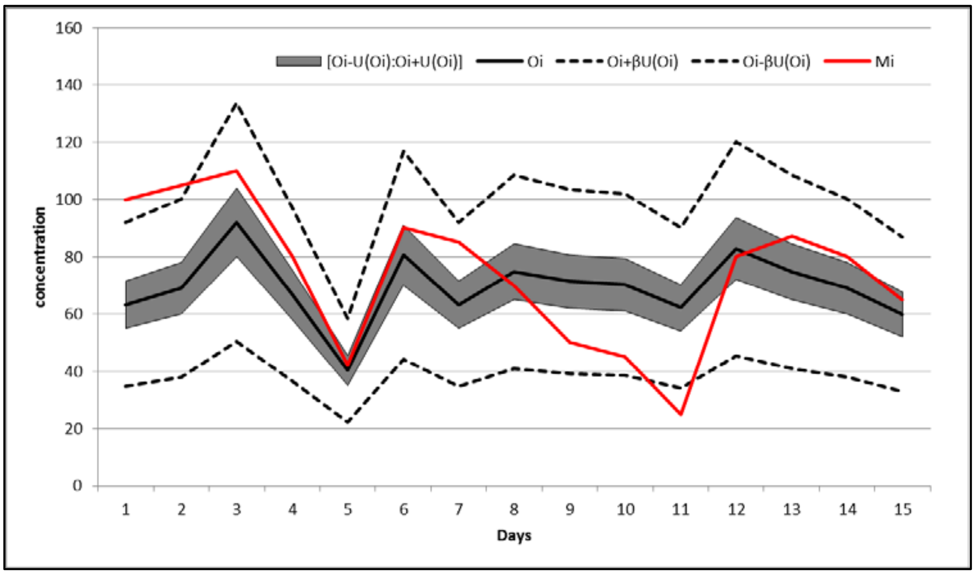
The formulation of the expanded 95th percentile measurement uncertainty U(Oi) is derived in the Annex 1. β is the coefficient of proportionality.

The MQO is the criteria for the MQI. The MQO is fulfilled when the MQI is less or equal to 1, i.e.:

$$\mathrm{MQO\ is\ fulfilled\ when\ MQI} \leq 1 \qquad (6)$$

In Figure 1, the MQO is fulfilled for example on days 3 to 10, whereas it is not fulfilled on days 1, 2 and 11. We will also use the condition $|\mathrm{O_i} - \mathrm{M_i}| \leq \mathrm{U(O_i)}$ in the MQO related diagrams (see Section 6) to indicate when model-measurement differences are within the measurement uncertainty (e.g. days 5 and 12 in Figure 1).

**Figure 1**: MQI and MQO explained on PM10 time series: measured (bold black) and modelled (bold red) concentrations are represented for a single station. The grey shaded area is indicative of the measurement uncertainty whereas the dashed black lines represent the MQO limits (proportional to the measurement uncertainty with β=2). Modelled data fulfilling the MQO must be within the dashed lines.



Source: JRC

Equation (5) and (6) can then be used to generalise the MQI and MQO to a time series:

$$MQI = \frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}}{\beta\sqrt{\frac{1}{N}\sum_{i=1}^{N}U(\mathrm{O}_i)^2}} = \frac{\mathrm{RMSE}}{\beta RMS_U} \quad \text{and the MQO is fulfilled when MQI} \leq 1 \qquad (7)$$

With this MQO formulation, the RMSE between measured Oi and modelled Mi values (numerator) is compared to a value representative of the maximum allowed uncertainty (denominator). The value of β determines the stringency of the MQO. From now on in this Guidance Document (and in the DELTA tool) β is arbitrarily set equal to 2, allowing thus deviation between modelled and measured concentrations as twice the measurement uncertainty.

### 5.2.2 MQI and MQO for yearly average model results

For air quality models that provide yearly averaged pollutant concentrations, the MQI is defined as the mean bias between modelled and measured annual averaged concentrations normalised by the expanded measurement uncertainty of the mean concentration:

$$MQI = \frac{|\bar{O} - \bar{M}|}{\beta U(\bar{\mathrm{O}})} \qquad and\ MQO\ is\ fulfilled\ when\ MQI \leq 1 \qquad (8)$$

The MQO is fulfilled when the MQI is less than or equal to 1, as in the case of time series data.

### 5.2.3 The 90% principle

For all statistical indicators used in DELTA for benchmarking purposes the approach currently used in the AQD has been followed. This means that the MQO must be fulfilled for at least 90% of the available stations (discussed in Section 5.7). The practical implementation of this approach consists in calculating the MQI associated to each station, rank them in ascending order and inferring the 90th percentile value according to the following linear interpolation (for N station):

$$\mathrm{MQI}_{90\mathrm{th}} = \mathrm{MQI}(\mathrm{S}_{90}) + [\mathrm{MQI}(\mathrm{S}_{90}+1) - \mathrm{MQI}(\mathrm{S}_{90})] * \mathrm{dist} \qquad (9)$$

where S90 = integer(N*0.9) and dist= $[\mathrm{N} * 0.9 - \mathrm{integer}(\mathrm{N} * 0.9)]$. If only one station is used in the benchmarking, $\mathrm{MQI}_{90\mathrm{th}} = \mathrm{MQI}(\mathrm{station}) * 0.9$.

The MQO is then expressed as:

$$\text{MQO is fulfilled when } \mathrm{MQI}_{90\mathrm{th}} \leq 1 \tag{10}$$

## 5.3 Fulfilment of the hourly/daily vs annual MQO

Modelling applications delivering hourly and/or daily data can be assessed both in terms of the yearly and hourly/daily MQO. Results might pass the hourly/daily objective but fail to meet the criteria when annual averaged values of the time series are used in the annual MQI procedure, and vice-versa.

The difference between the two indicators is related to the auto-correlation in both the monitoring data and the model results and the way those auto-correlations are affecting the uncertainty of the annual averaged values. At this stage, it is suggested that model applications with hourly/daily output should also comply with the annual MQO.

## 5.4 Calculation of the associated model uncertainty

To give some information about the model uncertainty, we will use the normalised deviation indicator En (ref: ISO 13528). It scales the model (M)-measurement (O) difference with the measurement and model uncertainties [$U(\mathrm{O_i})$ and $\mathrm{U(M_i)}$] associated to this difference:

$$\mathrm{E_n} = \frac{|\mathrm{O_i} - \mathrm{M_i}|}{\sqrt{U(\mathrm{O_i})^2 + \mathrm{U(M_i)}^2}} \tag{11}$$

En equals to unity implies that the model and measurement uncertainties are compatible with the model-measurement bias. We use this relation, i.e. En=1, in DELTA to estimate the minimum model uncertainty compatible with the resulting model-measurement bias as follows:

$$\mathrm{E_n} = 1 \Rightarrow \mathrm{U(M_i)} = U(\mathrm{O_i})\sqrt{\left(\frac{\mathrm{O_i} - \mathrm{M_i}}{\mathrm{U(O_i)}}\right)^2 - 1} \tag{12}$$

Relation (12) does not apply to cases in which $|O_i - M_i| < U(O_i)$, i.e. when the bias is inferior to the measurement uncertainty, cases in which no meaningful improvement of the model can be made. It is interesting to note that the fulfilment of the MQO proposed in (6) and (8) implies therefore that the model uncertainty must not exceed 1.75 times the measurement one [this value if obtained by substituting the bias term in (12) by its maximum allowed value in the MQO, i.e. βU(Oi) with β=2].

We can generalise equation (12) for a time series and for time averaged values as:

$$RMS_{U_M} = RMS_U\sqrt{\left(\frac{\mathrm{RMSE}}{RMS_U}\right)^2 - 1} \tag{13}$$

and

$$U(\bar{M}) = U(\bar{O})\sqrt{\left(\frac{\mathrm{Bias}}{U(\bar{O})}\right)^2 - 1} \tag{14}$$

In DELTA the value of the ratio ($RMS_{U_M}/RMS_U$) or ($U(\bar{M})/U(\bar{O})$) is used to scale the measurement uncertainty around the reference value U(RV) and to provide information about the minimum model uncertainty reached around the reference value.

The 90% principle (see Section 5.2.3) is also applied to the corresponding model uncertainty. The minimum model uncertainty is the value of the uncertainty associated to the 90th percentile station. This information is provided in some benchmarking diagrams (see Section 6).

## 5.5 Comparison to values in the AQD

With the uncertainty parameters as defined in Annex 1, the Table below lists the values currently used in FAIRMODE as compared to those available in the AQD. The data quality objective (DQO) and the maximum bias at limit value (defined as model quality objective in the AQD) can be compared with the reference measurement uncertainty around the limit value LV, $U_{O,r}(LV)$ and the maximum bias used in FAIRMODE. Obviously, the FAIRMODE maximum bias is concentration dependent and applies to the whole range of concentration (equal to $2U_O$) but is only reported here around the limit value. The last column shows the modelling uncertainty. Note that the values are obtained with β fixed to 2.

**Table 2**: Comparison to AQD values.

| | | | **2008 AQ Directive** | | **FAIRMODE** | | |
|---|---|---|---|---|---|---|---|
| | *Frequency* | *Limit value*<br>ug/m$^3$ | *DQO*<br>*at LV* | *Max bias*<br>*at LV* | $U_{O,r}(LV)$ | *Max bias at LV* | $U_{M,r}(LV)$ |
| **NO2** | Hour | 200 | 15% | 50% | 24.0% | 48% | 42% |
| | Year | 40 | - | 30% | 14.5% | 29% | 25% |
| **O3** | 8h | 120 | 15% | 50% | 18% | 36% | 18% |
| **PM10** | day | 50 | 25% | - | 28% | 56% | 49% |
| | year | 40 | - | 50% | 6.4% | 13% | 11% |
| **PM25** | Day | 25 | 25% | - | 36% | 72% | 63% |
| | year | | - | 50% | 10% | 20% | 17% |

Source: JRC

## 5.6 Modelling performance indicators (MPI) and criteria (MPC)

Modelling performance indicators are statistical indicators that describe certain aspect of the discrepancy between measurement and modelling results. The MQI can be treated as a kind of MPI related to one of the core statistical parameters defined in 5.1, namely the RMSE. We define here MPI related to correlation, bias and standard deviation (i.e. the remaining core statistical parameters). Furthermore, we define also MPI related to the spatial variability. The criteria that MPI are expected to fulfil are defined as modelling performance criteria (MPC). MPI do not apply to yearly average concentrations.

### 5.6.1 Temporal MPI and MPC

A characteristic of the proposed MQI, through its link to RMSE, is that errors in BIAS, σM and R are condensed into a single number. These three different statistics are however related as follows:

$$\mathrm{MQI}^2 = \frac{\mathrm{RMSE}^2}{(\beta \mathrm{RMS_U})^2} = \frac{\mathrm{BIAS}^2}{(\beta \mathrm{RMS_U})^2} + \frac{(\sigma_\mathrm{M} - \sigma_\mathrm{O})^2}{(\beta \mathrm{RMS_U})^2} + \frac{2\sigma_\mathrm{O}\sigma_\mathrm{M}(1-\mathrm{R})}{(\beta \mathrm{RMS_U})^2} \qquad (15)$$

By considering ideal cases where two out of three indicators perform perfectly, separate MPI and respective MPC can be derived from (15) for each of the three statistics. For example, assuming R=1 and σM= σO in equation (15) leads to an expression for the bias model performance indicator (MPI) and bias model performance criterion (MPC) as:

$$\mathrm{MPI} = \frac{\mathrm{BIAS}}{(\beta \mathrm{RMS_U})} \qquad \text{and} \qquad \mathrm{MPC}\text{:}\ \frac{\mathrm{BIAS}}{\beta \mathrm{RMS_U}} \leq 1$$

This approach can be generalised to the other two temporal MPI (see Table 3).

**Table 3:** Model performance indicators and criteria for temporal statistics.

| | **MPI - Model Performance Indicator**<br>**(the two parameters assumed to be ideal)** | **Equation** | **MPC - Modelling performance criterion** |
|---|---|---|---|
| BIAS | $MPI = \frac{\lvert BIAS \rvert}{\beta\ RMS_U}$<br>$(R = 1, \sigma_O = \sigma_M)$ | (16) | The MPC is fulfilled when<br>$\mathrm{MPI} \leq 1$ |
| R | $\mathrm{MPI} = \frac{1-\mathrm{R}}{0.5\beta^2 \frac{RMS_U{}^2}{\sigma_O \sigma_M}}$<br>$(BIAS = 0, \sigma_O = \sigma_M)$ | (17) | |
| Standard deviation | $MPI = \frac{\lvert \sigma_M - \sigma_O \rvert}{\beta RMS_U}$<br>$(BIAS = 0, R = 1)$ | (18) | |

Source: JRC

One of the main advantages of this approach for deriving separate MPI is that it provides a selection of statistical indicators with a consistent set of performance criteria based on one single input: the measurement uncertainty U(Oi). While the $MQI$, based on the RMSE indicator, provides a general overview of the model performance, the associated MPI for correlation, standard deviation and bias can be used to highlight which of the model performance aspects need to be improved. It is important to note that **the MPC for bias, correlation, and standard deviation represent necessary but not sufficient conditions to ensure fulfilment of the $MQO$.**

### 5.6.2 Spatial MPI and MPC

Spatial statistics are calculated in the benchmarking performance report (see Chapter 6). For hourly frequency, the model results are first averaged yearly at each station. A correlation and a standard deviation indicator are then calculated for this set of averaged values. Formulas (17) and (18) are still used but RMSU is substituted by $\mathrm{RMS_{\bar{U}}}$ where $\mathrm{RMS_{\bar{U}}} = \sqrt{\frac{1}{\mathrm{N}} \sum \mathrm{U}(\bar{\mathrm{O}})^2}$. The same approach holds for yearly frequency output.

**Table 4:** Model performance indicators and criteria for spatial statistics

| | MPI | Equation | MPC |
|---|---|---|---|
| R | $\mathrm{MPI} = \frac{1 - \mathrm{R}}{0.5\beta^2 \frac{{RMS_{\bar{U}}}^2}{\sigma_O \sigma_M}}$<br>$(BIAS = 0, \sigma_O = \sigma_M)$ | (19) | $\mathrm{MPC: MPI} \leq 1$ |
| Standard deviation | $MPI = \frac{\lvert\sigma_M - \sigma_O\rvert}{\beta RMS_U}$<br>$(BIAS = 0, R = 1)$ | (20) | |

Source: JRC

## 5.7 Data requirements for application of the MQI and MPI

FAIRMODE recommends that a fit-for-purpose modelling system application (assessment, planning, forecast, source apportionment) should be able to capture both the spatial and temporal variability of the environmental indicator under investigation, in the region (zone or agglomeration) covered by the application.

Therefore, in general, the resolution of the modelling system results should be such that measurements of environmental indicators (e.g. time averaged pollutant concentration levels) within the scope of the application can be reproduced, irrespective of the spatial representativeness or classification of the monitoring locations.

To reduce the ambiguity of the spatial scale in the fitness-for-purpose definition, FAIRMODE proposes, as a general guidance, that the spatial scale(s) of the modelling system should be such that all measurements of pollutant concentration levels within the scope of the application can be reproduced:

1. A modelling system application that produces data used as supplementary information to measurements should be consistent with the classification of station and pollutant that is supplemented. This means that, for rural stations, (coarser) regional scale model results can be sufficient, whereas for traffic stations more detailed street level models have to be applied. The MQO should be evaluated by making use of all measurements from monitoring stations that are complemented by the modelling application.
2. A modelling system application that is used to assess the spatial extent of the exceedances (area of exceedance, length of road in exceedance, population in the exceedance area), should be evaluated using the MQO for all available measurements in the zones and agglomeration of interest.
3. When assessment using modelling is performed as a starting point for planning under the Air Quality Directives, the ambition should be to reproduce what is measured in the atmosphere within the air quality zone under investigation. Therefore, it is recommended that all measurements in the air quality zone are used in the MQO evaluation of the assessment results. This will ensure that the starting point of the plan reproduces (most of) the complex structures and gradients as described by measurements.

### 5.7.1 Measurement data quality

A minimum of data availability is required for statistics to be produced at a given station. Presently the requested percentage of available data over the selected period is 75%. Statistics for a single station are only produced when data availability of paired modelled and observed data is for at least 75% of the time period considered. When time averaging operations are performed the same availability criteria of 75% applies. For example, daily averages will be performed only if data for 18 hours, at least, are available. Similarly, an 8-hour average value for calculating the O3 daily maximum 8-hour means is only calculated for the 8-hour periods in which 6 hourly values, at least, are available.

### 5.7.2 Minimum number of stations for application of the MQO

The minimum number of stations (Nmin) for the modelling validation is 5. The 90%-rule applies regardless of the number of stations (see Section 5.2.3). If Nmin is less than 5, the following approaches are recommended to reach 5:

- Expansion of the modelling domain with contiguous zones and/or agglomerations, i.e. enlarge the modelling domain size to include at least the minimum number of stations.
- Combination of non-contiguous zones and agglomerations that are expected to have similar characteristics. Note that the same modelling system should be applied to all zones/agglomerations.
- Expansion of the monitoring network by (a) increasing reference or equivalent stations or (2) allowing for indicative measurements (e.g. measurement campaigns) if comparable in terms of measurement uncertainty. In both cases, the network design should be prioritised based on understanding of significant sources.

### 5.7.3 Data assimilation / data fusion

The AQD suggests the integrated use of modelling techniques and measurements if the pollutant concentrations are below the upper assessment threshold to provide suitable information about the spatial and temporal distribution of pollutant concentrations. This combination is designated as data fusion or data assimilation.

Incorporating measurement data in the modelling approach poses a challenge in terms of assuring an evaluation of the modelling system by measurements.

All available independent measurement data (measurements not used in the model and complying the data quality objectives) should be included in the validation.

Different approaches found in literature are based on dividing the set of measurement data into two groups, one for the data assimilation or data fusion and one for the evaluation of the integrated modelling results. Cross-validation approaches consist in repeating this process in an iterative manner.

We recommend the “leaving one out” cross-validation strategy as a methodology for the evaluation of data assimilation or data fusion results. For complex data assimilation methodologies such as 4D VAR, it is recommended to retain an independent set of measurement sites in the modelling domain for evaluation and do the assimilation without considering these measurement sites. The number of measurement sites which are retained should be at least 20% of the total number and should be randomly selected while ensuring that all the considered scales are represented (stratified random selection). However, the modeller should be aware of the fact that this a priori selection of validation stations will have an impact on the final result of the evaluation of the model application. If the 20% - criteria is not suitable because of the lack of measurement sites in the modelling domain, independent external data should be used for the validation procedure, e.g. from dedicated measurement campaigns.

### 5.7.4 Exceptional cases

In exceptional situations, stations can be left out from the evaluation of the modelling system application when the situation is arguably too complex to be captured in a model. However, in reporting the results of the assessment, the regions/situations where the modelling system application may not be ‘fit-for-purpose’ should be clearly indicated and described.

**Modelling Quality Indicator (MQI) and Modelling Quality Objective (MQO):**

- MQI is the main modelling performance indicator
- MQI is defined as the ratio of the RMSE between measured and modelled values and a value proportional to the measurement uncertainty RMSU
- The proportionality coefficient β is arbitrarily set equal to 2, allowing thus deviation between modelled and measured concentrations as large as twice the measurement uncertainty.
- MQI for time series is given by (7), for yearly averaged data by (8). Values for MQI based on time series or annual data may differ
- MQO is the criteria for MQI. MQO is fulfilled when MQI is less than or equal to 1
- MQO does not depend on pollutant , scale and data frequency
- MQO must be fulfilled for at least 90% of the available stations

**Modelling Performance Indicators (MPI) and Modelling Performance Criteria (MPC):**

- MPI are performance indicators additional to the main MQI, highlighting which aspect of the modelling result needs to be improved
- MPI are related to temporal correlation, bias, and standard deviation, spatial correlation and bias. They all depend on measurement uncertainty
- MPC are the criteria for MPI, defined in Table 3
- MPC represent necessary but not sufficient conditions to ensure fulfilment of the MQO

**Data requirements on measurements:**

- The spatial scale(s) and resolution(s) of the modelling system should be such that all measurements of pollutant concentration levels within the scope of the application can be reproduced
- The minimum number of stations for the modelling validation is 5. If the number of available stations is less than 5, some approaches are recommended
- The “leaving one out” cross-validation strategy is recommended as a methodology for the evaluation of data assimilation or data fusion results

# 6 REPORTING MODEL PERFORMANCE

Benchmarking reports are currently defined for the hourly NO2, the 8h daily maximum O3 and daily PM10 and PM2.5 concentrations. The reports for the evaluation of hourly and yearly average model results are different. Below we present details for these two types of reports.

## 6.1 Hourly data

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

### 6.1.1 Target Diagram

The MQI as described by Eq (7) is used as main indicator. In the uncertainty normalised Target diagram, the MQI represents the distance between the origin and a given station point, for this reason in previous documents the MQI was called also target indicator. The performance criterion for the MQI, defined as MQO, is set to unity regardless of spatial scale and pollutant and it is expected to be fulfilled for at least 90% of the available stations. A MQI value representative of the 90th percentile is calculated according to (9).

In the Target diagram, Figure 2, the X and Y axis correspond to the BIAS and $CRMSE$ which are normalised by the measurement uncertainty, $RMS_U$. The $CRMSE$ is defined as:

$$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}[(M_i - \bar{M}) - (O_i - \bar{O})]^2} \qquad (21)$$

and is related to RMSE and BIAS as follows:

$$RMSE^2 = BIAS^2 + CRMSE^2 \qquad (22)$$

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

$$CRMSE^2 = \sigma_o^2 + \sigma_m^2 - 2\sigma_o\sigma_m R \qquad (23)$$

For each point representing one station on the diagram the ordinate is then $BIAS/\beta RMS_U$, the abscissa is $CRMSE/\beta RMS_U$ and the radius is proportional to $RMSE/\beta RMS_U$. The green area on the Target plot identifies the area of fulfilment of the MQO, i.e. MQI less than or equal to 1.

Because $CRMSE$ is always positive only the right-hand side of the diagram would be needed in the Target plot. The negative X axis section can then be used to provide additional information. This information is obtained through relation (24) which is used to further investigate the $CRMSE$ related error and see whether it is dominated by $R$ or by σ. The ratio of two $CRMSE$, one obtained assuming a perfect correlation ($R = 1$, numerator), the other assuming a perfect standard deviation ($\sigma_M = \sigma_O$, denominator) is calculated and serves as basis to decide on which side of the Target diagram the point will be located:

$$\frac{CRMSE(R=1)}{CRMSE\ (\sigma_M = \sigma_{O)}} = \frac{|\sigma_M - \sigma_O|}{\sigma_O\sqrt{2(1-R)}} \begin{Bmatrix} >1 : \ \sigma\ dominates\ R : right \\ <1 : \ R\ dominates\ \sigma : left \end{Bmatrix} \qquad (24)$$

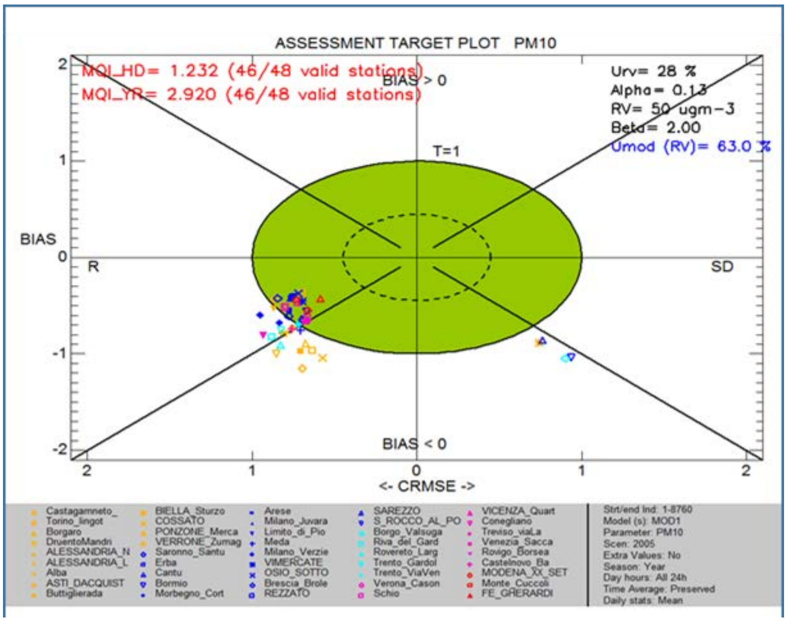
For ratios larger than one the σ error dominates and the station is represented on the right, whereas the reverse applies for values smaller than one.

The MQI associated to the 90th percentile worst station is calculated (see section 5.2.3) and indicated in the upper left corner. It is meant to be used as the main indicator in the benchmarking procedure and should be less or equal to one. Below this main indicator, also the MQI when using yearly average data is provided. Note that both MQI's have to fulfil the MQO (MQI < 1). The measurement uncertainty parameters (α, β, $U_r(RV)$ and RV) used to produce the diagram are listed on the top right-hand side. In blue color, the resulting model uncertainty is calculated according to equation (14) and is provided as output information.

In addition to the information mentioned above the proposed Target diagram also provides the following information:

- A distinction between stations according to whether their error is dominated by bias (either negative or positive), by correlation or standard deviation. The sectors where each of these dominates are delineated on the Target diagram by the diagonals in Figure 2.
- Identification of performances for single stations or group of stations by the use of different symbols and colours.

**Figure 2**: Example of Target diagram to visualise the main aspects of model performance. Each symbol represents a single station.



Source: JRC

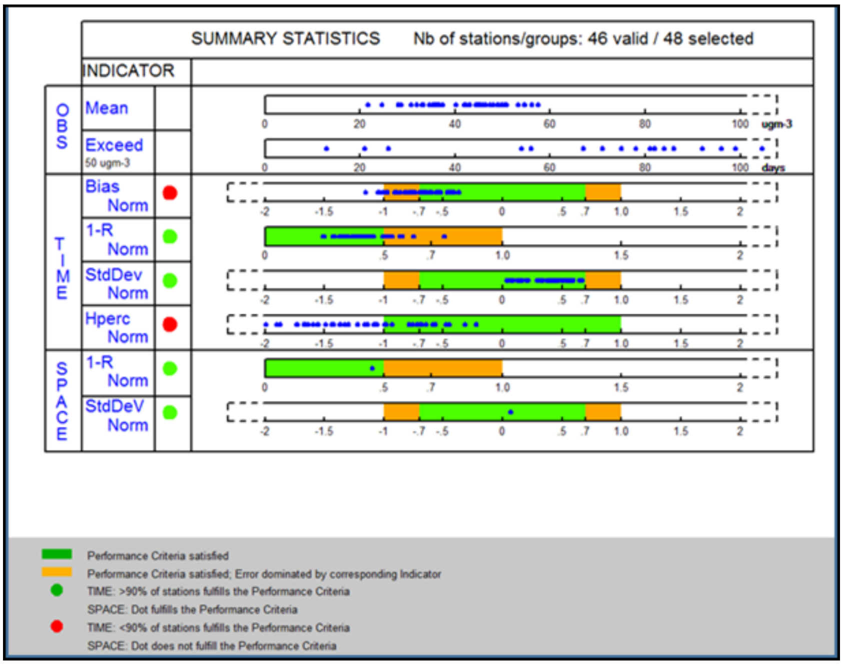
### 6.1.2 Summary Report

The summary statistics table, Figure 3, provides additional information on model performances. It is meant as an **additional** and **complementary** source of information to the MQO (Target diagram) to identify model strengths and weaknesses. The summary report is structured as follows:

- ROWS 1-2 provide the measured observed yearly means calculated from the hourly values and the number of exceedances for the selected stations. In benchmarking mode, the threshold values for calculating the exceedances are set automatically to 50, 200 and 120 µg/m3 for the daily PM10, the hourly NO2 and the 8h daily O3 maximum, respectively. For other variables (PM2.5, WS…) no exceedances are shown.
- ROWS 3-6 provide an overview of the temporal statistics for bias (row 3), correlation (row 4) and standard deviation (row 5) as well as information on the ability of the model to capture the highest range of concentration values (row 6). Values for the three first parameters are estimated using equations (16), (17) and (18). The fourth indicator is discussed in Section 7.1. Note that for the correlation a normalised indicator based on "1 – correlation" is plotted. As a result excellent stations have a value close to zero. Each point represents a specific station. The points for stations for which the model performance criterion is fulfilled lie within the green and the orange shaded areas. If a point falls within the orange shaded area the error associated with the particular statistical indicator is dominant. Note again that fulfilment of the bias, correlation, standard deviation and high percentile related indicators does not guarantee that the overall MQO based on the MQI (or RMSE, visible in the Target diagram) is fulfilled.
- ROWS 7-8 provide an overview of spatial statistics for correlation and standard deviation. Average concentrations over the selected time period are first calculated for each station and these values are then used to compute the averaged spatial correlation and standard deviation [as per equations (19) and (20)]. As a result, only one point representing the spatial correlation of all selected stations is plotted. Colour shading follows the same rules as for rows 3-5.

Note that for indicators in rows 3 to 8, values beyond the proposed scale will be represented by the station symbol being plotted in the middle of the dashed zone on the right/left side of the proposed scale. For all indicators, the second column with the coloured circle provides information on the number of stations fulfilling the performance criteria: the circle is coloured green if more than 90% of the stations fulfil the criterion and red if the number of stations is lower than 90%.

**Figure 3**: Example of a summary report based on hourly model results.



Source: JRC

## 6.2 Yearly average data

For the evaluation and reporting of yearly averaged model results, a Scatter diagram is used to represent the MQO instead of the Target plot because the CRMSE is zero for yearly averaged results so that the RMSE is equal to the BIAS in this case. The report then consists of a Scatter Diagram followed by the Summary Statistics (Figure 4).

### 6.2.1 Scatter Diagram

Equation (8) for yearly averaged results (i.e. based on the bias) is used as main model quality indicator. In the scatter plot, it is used to represent the distance from the 1:1 line. As mentioned above it is expected to be fulfilled (points are in the green area) by at least 90% of the available stations and a MQI value representative of the 90th percentile is calculated according to (9). The uncertainty parameters (α, β, $U_r(RV)$, $\mathrm{N_{np}}$, $\mathrm{N_p}$ and RV) used to produce the diagram are listed on the top right-hand side together with the associated model uncertainty calculated from (14).

The Scatter diagram also provides information on performances for single stations or groups of stations (e.g. different geographical regions in this example below) by the use of symbols and colours. The names of the stations are given as legend below the scatterplot.

### 6.2.2 Summary Report

The summary statistics table provides additional information on the model performance. It is meant as an additional and complementary source of information to the bias-based MQI to identify model strengths and weaknesses. It is structured as follows:

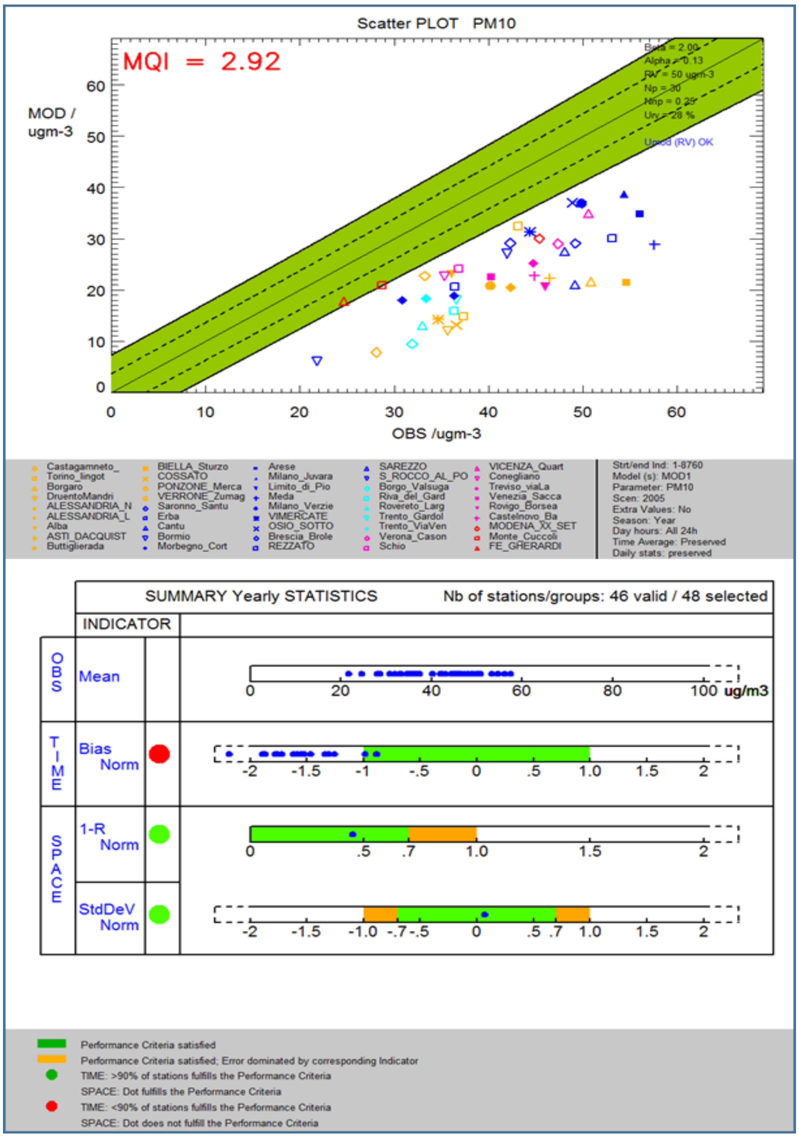
- ROW 1 provides the measured observed means for the selected stations.

- ROW 2 provides information on the fulfilment of the bias-based MQO for each selected station. Note that this information is redundant as it is already available from the scatter diagram but this is kept in the summary report so that it can be used independently of the scatter diagram.
- ROWS 3-4 provide an overview of spatial statistics for correlation and standard deviation. Annual values are used to calculate the spatial correlation and standard deviation. Equations (19) and (20) are used to check fulfilment of the performance criteria. The green and the orange shaded zones represent the area where the model performance criterion is fulfilled. If the point is in the orange shaded area, the error associated to the particular statistical indicator is dominant.

Note that for the indicators in rows 2 to 4, values beyond the proposed scale will be represented by plotting the station symbol in the middle of the dashed zone on the right/left side of the proposed scale.

The second column with the coloured circle provides information on the number of stations fulfilling the performance criteria: a green circle indicates that more than 90% of the stations fulfil the performance criterion while a red circle is used when this is less than 90% of the stations.

**Figure 4**: Example of main diagram (scatter) and a summary report based on yearly average model results.



Source: JRC

# 7 OPEN ISSUES

In this section, some topics are introduced on which there is no consensus yet within the FAIRMODE community and which merit further consideration.

## 7.1 Performance criteria for high percentile values

The MQI and MPI described in this document provide insight on the quality of the model average performances but do not inform on the model capability to reproduce extreme events (e.g. exceedances). For this purpose, a specific $MPI$ indicator is proposed as:

$$MPI_{perc} = \frac{\left|M_{perc} - O_{perc}\right|}{\beta U\left(O_{perc}\right)} \quad and \quad MPC\ is\ fulfilled\ when\ MPI_{perc} \leq 1 \tag{25}$$

where “perc” is a selected percentile value and Mperc and Operc are the modelled and observed values corresponding to this selected percentile. The denominator, U(Operc) is directly given as a function of the measurement uncertainty characterising the Operc value. For pollutants for which exceedance limit values exist in the legislation this percentile is chosen according to legislation.

For hourly NO2 this is the 99.8% (19th occurrence in 8760 hours), for the 8h daily maximum O3 92.9% (26th occurrence in 365 days) and for daily PM10 and PM2.5 90.1 (36th occurrence in 365 days). For general application, when e.g. there is no specific limit value for the number of exceedances defined in legislation, the 95% percentile is proposed. To calculate the percentile uncertainty used in the calculation of $MQI_{perc}$, eq. (37) is used with $O_i = O_{perc}$.

Current status: It is still under debate if the uncertainty of the percentile value can be calculated as the measurement uncertainty at $O_i = O_{perc}$. It can be assumed that a lower uncertainty value might be more appropriate since $O_{perc}$ is not a single observation but a statistical indicator of a time series. Apart from the extension of the MQI for percentiles as described above, the threshold evaluation criteria as implemented for forecast models (see further in Section 8) can also be used for the evaluation of high percentiles and episodes.

## 7.2 Application of the procedure to other parameters

Currently only PM, O3 and NO2 have been considered but the methodology could be extended to other pollutants such as heavy metals and polyaromatic hydrocarbons which are considered in the Ambient Air Quality Directive 2004/107/EC.

The focus in this document is clearly on applications related to the AQD 2008. However, the procedure can be extended to other variables including meteorological data as proposed in Pernigotti et al. (2014)

Current status: In the table below values are proposed for the parameters in (37) for wind speed and temperature data.

**Table 5:** List of the parameters used to calculate the uncertainty for the variables wind speed (WS) and temperature (TEMP).

| | $\gamma$ | $U_r^{RV}$ | $RV$ | A | $N_p$ | $N_{np}$ |
|---|---|---|---|---|---|---|
| **WS (test)** | 1.75 | 0.260 | 5 m/s | 0.89 | NA | NA |
| **TEMP (test)** | 1.75 | 0.05 | 25 K | 1.00 | NA | NA |

Source: JRC

When performing validation using the DELTA Tool, it is helpful to look at both NOx as well as NO2, as the former pollutant is less influenced by chemistry, and is therefore a better measure of the models’ ability to represent dispersion processes. The NOx measurement uncertainty is not available but could be approximated by the NO2 uncertainty for now.

# 8 FORECASTING & EXCEEDANCES INDICATORS

In this chapter, indicators and diagrams are proposed for the evaluation of model results in forecast mode. The main objective is to offer a common standardised template to facilitate the screening and comparison of forecast results.

First, it should be mentioned that the proposed Forecast Modelling Quality and Performance Indicators come on top of FAIRMODE's assessment MQO as defined in the previous chapters of this document. Therefore, it is recommended that forecast models fulfil the standard assessment MQO as well as the additional forecast MQO as defined here.

When evaluating a forecast model, two additional features of the model should be tested:

1. Sudden changes in the concentration levels (episodes) should be captured by the model
2. The exceedance of specific thresholds should be modelled well as such threshold exceedance can be used as trigger for short term action plans

To account for this, we will benchmark the forecast model with the so called "persistence model", which is the simplest method for predicting the future behaviour if no other information is available. The persistence model uses the measurements of the previous (day -1) as an estimate for the full forecast horizon and is by default not able to capture any changes in the concentration levels.

The methodology in its current form supports the following pollutants and time averages: the hourly NO2 daily maximum, the 8h O3 daily maximum and the daily PM10 and PM2.5 averaged concentrations. Note that only one value per day is used and that no evaluation of the entire hourly time profile is made.

In section 8.1 and 8.2 the Modelling Quality Indicator and the Modelling Performance Indicators are defined, respectively. Section 8.4 deals with the threshold indicators and in 8.4 the benchmarking diagrams as currently implemented in DELTA (vs7.0) are described in detail.

## 8.1 The forecast Modelling Quality Indicator

In forecast mode we are mostly interested to check the model ability to accurately reproduce daily forecasts and especially sudden changes in the concentration levels (episodes). The main evaluation assessment of the 'fitness for purpose' of a forecast application is based on the comparison with the persistence model. The MQI is defined as the difference between measured and modelled values, normalised by the root mean square error of the persistence model with respect to the measurements, i.e.

$$MQI_{forecast} = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}} \quad \text{and} \tag{26}$$

$$MQO_{forecast} \text{ is fulfilled if } MQI_{forecast} \leq 1 \text{ ,}$$

where Mi represents the forecasted value of model M for day i, where Pi is the value of the persistence model P on day i, Oi the measurement on day i, and N the number of days included in the time series. It is clear from the formula that MQI becomes 1 when the persistence model P is used as forecast model M. MQI values lower than 1 indicate better capabilities than the persistence model, whereas values larger than 1 indicate poorer performances.

Note that the persistence model is using the available observations from "the day before" as an estimate for all forecast horizons. So if today is 7 February, persistence model uses data referring to yesterday (6 February) for all forecast data produced today. Considering as an example a 3 day-forecast, modelled data for 7, 8, and 9 February are produced and all of them will be compared to 6 February measurement data. In other word, Pi refers to Oi-1 for "today forecast" (7 February), it refers to Oi-2 for tomorrow forecast (8 February) and it refers to Oi-3 for the day after tomorrow forecast (9 February). More generally, the persistence model is related to the forecast horizon as:

$$P_i = O_{i-1-forecast\ horizon} \pm U\left(O_{i-1-forecast\ horizon}\right) \tag{27}$$

where forecast horizon ranges from 0 for a “today forecast” or “now cast” up to typical 3 to 5 for longer term forecasts.

Note, in equation 27, that current Pi formulation includes the measurement uncertainty, and this prevents the denominator to tend to zero (i.e. the overall indicator cannot tend to infinity). Moreover, since the measurement uncertainty is concentration dependent (larger relative uncertainties are expected in the low concentration range) model performances at low concentration levels have less impact on MQI calculation.

## 8.2 The forecast Modelling Performance Indicators

In addition to the main MQI, Modelling Performance Indicators (MPI) are defined in order to support the interpretation of results. More in detail, we define a MPI related to the Mean Fractional Error (MFE) statistical indicator, as:

$$MFE = \frac{2}{N}\sum_{i=1}^{N}\frac{|M_i - O_i|}{(M_i + O_i)} \qquad (28)$$

MFE is chosen because it is a normalized error. It helps in interpreting the outcomes, since it does not depend on the magnitude of the corresponding concentration values. Moreover, MFE is already used in the framework of FAIRMODE activities5 and performance criteria and goals, based on this indicator, have been defined in literature for PM (Boylan and Russell, 2006) and O3 (Chemel et al., 2010).

Once the MFE indicator is chosen, two different MPIs are formulated following two rules.

1. Consistently with the forecast MQI formulation, an indicator is constructed to compare MFEf from the forecast model to the MFEp from the persistence model.
2. Forecast performances are also evaluated, regardless of persistence aspects, using an acceptability threshold based on measurement uncertainty. More in detail MFE is compared with the Mean Fractional Uncertainty (MFU), defined as follow

$$MF_U = \frac{1}{N}\sum_{i=1}^{N}\frac{2U(O_i)}{O_i} \qquad (29)$$

The definition of MFU as the acceptability threshold derives from considering

$$MFE = \frac{2}{N}\sum_{i=1}^{N}\frac{|M_i - O_i|}{(M_i + O_i)} \sim \frac{1}{N}\sum_{i=1}^{N}\frac{|M_i - O_i|}{O_i}$$

and then setting the condition $|M_i - O_i| \leq 2U(O_i)$ consistently with the assumptions described in section 5.2.1.

Rules 1 and 2 turn into the following formulas (30) and (31), respectively.

**Table 6:** Model performance indicators and criteria for MFE.

| MPI | Equation | MPC |
|---|---|---|
| $MPI = \frac{MFE_f}{MFE_p}$ | (30) | $\text{MPC: MPI} \leq 1$ |
| $MPI = \frac{MFE_f}{MF_U}$ | *(31)* | |

Source: JRC

[5] https://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/WG2_SG4_benchmarking_V2.pdf

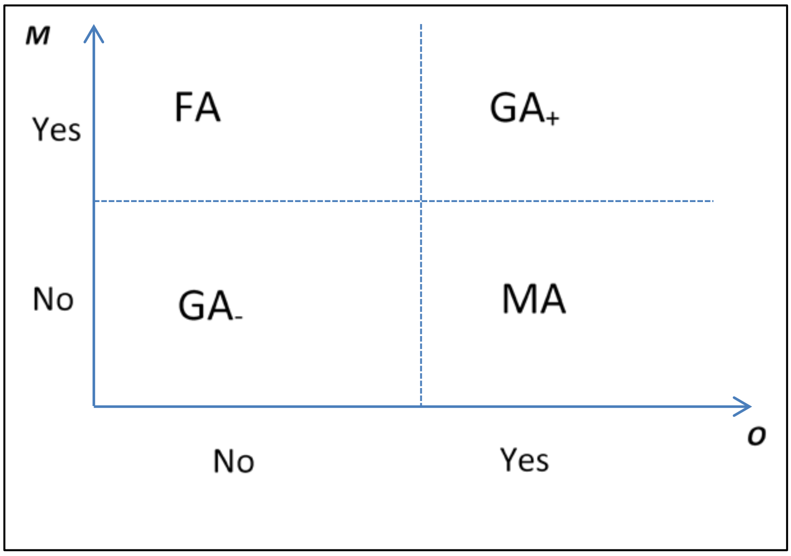
Using uncertainty parameters set in Table 7, double relative uncertainties $2U(O_i)/O_i$ as a function of concentration values show larger values in the low concentration range and then tend towards 0.5 (for NO2), 0.3 (for O3), 0.55 (for PM10), 0.7 (for PM2.5) at higher concentration values. So the choice of MFU as acceptability threshold is consistent with performances criteria and goals defined in literature for PM and O3 (Boylan and Russell, 2006; Chemel et al., 2010) but it has the advantage that it does not introduce any additional free (and arbitrary) parameters. In addition, its formulation, based on measurement uncertainty, is consistent with FAIRMODE approach and can be applied to all pollutants for which uncertainty parameters in Table 7 are set.

## 8.3 Threshold exceedances indicators

In addition to the main MQI and MPIs, based on the comparison with the persistence model, some commonly used indicators related to threshold exceedances are defined, based on the 2x2 contingency table representing the joint distribution of categorical events (below or above the threshold) by the model (M) and by the observations (O) as presented in Figure 5. In this framework, four quantities can be defined:

- False Alarms (FA): Model values are above the threshold but not the observations
  - M > threshold & O ≤ threshold
- Missed Alarms (MA): Model values are below the threshold but observed values are above it
  - M ≤ threshold & O > threshold
- Good values below threshold (GA-): both model and observation are below or equal to the threshold.
  - M & O ≤ threshold
- Good values above threshold (GA+): both model and observations are above the threshold.
  - M & O > threshold

**Figure 5**: Schematic outline of the threshold exceedance quantities GA+, GA-, FA and MA. The threshold is indicated by the dashed line.



Source: JRC

As a consequence, the counted alarms CA = GA+ + MA include all cases where O > threshold.

For a good forecast both FA and MA are small compared to GA+ and GA-. Based on these quantities the following indicators can be calculated:

- Probability of Detection: POD = GA+/(MA + GA+)
- Success Ratio: SR = 1 - False Alarm Ratio: 1 - FAR = 1 - FA/(FA + GA+) = GA+/(FA + GA+)

The POD indicator is comparing the correct modelled alerts with the **observed** alerts whereas the SR indicator is comparing the correct modelled alerts with all alerts **issued** by the model.

We also define four additional indicators as:

- FBias score: FBIAS= (GA+ + FA) / (MA + GA+)
- Accuracy: ACC = (GA+ + GA-) / Total
- Threat score: TS = GA+ / (MA + FA + GA+) = GA+/(FA + CA)
- Gilbert Skill score: GSS = (GA+ - Hrandom) / (MA + FA + GA+ - Hrandom)

  with Hrandom = (GA+ + MA)(GA+ + FA) / Total

## 8.4 Diagrams in the DELTA tool

When the DELTA tool is used in forecast mode, a number of specific forecast diagrams are produced that reflect the evaluation framework described above.

### 8.4.1 Forecast Target Plot

In the Forecast target plot (see Figure 6) information is included on the following quantities (all normalised by the root mean squared error of the persistence model):

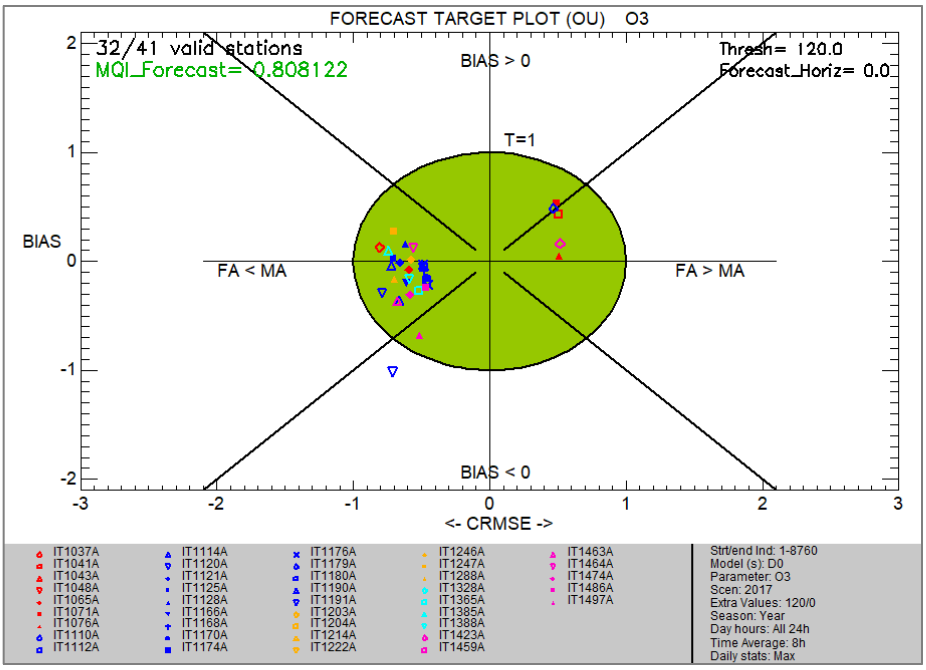
- Target forecast (RMSE): distance from the origin to the point (if distance is inferior to one, then the model behaves better than the persistence approach)
- BIAS: the bias can be either positive or negative and is represented along the vertical axis (Y)
- CRMSE: The CRMSE is always positive and given by the distance from the origin to the point along the X axis.
- False Alarm (FA) vs. Missed Alarm (MA): CRMSE is still on the X axis but we use the FA/MA ratio to differentiate the negative and positive portions of the X axis. This ratio is used to differentiate the right and left parts of the target diagram:

$$If\ \frac{FA}{MA} < 1 \Rightarrow Left$$

$$If\ \frac{FA}{MA} \geq 1 \Rightarrow Right$$

Values lower than one (within the green circle) indicate better capabilities than the persistence model whereas values larger than one indicate poorer performances. An example of the Forecast target plot (see Figure 6) is provided below for the case of a single model forecast. The MQI_Forecast value corresponding to the 90th largest percentile is printed in the left upper corner and should be lower than 1.

**Figure 6**: Forecast Target plot. The options selected (threshold and forecast horizon) are reported in the right hand top corner of the figure. The stations are represented by various symbols.

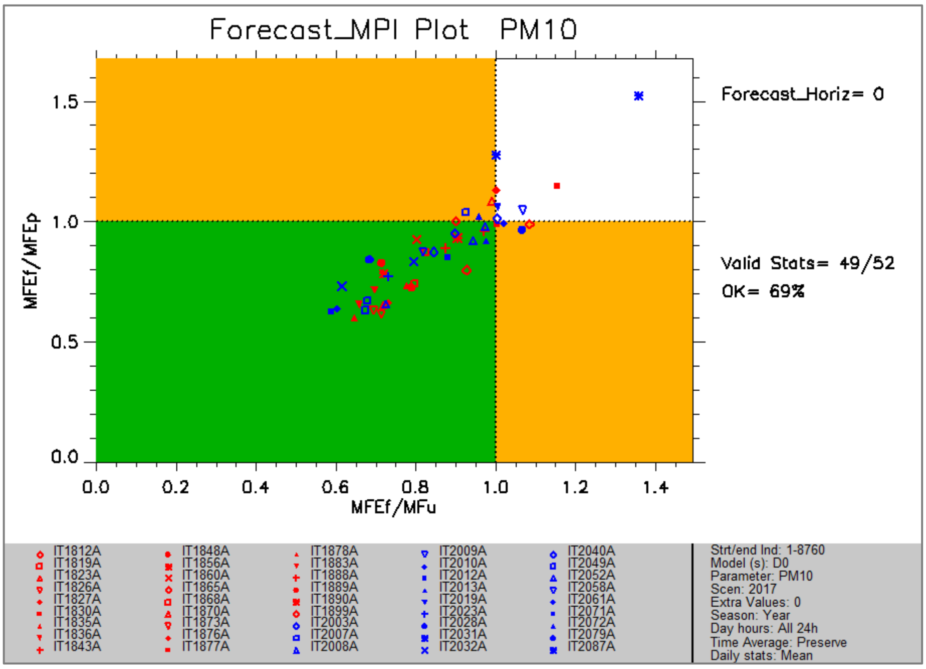


Source: JRC

### 8.4.2 Forecast MPI Plot

In addition to the MQI Target Plot, the MPI indicators defined in Section 8.2 help interpreting the comparison of forecast model performances with persistence model ones. The fulfilment of the MPC defined in Table 6 is provided by the Forecast MPI Plot (see Figure 7), where forecast performances (MFEf) are compared to Mean Fractional Uncertainty (MFU) along the X axis (by equation 31) and to the persistence model performances (MFEp) along Y axis (by equation 30). The green area identifies the area of fulfilment of both criteria. The orange areas indicate where only one of them is fulfilled.

**Figure 7**: Forecast MPI plot. Forecast horizon value is reported together with the fraction of valid station and the percentage of stations were both the MPCs are fulfilled.



Source: JRC

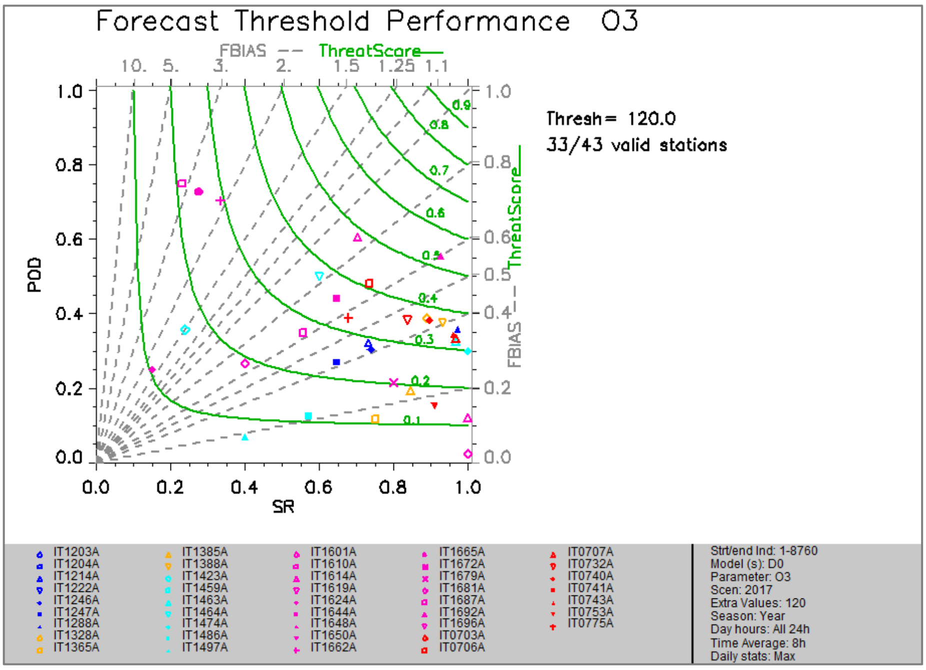
### 8.4.3 Forecast Threshold Performance Plot

In addition to the comparison with the persistence model, indicators related to threshold exceedances are needed as well to evaluate the performance of the forecast model. They are defined in Section 8.3. The four forecast indicators POD (Probability of Detection), SR (Success Ratio), FBIAS (Fractional bias) and TS (Threat score) can be put into one Forecast Threshold Performance Plot which is depicted in Figure 8. It is based on the SR values on the X axis and POD values on the Y axis. Since FBIAS and TS are indicators related to POD and SR, they are represented by additional isolines. Good forecasts with a high POD and SR are situated in the upper right corner.

Inspection of this Forecast Threshold Performance Plot still does not indicate whether a forecast model is "good enough". Therefore, the POD and SR values obtained with the persistence model are used again as a benchmark. This normalisation gives rise to the Forecast Threshold Normalized Performance Plot (Figure 9). In this plot, the green area represents forecasts with better POD and SR threshold indicators than the persistence model. In the white zone, the model performs worse than the persistence model on both indicators. In the orange zone, one of the two indicators is better than the benchmark.

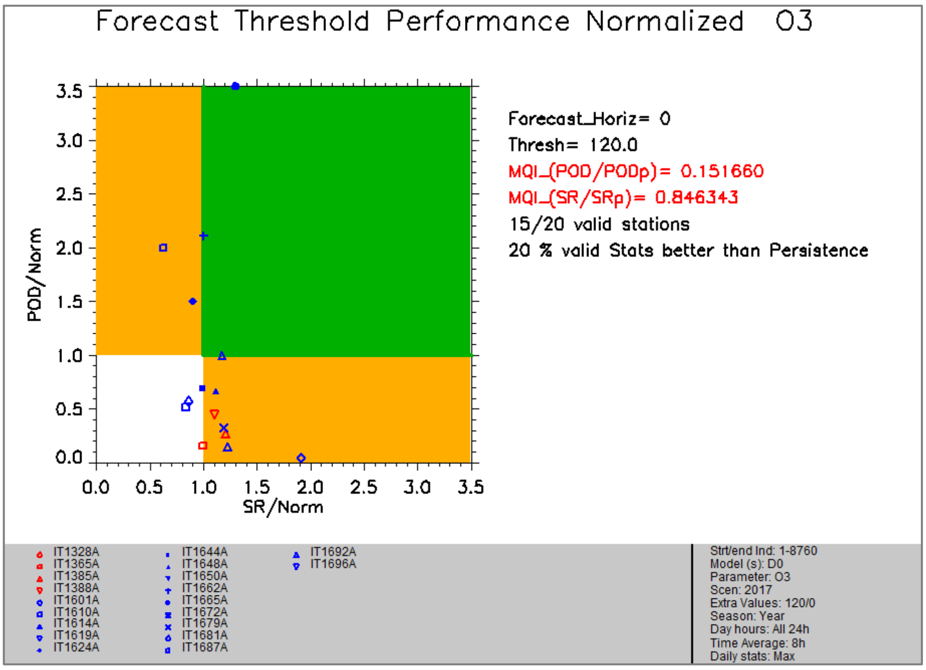
The normalised POD and SR values (i.e. POD/PODp and SR/SRp) are also given as indicative Modelling Quality Indicators in the Forecast Threshold Normalized Performance Plot. These MPI correspond to the 10th largest percentile value and should be larger than one for a "good enough" forecast.

**Figure 8:** Forecast Threshold Performance plot including the indicators POD, SR, FBIAS, and TS.



Source: JRC

**Figure 9**: Forecast Threshold Normalized Performance Plot as a normalised version of the Forecast Performance Diagram. POD and SR values obtained with the persistence model are used as benchmark and normalisation factor for the X and Y axis.

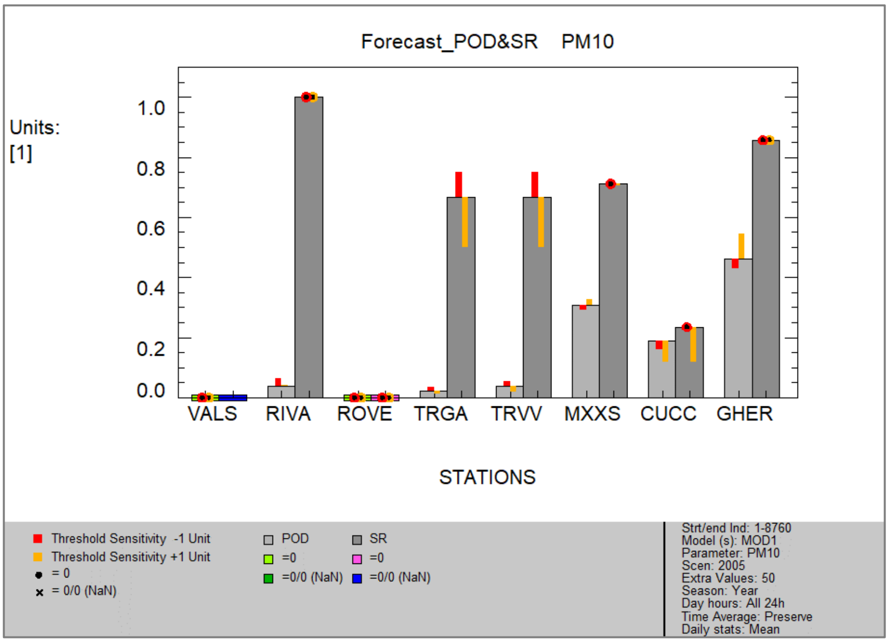


Source: JRC

### 8.4.4 Diagrams for exceedance indicators

In the Forecast Threshold Performance Plot and in the Forecast Threshold Normalized Performance Plot a combination of exceedance indicators is given. However, it remains interesting to look at individual indicators as well. Here we focus on the POD and SR indicators since they are independent quantities. The bar plot for POD & SR (Figure 10) shows the values for POD (light grey) and SR (dark grey) for each station, together with their negative and positive sensitivities with respect to the threshold in red and yellow. The red bar indicates the change in the indicator when the threshold is reduced by 1 unit, the yellow bar the change when the threshold is increased by 1 unit.

**Figure 10**: POD & SR diagram. Per station the POD (grey) and SR (dark grey) values are given. The sensitivity with respect to the threshold value is expressed by the red and yellow bars. Values of 0 (zero) and 0/0 (NaN) are indicated by different colours.



Source: JRC

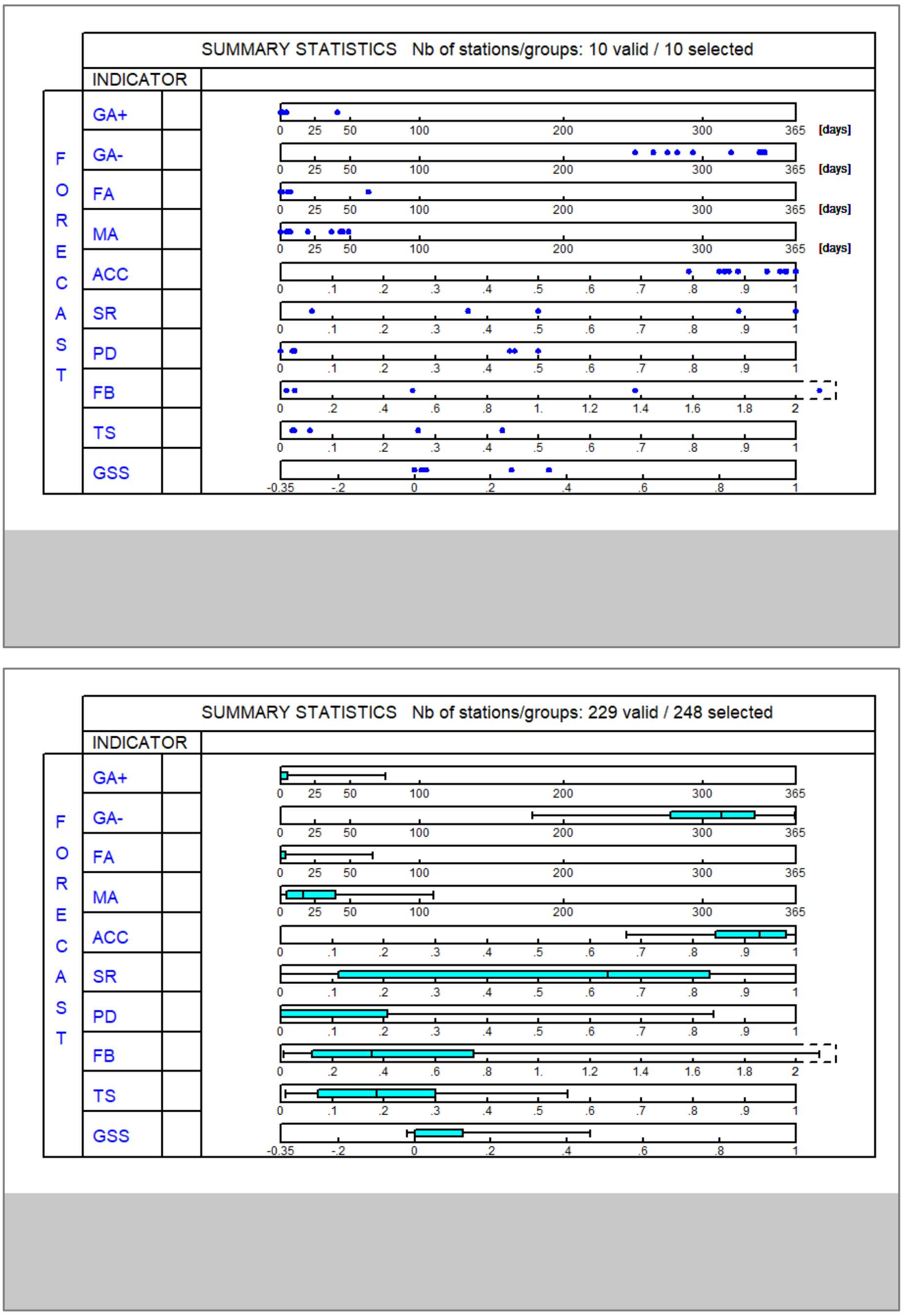
### 8.4.5 Summary report

As with the assessment benchmarking report, the Target Plot is complemented by a Summary Report. The following indicators are included in this report:

| Indicator | Optimal value |
|---|---|
| **GA+** | → Counted exceedances |
| **GA-** | → Counted non-exceedance |
| **FA** | → 0 |
| **MA** | → 0 |
| **ACC** | → 1 |
| **SR=1-FAR** | → 1 |
| **PD (POD)** | → 1 |
| **FB (FBIAS)** | → 1 |
| **TS** | → 1 |
| **GSS** | → 1 |

The Forecast Summary Report is graphically presented in Figure 11. A different graphical layout is applied depending on the number of stations taken into account in the analysis: if the number of stations is below 15, each of the dots represents a station for which the forecast indicators are evaluated (Figure 11, top); if the number of stations is above or equal 15, boxplots are used to represent the statistical distribution of the indicators values (Figure 11, bottom).

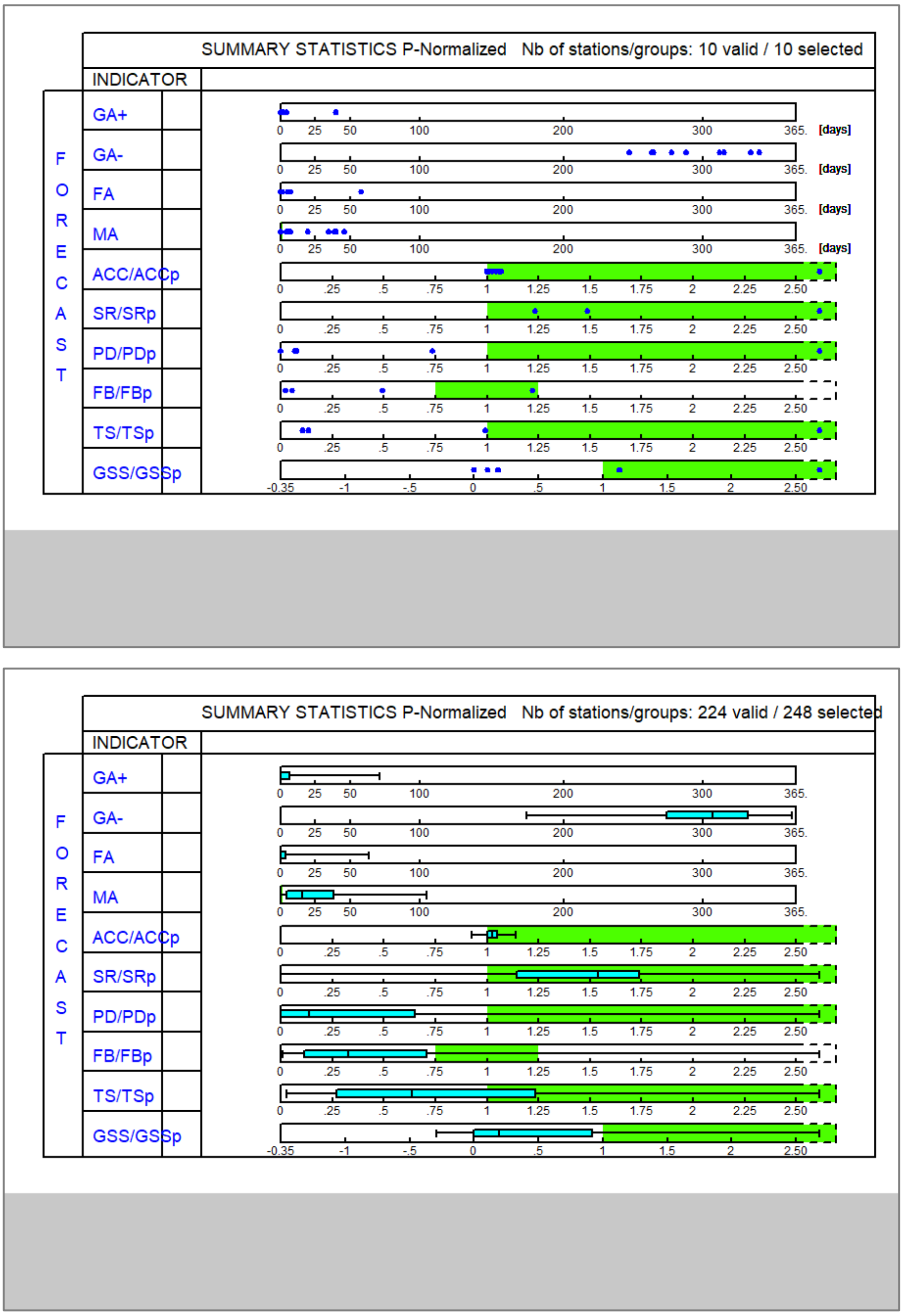
**Figure 11**: Forecast Summary Report, when the number of number of station is below 15 (top) and when the number of stations is above or equal 15 (bottom).



Source: JRC

In order to evaluate whether a forecast model is “good enough”, indicator values obtained with the persistence model can be used as a benchmark. A normalized version of the Forecast Summary Report can be produced, where the “good enough” zone is shaded in green and indicates that the model performs better than the persistence model for this particular indicator. An example of the Forecast Summary P-Normalized Report is provided in Figure 12 when the number of station is below 15 (top) and when the number of stations is above or equal 15 (bottom).

**Figure 12**: Forecast Summary P-Normalized Report, when the number of number of station is below 15 (top) and when the number of stations is above or equal 15 (bottom)..



Source: JRC

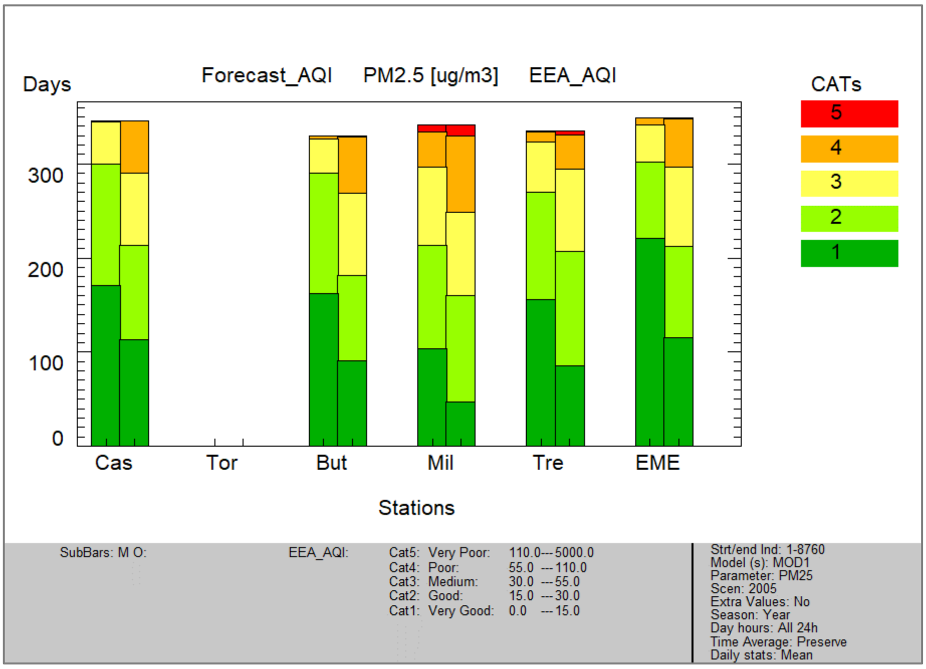
### 8.4.6 Forecast diagrams using Air Quality Indices

The Forecast diagrams in the previous sections were all based on a single value of the threshold. In this section we propose a plot based on multiple thresholds as they appear in the Air Quality Categories and their Indices, like EEA, UK or US EPA indicators. In this diagram we compare the number of days that the forecast model M, and the Measurements have in common in each of the Air Quality Categories.

In Figure 13 we show for each station two bars. For the forecast model M, and the measurement O, the vertical bar shows the number of days in one of the 5 indices (categories) of the EEA Air Quality Index table. The Index table itself is shown in the grey area below, the corresponding colours on the right-hand side of the graphic.

In the current version of the DELTA Tool (vs7.0, the following AQI tables are available: EEA (5 indices), UK4 (4 indices), UK10 (10 indices), USEPA (7 indices), and can be selected in the aqibounds.dat file in the DELTA tool configuration folder.

**Figure 13**: Forecast Air Quality Index diagram showing the number of days in each index category. Two bars (Model and Observation) per station.



Source: JRC

**Forecast Modelling Quality Indicator (MQI) and Objective (MQO):**

- The Forecast MQI is defined as the ratio of the RMSE between measured and modelled values and the same indicator computed for the persistence model
- MQO is the criteria for MQI and is fulfilled when MQI is less than or equal to 1
- MQI values lower than 1 indicate better capabilities than the persistence model, whereas values larger than 1 indicate poorer performances
- MQO must be fulfilled for at least 90% of the available stations
- Persistence model formulation includes the measurement uncertainty

**Forecast Modelling Performance Indicators (MPI) and Criteria (MPC):**

- Forecast MPIs are performance indicators that complement the MQI, supporting the interpretation of the result
- Forecast MPIs are related to the Mean Fractional Error (MFE)
- The MFE of the forecast model is compared with both the MFE of the persistence model and the Mean Fractional Uncertainty (MFU)
- MPCs are the criteria for MPIs, defined in Table 6

**Threshold exceedances indicators:**

- Four quantities are defined: False Alarms (FA); Missed Alarms (MA); Good values below threshold (GA-); Good values above threshold (GA+)
- Commonly used indicators are defined combining this quantities
- The Probability of Detection (POD) compares the correct modelled alerts with all observed alerts
- The Success Ratio (SR) compares the correct modelled alerts with all model alerts

# 9 CONCLUSIONS

This guidance document describes a procedure for air quality model benchmarking in the context of the Air Quality Directive 2008/50/EC (AQD). It defines proper modelling quality indicators and criteria to be fulfilled in order to allow sufficient level of quality for a given model application under the AQD.  The focus initially on applications related to air quality assessment has gradually been expanded to other applications and forecasting applications are now addressed as well. We explain and summarise the current concepts of the modelling quality objective methodology, elaborated in various papers and documents in the FAIRMODE community, addressing model applications for air quality assessment and forecast. We also present and explain templates for harmonised reporting of modelling results. Finally, remaining open issues in the implementation of the presented methodology are discussed to trigger further research and discussions.

# 10 REFERENCES

## 10.1 Peer reviewed articles

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

Carnevale C., G. Finzi, A. Pederzoli, E.Pisoni, P. Thunis, E.Turrini, M.Volta (2014), Applying the Delta tool to support AQD: The validation of the TCAM chemical transport model, Air Quality, Atmosphere and Health , 10.1007/s11869-014-0240-4

Carnevale C., G. Finzi, A. Pederzoli, E. Pisoni, P. Thunis, E. Turrini, M. Volta. (2015), A methodology for the evaluation of re-analysed PM10 concentration fields: a case study over the Po Valley, Air quality Atmosphere and Health, 8, p.533-544

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

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

Joliffe I.T. and David B. Stephenson (Eds), Forecast verification: a practitioner's guide in atmospheric science, ISBN 0-471-49759-2, 247pp, John Wiley & Sons Ltd, 2003

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

Monteiro, A., Durka, P., Flandorfer, C., Georgieva, E., Guerreiro, C., Kushta, J., Malherbe, L., Maiheu, B., Miranda, A.I., Santos, G., Stocker, J., Trimpeneers, E., Tognet, F., Stortini, M., Wesseling, J., Janssen, S., Thunis, P. (2018), Strengths and weaknesses of the FAIRMODE benchmarking methodology for the evaluation of air quality models, Air Quality, Atmosphere & Health, 11, p.373–383.

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

Stidworthy A., D. Carruthers, J. Stocker, D. Balis, E Katragkou, J Kukkonen, (2013), Myair toolkit for model evaluation, Proceedings of the 15th International Conference on Harmonisation, Madrid, Spain, 2013.

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

Thunis P., A. Pederzoli, D. Pernigotti (2012), Performance criteria to evaluate air quality modelling applications. Atmospheric Environment, 59, p.476-482

Thunis P., E. Georgieva, A. Pederzoli (2012), A tool to evaluate air quality model performances in regulatory applications, Environmental Modelling & Software 38, p.220-230

## 10.2 Reports/ working documents / user manuals

P. Thunis, A. Pederzoli, D. Pernigotti (2012) FAIRMODE SG4 Report Model quality objectives Template performance report & DELTA updates, March 2012. http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_March2012.pdf

D. Pernigotti, P. Thunis and M. Gerboles (2014), Modeling quality objectives in the framework of the FAIRMODE project: working document, http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Working%20note_MQO.pdf

P. Thunis, E. Georgieva, A. Pederzoli (2011), The DELTA tool and Benchmarking Report template Concepts and User guide, Joint Research Centre, Ispra Version 2 04 April 2011 http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/FAIRMODE_SG4_Report_April2011.pdf

P. Thunis, E. Georgieva, S. Galmarini (2011), A procedure for air quality models benchmarking, Joint Research Centre, Ispra Version 2 16 February 2011 http://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/WG2_SG4_benchmarking_V2.pdf

P. Thunis, C. Cuvelier, A. Pederzoli, E. Georgieva, D. Pernigotti, B. Degraeuwe (2016), DELTA Version 5.3 Concepts / User's Guide / Diagrams Joint Research Centre, Ispra, September 2016

ISO 13528: Statistical methods for use in proficiency testing by interlaboratory comparison.

JCGM 200, International vocabulary of metrology — Basic and general concepts and associated terms (VIM), 2008.

## List of abbreviations and definitions

| | |
|---|---|
| ACC | ACCuracy score (forecast) |
| AQD | EU Air Quality Directive |
| AQUILA | EU network for air quality monitoring |
| CA | Counted Alarms (forecast) |
| CAFÉ | EU Clean Air For Europe programme |
| CTM | Chemistry Transport Model |
| DQO | Data Quality Objective |
| EEA | European Environmental Agency |
| EU | European Union |
| FA | False Alarms (forecast) |
| FAIRMODE | EU Forum for AIR quality MODElling |
| FBIAS | Fractional bias |
| GA | Good Alarms (forecast) |
| GDE | Guide for Demonstration of Equivalence |
| GSS | Gilbert Skill Score (forecast) |
| GUM | Guide to the expression of Uncertainty in Measurement |
| LV | Limit Value |
| M | Model |
| MA | Missed Alarms (forecast) |
| MFE | Mean Fractional Error |
| MPC | Modelling Performance Criteria |
| MPI | Modelling Performance Indicator |
| MQI | Model Quality Indicator |
| MQO | Model Quality Objective |
| NO2 | Nitrogen Dioxides |
| O | Observations |
| O3 | Ozone |
| PM | Particulate Matter |
| POD | Probability Of Detection |
| RMSE | Root Mean Square Error |
| SR | Success Ratio (forecast) |
| TS | Threat Score (forecast) |
| U | Measurement Uncertainty |
| UK | United Kingdom |
| USEPA | United States Environmental Protection Agency |
| UV | Ultra Violet |
| 4DVAR | 4-dimensional data assimilation |

## List of figures

**Figure 1**: MQI and MQO explained on PM10 time series: measured (bold black) and modelled (bold red) concentrations are represented for a single station. The grey shaded area is indicative of the measurement uncertainty whereas the dashed black lines represent the MQO limits (proportional to the measurement uncertainty with β=2). Modelled data fulfilling the MQO must be within the dashed lines. ....................11

**Figure 2**: Example of Target diagram to visualise the main aspects of model performance. Each symbol represents a single station. ....................19

**Figure 3**: Example of a summary report based on hourly model results. ....................20

**Figure 4**: Example of main diagram (scatter) and a summary report based on yearly average model results. ....................21

**Figure 5**: Schematic outline of the threshold exceedance quantities GA+, GA-, FA and MA. The threshold is indicated by the dashed line. ....................25

**Figure 6**: Forecast Target plot. The options selected (threshold and forecast horizon) are reported in the right hand top corner of the figure. The stations are represented by various symbols. ....................27

**Figure 7**: Forecast MPI plot. Forecast horizon value is reported together with the fraction of valid station and the percentage of stations were both the MPCs are fulfilled. ....................27

**Figure 8**: Forecast Threshold Performance plot including the indicators POD, SR, FBIAS, and TS. ....................28

**Figure 9**: Forecast Threshold Normalized Performance Plot as a normalised version of the Forecast Performance Diagram. POD and SR values obtained with the persistence model are used as benchmark and normalisation factor for the X and Y axis. ....................29

**Figure 10**: POD & SR diagram. Per station the POD (grey) and SR (dark grey) values are given. The sensitivity with respect to the threshold value is expressed by the red and yellow bars. Values of 0 (zero) and 0/0 (NaN) are indicated by different colours. ....................29

**Figure 11**: Forecast Summary Report, when the number of number of station is below 15 (top) and when the number of stations is above or equal 15 (bottom). ....................31

**Figure 12**: Forecast Summary P-Normalized Report, when the number of number of station is below 15 (top) and when the number of stations is above or equal 15 (bottom).. ....................32

**Figure 13**: Forecast Air Quality Index diagram showing the number of days in each index category. Two bars (Model and Observation) per station. ....................33

## List of Tables

**Table 1:** Core set of statistical indicators ........................................................................... 9

**Table 2**: Comparison to AQD values. ..........................................................................13

**Table 3:** Model performance indicators and criteria for temporal statistics......................................14

**Table 4:** Model performance indicators and criteria for spatial statistics ........................................15

**Table 5**: List of the parameters used to calculate the uncertainty for the variables wind speed (WS) and temperature (TEMP)...........................................................................................22

**Table 6:** Model performance indicators and criteria for MFE. .....................................................24

**Table 7**: List of the parameters used to calculate the measurement uncertainty. ..............................42

# Annexes

## ANNEX 1. MEASUREMENT UNCERTAINTY

### A simple expression for the measurement uncertainty U(Oi) for time series

We derive here a simplified and general expression for the measurement uncertainty U(Oi). U(Oi) represents the expanded measurement uncertainty which can be expressed in terms of the combined uncertainty, $u(O_i)$ by multiplying with a coverage factor $k$:

$$U(O_i) = ku(O_i). \tag{32}$$

Each value of $k$ gives a particular confidence level so that the true value is within the confidence interval bounded by $O_i \pm ku(O_i)$. Coverage factors of $k$ = 2.0 and $k$ = 2.6 correspond to confidence levels of around respectively 95 and 99%, so that the unknown true value lies within the estimated confidence intervals.

In Thunis et al., 2013 a general expression for the combined measurement uncertainty is derived by considering that $u(O_i)$ of a measurement $O_i$, can be decomposed into a component that is proportional, $u_p(O_i)$ to the concentration level and a non-proportional contribution $u_{np}(O_i)$:

$$u^2(O_i) = {u_p}^2(O_i) + {u_{np}}^2(O_i) \tag{33}$$

The non-proportional contribution $u_{np}(O_i)$ is by definition independent of the concentration and can therefore be defined as a fraction α (ranging between 0 and 1) of the uncertainty at the reference value:

$${u_{np}}^2(O_i) = \alpha^2 u^2(RV) \tag{34}$$

On the other hand, the proportional component $u_p(O_i)$ can be estimated from

$${u_p}^2(O_i) = \eta^2\, O_i^2 \tag{35}$$

where n is a fraction of the measurements. Applying (32), (33) and (34) to the reference value Oi = RV leads to $\eta^2 = (1-\alpha^2)\cdot u_r^2(RV)$, where $u_r(RV)$ is the relative standard uncertainty around the reference value: $u_r^2(RV) = u^2(RV)/RV^2$ hence:

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

As representative values of the measurement uncertainty, the 95th percentile highest value among all uncertainty values is selected. To derive expressions for the uncertainty estimation for PM10 and PM2.5 the results of a JRC instrument inter-comparison (Lagler et al. 2011) have been used, whereas a set of EU AIRBASE stations available for a series of meteorological years has been used for NO2 and analytical relationships have been used for O3. Using the relation

$$ku(RV) = U(RV) = U_r(RV) * RV \tag{36}$$

and (32) to (34), $U$ can be expressed as:

$$U(O_i) = U_r(RV)\sqrt{(1-\alpha^2)O_i^2 + \alpha^2 RV^2} \tag{37}$$

From Equation (37) it is possible to derive an expression for $RMS_U$ as:

$$RMS_U = \sqrt{\frac{\sum_{i=1}^{N}\left(U(O_i)\right)^2}{N}} = U_r(RV)\sqrt{(1-\alpha^2)(\bar{O}^2 + \sigma_o^2) + \alpha^2 RV^2} \tag{38}$$

in which $\bar{O}$ and σ0 are the mean and the standard deviation of the measurement time series, respectively.

For yearly averaged measurements, the following expression for the 95th percentile uncertainty is derived:

$$U(\overline{\mathrm{O}}) = U_r(RV)\sqrt{\frac{(1-\alpha^2)}{\mathrm{N}_\mathrm{p}^*}(\overline{\mathrm{O}}^2 + \sigma_\mathrm{o}^2) + \frac{\alpha^2.\mathrm{RV}^2}{\mathrm{N}_\mathrm{np}}} \cong U_r(RV)\sqrt{\frac{(1-\alpha^2)}{\mathrm{N}_\mathrm{p}}\overline{\mathrm{O}}^2 + \frac{\alpha^2.\mathrm{RV}^2}{\mathrm{N}_\mathrm{np}}} \qquad (39)$$

where Np and Nnp are two coefficients that are only used for annual averages and that account for the compensation of errors (and therefore a smaller uncertainty) due to random noise and other factors like periodic re-calibration of the instruments. Details on the derivation of (39) and in particular the parameters Np and Nnp are provided in Pernigotti et al. (2013).

### Parameters for calculating the measurement uncertainty

Table 7 presents the values for the parameters in equations (38) and (39). All values are as reported in Pernigotti et al. (2013) and Thunis et al. (2012) with the exception of the Np and Nnp parameters for PM10 that have been updated to better account for the yearly average measurement uncertainty range with current values set to reflect uncertainties associated to the β-ray measurement technique. Because of insufficient data for PM2.5, values of Np and Nnp similar to those for PM10 have been set. The value of $U_r(RV)$ has also been updated for O3 where the coverage factor (k) has been updated to 2 (not 1.4 as in Thunis et al. 2012).

Note also that the value of α for PM2.5 referred to in the Pernigotti et al. (2014) working note has been arbitrarily modified from 0.13 to 0.30 to avoid larger uncertainties for PM10 than PM2.5 in the lowest range of concentrations.

**Table 7**: List of the parameters used to calculate the measurement uncertainty.

| | $U_r(RV)$ | $RV$ | α | $N_p$ | $N_{np}$ |
|---|---|---|---|---|---|
| **NO2** | 0.24 | 200 µg/m3 | 0.20 | 5.2 | 5.5 |
| **O3** | 0.18 | 120 µg/m3 | 0.79 | 11 | 3 |
| **PM10** | 0.28 | 50 µg/m3 | 0.25 | 20 | 1.5 |
| **PM2.5** | 0.36 | 25 µg/m3 | 0.50 | 20 | 1.5 |

Source: JRC

### Practical derivation of the measurement uncertainty parameters

To be able to apply Equation (37) it is necessary to estimate the relative uncertainty around a reference value and α, the non-proportional fraction around that reference value. Equation (37) is rewritten as:

$$U^2(O_i) = \alpha^2 U^2(RV) + U_r^2(RV)(1-\alpha^2)O_i^2 \qquad (40)$$

This is a linear relationship with slope, $m = (1-\alpha^2)U_r^2(RV)$ and intercept, $q = \alpha^2\, U^2(RV)$ which can be used to derive values for $U^2(RV)$ and α by fitting measured squared uncertainties $U^2(O_i)$ to squared observed values $(O_i)^2$.

An alternative procedure for calculating $U^2(O_i)$ and α is to use the generic equation formulation of a line passing through two points: L2 (for a low range concentration) and RV2 with $U(L)$ and $U(RV)$ their associated expanded uncertainties:

$$U^2(O_i) = U^2(L) + \frac{U^2(RV) - U^2(L)}{RV^2 - L^2}(O_i^2 - L_i^2) \qquad (41)$$

While Equation (40) requires defining values for both α and $U(RV)$ around a reference value ($RV$), Equation (41) requires defining uncertainties around two arbitrarily fixed concentrations ($RV$ and $L$).

## ANNEX 2. OVERVIEW OF EXISTING LITERATURE

The development of the procedure for air quality model benchmarking in the context of the AQD has been an on-going activity in the context of the FAIRMODE community. The JRC developed the DELTA tool in which the Modelling Performance Criteria (MPC) and Modelling Quality Objective (MQO) are implemented. Other implementations of the MPC and MQO are found in the CERC Myair toolkit and the on-line ATMOSYS Model Evaluation tool developed by VITO.

In the following paragraphs a chronological overview is given of the different articles and documents that have led to the current form of the MQO and MPC. Starting from a definition of the MPC and MQO in which the measurement uncertainty is assumed constant (Thunis et al., 2012) this is further refined with more realistic estimates of the uncertainty for O3 (Thunis et al., 2013) and NOx and PM10 (Pernigotti et al., 2013). The DELTA tool itself and an application of this tool are respectively described in Thunis et al., 2013, Carnevale et al., 2013 and Carnevale et al., 2014, , Georgieva et al., 2015. Full references to these articles can be found at the end of this document.

**Measurement uncertainty for a given pollutant:**

- Depends on the concentration level
- is estimated as 95th percentile highest value, based on: JRC instrument inter-comparison results (for PM); data from EU AIRBASE stations for series of meteorological years (for NO2) and on analytical relationships (for O3)
- is expressed by (38) for time series
- is given by (39) for yearly averaged values
- uses parameters in its calculation as defined in Table 7

**<u>Literature on how  MQO and MPC are defined</u>.**

### *Thunis et al., 2012: Performance criteria to evaluate air quality modelling applications*

This article introduces the methodology in which the root mean square error (RMSE) is proposed as the key statistical indicator for air quality model evaluation. A Modelling quality objective (MQO) and Model Performance Criteria (MPC) to investigate whether model results are 'good enough' for a given application are calculated based on the measurement uncertainty (U). The basic concept is to allow the same margin of tolerance (in terms of uncertainty) for air quality model results as for observations. As the objective of the article is to present the methodology and not to focus on the actual values obtained for the MQO and MPC, U is assumed to be independent of the concentration level and is set according to the data quality objective (DQO) value of the Air Quality Directive (respectively 15, 15 and 25% for O3, NO2 and PM10). Existing composite diagrams are then adapted to visualise model performance in terms of the proposed MQO and MPC. More specifically a normalised version of the Target diagram, the scatter plot for the bias and two new diagrams to represent the standard deviation and the correlation performance are considered. The proposed diagrams are finally applied and tested on a real case.

***Thunis et al., 2013: Model quality objectives based on measurement uncertainty. Part I: Ozone***

Whereas in Thunis et al., 2012 the measurement uncertainty was assumed to remain constant regardless of the concentration level and based on the DQO, this assumption is dropped in this article. Thunis et al., 2013 propose a formulation to provide more realistic estimates of the measurement uncertainty for O3 accounting for dependencies on pollutant concentration. The article starts from the assumption that the combined measurement uncertainty can be decomposed into non-proportional (i.e. independent from the measured concentration) and proportional fractions which can be used in a linear expression that relates the uncertainty to known quantities specific to the measured concentration time series. To determine the slope and intercept of this linear expression, the different quantities contributing to the uncertainty are analysed according to the direct approach or GUM6 methodology. This methodology considers the individual contributions to the measurement uncertainty for O3 of the linear calibration, UV photometry, sampling losses and other sources. The standard uncertainty of all these input quantities is determined separately and these are subsequently combined according to the law of propagation of errors. AIRBASE data for 2009 have been used in obtaining the measurement uncertainty. Based on the new linear relationship for the uncertainty more accurate values for the MQO and MPC are calculated for O3. MPC are provided for different types of stations (urban, rural, traffic) and for some geographical areas (Po Valle, Krakow, Paris).

***Pernigotti et al., 2013: Model quality objectives based on measurement uncertainty. Part II: PM10 and NO2***

The approach presented for O3 in Thunis et al., 2013 is in this paper applied to NO2 and PM10 but using different techniques for the uncertainty estimation. For NO2 which is not measured directly but is obtained as the difference between NOx and NO, the GUM methodology is applied to NO and NOx separately and the uncertainty for NO2 is obtained by combining the uncertainties for NO and NOx. For PM which is operationally defined as the mass of the suspended material collected on a filter and determined by gravimetry there are limitations to estimate the uncertainty with the GUM approach. Moreover, most of the monitoring network data are collected with methods differing from the reference one (e.g. automatic analysers), so-called equivalent methods. For these reasons the approach based on the guide for demonstration of equivalence (GDE) using parallel measurements is adopted to estimate the uncertainties related to the various PM10 measurements methods. These analyses result in the determination of linear expressions which can be used to derive the MQO and MPC. The Authors also generalise the methodology to provide uncertainty estimates for time-averaged concentrations (yearly NO2 and PM10 averages) taking into account the reduction of the uncertainty due to error compensations during this time averaging.

***Pernigotti et al., 2014: Modelling quality objectives in the framework of the FAIRMODE project: working document***

This document corrects some errors found in the calculation of the NO2 measurement uncertainty in Pernigotti et al., 2013 and assesses the robustness of the corrected expression. In a second part, the validity of an assumption underlying the derivation of the yearly average NO2 and PM10 MQO in which a linear relationship is assumed between the averaged concentration and the standard deviation is investigated. Finally, the document also presents an extension of the methodology for PM2.5 and NOx and a preliminary attempt to also extend the methodology for wind and temperature.

## Literature on the implementation and use of the Delta tool

### Thunis et al., 2012: A tool to evaluate air quality model performances in regulatory applications

The article presents the DELTA Tool and Benchmarking service for air quality modelling applications, developed within FAIRMODE by the Joint Research Centre of the European Commission in Ispra (Italy). The DELTA tool addresses model applications for the AQD, 2008 and is mainly intended for use on assessments. The DELTA tool is an IDL-based evaluation software and is structured around four main modules for respectively the input, configuration, analysis and output. The user can run DELTA either in exploration mode for which flexibility is allowed in the selection of time periods, statistical indicators and stations, or in benchmarking mode for which the evaluation is performed on one full year of modelling data with pre-selected statistical indicators and diagrams. The Authors also present and discuss some examples of DELTA tool outputs.

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

### Carnevale et al., 2014: 1. Applying the Delta tool to support AQD: The validation of the TCAM chemical transport model

This paper presents an application of the DELTA evaluation tool V3.2 and test the skills of the chemical transport model TCAM model by looking at the results of a 1-year (2005) simulation at 6km × 6km resolution over the Po Valley. The modelled daily PM10 concentrations at surface level are compared to observations provided by approximately 50 stations distributed across the domain. The main statistical parameters (i.e., bias, root mean square error, correlation coefficient, standard deviation) as well as different types of diagrams (scatter plots, time series plots, Taylor and Target plots) are produced by the Authors. A representation of the measurement uncertainty in the Target plot, used to derive model performance criteria for the main statistical indicators, is presented and discussed.

### Thunis and Cuvelier, 2020: DELTA Version 6.0 Concept / User's Guide / Diagrams

This is currently the most recent version of the user's guide for the DELTA tool. The document consists of three main parts: the concepts, the actual user's guide and an overview of the diagrams the tool can produce. The concepts part sets the application domain for the tool and lists the underlying ideas of the evaluation procedure highlighting that the tool can be used both for exploration and for benchmarking. The MQO and the MPCs that are applied are explained including a proposal for an alternative way to derive the linear expression relating uncertainty to measured concentrations. Examples of the model benchmarking report are presented for the cases model results are available hourly and as a yearly average. The actual user guide contains the information needed to install the tool, prepare input for the tool, and run the tool both in exploration and in benchmarking modes. Also details on how to customise certain settings (e.g. uncertainty) and how to use the included utility programs are given.

### Carnevale et al., 2014: A methodology for the evaluation of re-analysed PM10 concentration fields: a case study over the Po valley

This study presents a general Monte Carlo based methodology for the validation of Chemical Transport Model (CTM) concentration re-analysed fields over a certain domain. A set of re-analyses is evaluated by applying the measurement uncertainty (U) approach, developed in the frame of FAIRMODE. Modelled results from the Chemical Transport Model TCAM for the year 2005 are used as background values. The model simulation domain covers the Po valley with a 6 km x 6 km resolution. Measured data for both assimilation and evaluation are provided by approximately 50 monitoring stations distributed across the Po valley. The main statistical indicators (i.e. Bias, Root Mean Square Error, correlation coefficient, standard deviation) as well as different types of diagrams (scatter plots and Target plots) have been produced and visualised with the Delta evaluation Tool V3.6.

## ANNEX 3. RELATED TOOLS

### The DELTA Tool

The DELTA tool is an IDL evaluation software developed at EC-JRC, Ispra within the FAIRMODE activities. It was built upon the assets of the EuroDelta, CityDelta and POMI tools (Cuvelier et al., 2007, Thunis et al., 2007) and was designed for rapid diagnostics of air quality modelling applications under the EU Air Quality Directive 2008. The tool is based on pairs of measurement and modelled data at 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 model quality objective and model performance criteria. The main concept of the methodology is based on taking into account the measurement uncertainty while calculating model performance indicators related to RMSE, Correlation, BIAS and standard deviation. The main model performance indicator, called modelling quality indicator (MQI), is expected to fulfil a criteria (the model quality objective) easily viewable at the target diagram, part of the benchmarking template.

The tool (current version 7.0) is available upon request via the FAIRMODE website.

### The ATMOSYS benchmarking tool

The ATMOSYS (Policy support system for atmospheric pollution hotspots) system that was developed and evaluated in the context of a LIFE+ project (2010 – 2013) is an integrated Air Quality Management Dashboard that can be used for air pollution management and policy support in accordance with the 2008 EU CAFÉ Directive. ATMOSYS offers different tools to support air pollution forecasting and assessment one of which is an air quality model benchmarking tool that is based on the methodology developed in the context of FAIRMODE. The tool allows the user to upload comma separated (csv) text files with hourly modelled and observed concentration values and use these to calculate the target plot and summary statistics (see chapter 6). The benchmarking functionality is currently limited to hourly values. As ATMOSYS is based on a generic web-based interface it can easily be adopted in other regions and in 2015 the ATMOSYS model benchmarking tool was updated and implemented as the model evaluation service for the French national air quality monitoring system (http://www.lcsqa.org/).

### The MyAir Model Evaluation Toolkit

The MyAir Model Evaluation Toolkit has been designed to evaluate air quality models in terms of general performance. In addition, the MyAir Toolkit has specific features that assess the models' ability to calculate metrics associated with air quality forecasting, for example exceedances of daily limit values. The toolkit was developed as part of the GMES downstream service project, PASODOBLE, which produced local-scale air quality services for Europe under the name 'Myair'. The MyAir Toolkit consists of four tools: a questionnaire tool offering structured advice on the advisability of the proposed evaluation; a data input tool able to import a wide range of modelled and in-situ monitored data formats; a model evaluation tool that analyses the performance of the model at predicting concentrations and pollution episodes; and a model diagnostics tool that compares modelled and monitored data at individual stations in more detail. The Myair Toolkit is easy to use, produces statistical data and attractive graphs, and has a comprehensive User Guide. The tool is downloadable from http://www.cerc.co.uk/MyAirToolkit and further information can be found in Stidworthy et al. (2013).

KJ-NA-31068-EN-N

**The European Commission's science and knowledge service**
Joint Research Centre

**JRC Mission**
As the science and knowledge service of the European Commission, the Joint Research Centre's mission is to support EU policies with independent evidence throughout the whole policy cycle.

**EU Science Hub**
ec.europa.eu/jrc

@EU_ScienceHub

EU Science Hub - Joint Research Centre

EU Science, Research and Innovation

EU Science Hub

Publications Office
of the European Union

doi:10.2760/41988

ISBN 978-92-76-52425-0