21st International Conference on
Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes
27-30 September 2022, Aveiro, Portugal

QUALITY CONTROL INDICATORS FOR THE VALIDATION OF AIR QUALITY FORECAST APPLICATIONS IN THE FRAMEWORK OF FAIRMODE ACTIVITIES

Antonio Piersanti[1], Cornelis Cuvelier[2], Stijn Janssen[3], Alexandra Monteiro[4], Paweł Durka[5], Philippe Thunis[2] and Lina Vitali[1]

[1]National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA), Bologna, Italy
[2]European Commission - Joint Research Centre (JRC), Ispra, Italy
[3]Flemish Institute for Technological Research (VITO), Mol, Belgium
[4]CESAM, Department of Environment, University of Aveiro, Aveiro, Portugal
[5]Institute of Environmental Protection (IEP) - National Research Institute, Warsaw, Poland

Abstract: In the framework of FAIRMODE activities concerning the harmonization of model validation methodologies, a specific task was dedicated to the development of a common standardized template to facilitate the screening and comparison of air quality forecast results. The proposed approach assesses the forecast model performances using, as a benchmark, the so called "persistence model", which uses measurements of the previous day as an estimate for the full forecast prediction. Consistently with the FAIRMODE approach, the proposed formulation includes measurement uncertainty and relies on the definition of specific Model Quality Objective and Criteria. The main features of the methodology are described here, together with an example of its application to evaluate the quality of one year-long model data set produced by the Italian air quality forecast system FORAIR-IT.

Key words: *air quality forecast, regulatory model applications, model benchmarking, harmonized validation methodologies, model quality objective, measurement uncertainty***.**

INTRODUCTION

One of the main activities of FAIRMODE (Forum for Air Quality Modelling in Europe, http://fairmode.jrc.ec.europa.eu/) has been the development of harmonized procedures for the validation and the benchmarking of air quality model applications, especially under the implementation of the Ambient Air Quality Directive 2008/50/EC (AAQD). The main goal was the definition of common standardized Model Quality Objectives (*MQO*) and Model Performance Criteria (*MPC*) to be fulfilled in order to ensure a sufficient level of quality of a given model application. The methodology (Thunis et al., 2013; Pernigotti et al., 2013; Janssen and Thunis, 2022), consolidated in the DELTA Tool software (https://aqm.jrc.ec.europa.eu/index.aspx), has reached a good level of maturity and has been widely used and tested by model developers and users (Monteiro et al., 2018).
The approach was initially focused on applications related to air quality assessment, but was recently expanded to address additional issues typical of other model applications, such as forecasting. More in detail, the FAIRMODE working plan for the period 2020-2022 included a specific task (CT3, https://fairmode.jrc.ec.europa.eu/Activity/CT3) dedicated to the development and the testing of additional quality control indicators to be checked when evaluating a forecast application. The proposed methodology was tested in different national and geographical contexts and first outcomes sound promising, pointing out to the usefulness of the approach in highlighting shortcomings and strengths of forecasting applications. Here we present the main features of the methodology and an application for a case study (Italy).

METHODOLOGY AND APPLICATION EXAMPLE

The proposed methodology for forecast evaluation comes on top of FAIRMODE's approach applied for the validation of assessment modelling applications. Therefore, it is recommended that forecast models fulfil the standard assessment MQO, as defined in Janssen and Thunis (2022), as well as the additional forecast objectives and criteria as described here. More in detail, the specific forecast indicators investigate the capability to detect sudden changes of concentrations levels, to predict threshold exceedances and to

reproduce air quality indices. The methodology, as currently implemented in the DELTA Tool software (version 7.0), supports the following pollutants and time averages: NO_2 daily maximum and annual mean, O_3 daily maximum of 8-hour average, PM10 and PM2.5 daily and annual mean.

Comparison with the "persistence model"

When evaluating a forecast model, it is of main interest to verify its ability to accurately reproduce sudden changes in the pollutant's concentration levels. To account for this, the proposed approach assesses the forecast model performances using, as a benchmark, the so called "persistence model", which uses the measurements of the previous day as an estimate for the full forecast horizon and is by default not able to capture changes in the concentration levels (e.g. Mittermaier, 2008). More in detail, the forecast Model Quality Indicator (MQI_f) is defined as the ratio between the Root Mean Square Errors (RMSE) computed for both the forecast and the persistence models, i.e.

$$MQI_f = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(M_i - O_i)^2}{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}} \quad (1)$$

where M_i, P_i, O_i represents respectively the forecast, the persistence and the measured values for day i, and N is the number of days included in the time series. Since the persistence model uses the available observations from the day before as an estimate for all forecast days, it is related to the forecast horizon (FH) as following, where, consistently with the FAIRMODE approach, measurement uncertainty (Janssen and Thunis, 2022) is also taken into accout:

$$P_i = O_{i-1-FH} \pm U(O_{i-1-FH}) \quad (2)$$

The forecast Modelling Quality Objective (MQO_f) is fulfilled when MQI_f is less or equal to 1, indicating better capabilities of the forecast model than the persistence one for a specific application.

Modelling Quality Indicator values are provided by means of the Forecast Target Plots (Figure 1), where MQI_f is the distance between the origin and a given point (representing each monitoring station). The green area identifies the fulfilment of the MQO_f. The MQI_f associated to the 90th percentile worst station is reported in the upper left corner (Janssen and Thunis, 2022 for details).

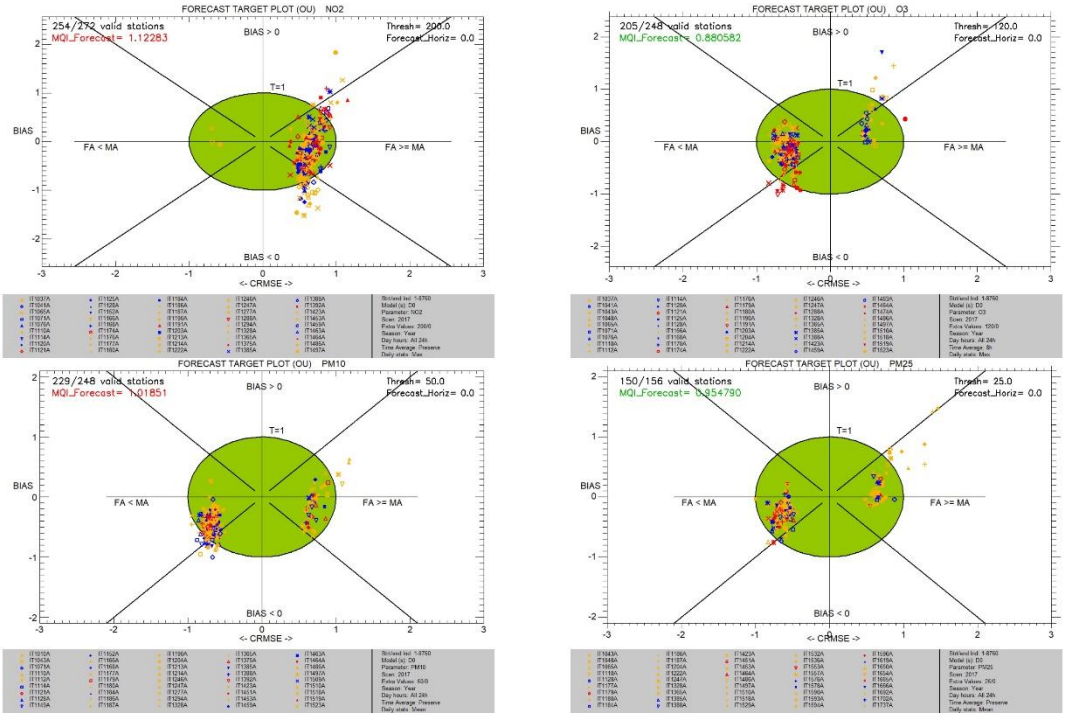


Figure 1. FORAIR-IT skills in forecasting 2017 year: Forecast Target Plots for NO_2 daily maximum (upper left), O_3 daily maximum of 8-hour average (upper right), PM10 daily mean (lower left) and PM2.5 daily mean (lower right) concentrations

As an example, Figure 1 shows, for all pollutans included within the methodology, the outcomes of the evaluation of one year-long model data set produced by the Italian air quality forecast system FORAIR-IT (Adani et al., 2022; https://impatti.sostenibilita.enea.it/en/research/activity/5479) against observations from background stations, the number of which is reported above the 90th percentile MQI_f. Results indicate a good level of quality of FORAIT-IT in simulating O_3 and PM2.5, and some room for improvement concerning NO_2 and PM10 (90th percentile MQI_f slightly higher than 1).

Additional Modelling Performance Indicators (*MPIs*) are defined based on the Mean Fractional Error (*MFE*), a normalized statistical indicator widely used in litterature (e.g. Boylan and Russell, 2006). Two different *MPIs* are defined as follows: 1) comparing the forecast model performances with the persistence model ones ($MPI_1 = MFE_f/MFE_p$); 2) evaluating forecast skills regardless of persistence aspects, using an acceptability threshold based on measurement uncertainty ($MPI_2 = MFE_f/MF_U$), where MF_U is the Mean Fractional Uncertainty, defined as follow

$$MF_U = \frac{1}{N}\sum_{i=1}^{N}\frac{2U(O_i)}{O_i} \quad (3)$$

For both *MPIs*, Modelling Performance Criteria (*MPC*) are defined, being fulfilled when *MPIs* are less or equal to 1.

MPIs based on *MFE* help in interpreting the outcomes. First of all, being *MFE* a normalized error, it does not depend on the magnitude of the absolute concentration values. Moreover, MPI_2 is formulated regardless of persistence aspects, providing, as an added value, an evaluation of the model performances quality itself.

As an example, Figure 2 shows how FORAIR-IT performances in simulating O_3 vary along with the forecast horizon. According to Forecast Target Plot outcomes (above), modelling performances get better from D0 (today forecast) to D2 (the day after tomorrow). MPI Plots (below) help to clarify that this unrealistic improvement is actually due to persistence model performances degradation. Indeed, forecast model performances get better along Y axis, where they are normalized to persistence model skills, but they slighty deteriorate along X axis, where they are considered regardless of persistence aspects. The green (orange) area indicates the fulfillment of MPC for both (one out of the two) MPIs.

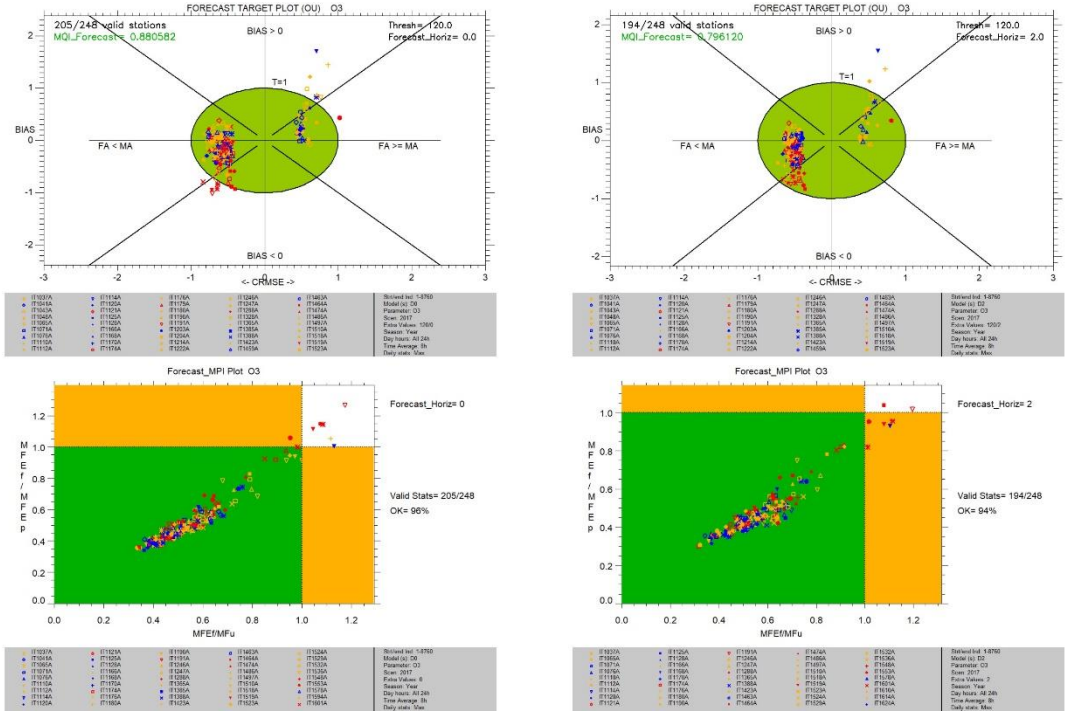
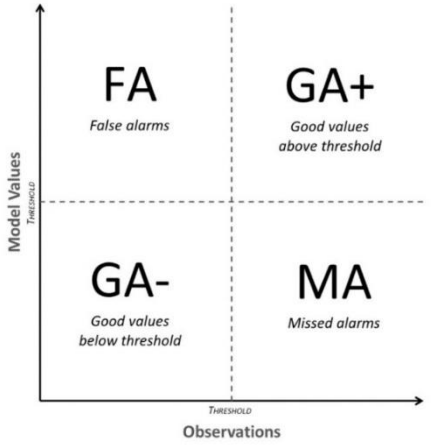


Figure 2. FORAIR-IT performances in forecasting O_3 along with the forecast horizon: skills variation from D0 (left plots) to D2 (right plots), according to Forecast Target Plot (upper plots) and Forecast MPI Plot (lower plots)

Assessment of model capability in predicting Threshold Exceedances

In addition to accurately reproducing sudden concentration changes, exceedances of specific thresholds levels (like limit values for daily concentrations) should be correctly estimated by a forecast model in order to support short-term action plans. Some commonly used threshold indicators (as defined in the right side of Figure 3) are included in the proposed validation approach, based on the 2x2 contingency table (Figure 3, left) representing the joint distribution of categorical events (below or above the threshold value) predicted by the model and observed by measurements.



Indicator	Acronym
Accuracy = (GA_+ + GA_-) / Total	ACC
Success Ratio = GA_+ / (FA + GA_+)	SR
Probability of Detection = GA_+ /(MA + GA_+)	PD
FBias score = (GA_+ + FA) / (MA + GA_+)	FB
Threat Score = GA_+ / (MA + FA + GA_+)	TS
Gilbert Skill Score= (GA_+ - H) / (MA + FA + GA_+ - H) with H = (GA_+ + MA)(GA_+ + FA) / Total	GSS

Figure 3. Left: contingency table and definition of the threshold exceedance quantities GA_+, GA_-, FA and MA. Right: Threshold exceedances indicators (definitions and acronyms)

The statistical distribution of all the quantities and indicators defined in Figure 3 are summarized in the Forecast Summary Report. Figure 4 shows an example of FORAIR-IT skills in predicting O_3 daily maximum of 8-hour average and PM10 daily mean, for which a daily limit value is set by AAQD.

A good performance level is reached for the Accuracy, i.e. the indicator measuring the global skills in predicting good categorical answers (below or above). Few False Alarms are predicted, conversely more Missed Alarms are observed consistently with the overall model underestimation (Figure 1).

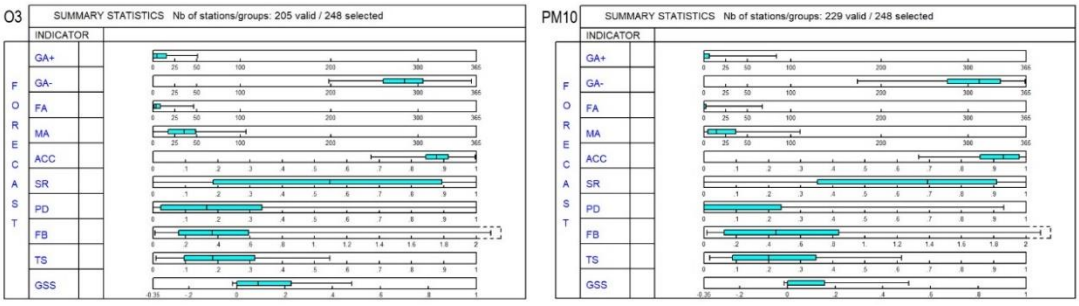


Figure 4. FORAIR-IT Forecast Summary Report for O_3 daily maximum of 8-hour average (left) and PM10 daily mean (right) concentrations

Assessment of modelling capability in predicting Air Quality Indices

The indicators presented in Figure 3 are based on a single threshold value. A simple multiple thresholds assessment is included in the proposed approach, based on Air Quality Indices, i.e a classification of concentrations levels into air quality categories commonly used for air quality forecasting purposes. More in detail, the number of days predicted by the forecast model in each category is compared with the corresponding number of measured ones. Figure 5 shows an example of the evaluation of FORAIR-IT, based on the EEA Air Quality Index table (https://www.eea.europa.eu/themes/air/air-quality-index/index). NO_2 and PM2.5 outcomes are presented at six monitoring stations, located in different geografical area and emission environments (i.e. rural, suburban, urban areas). In the context of a prevalent good agreement, a

general underestimation is observed, i.e. model values populate higher level categories to a lesser extent than the measured ones.

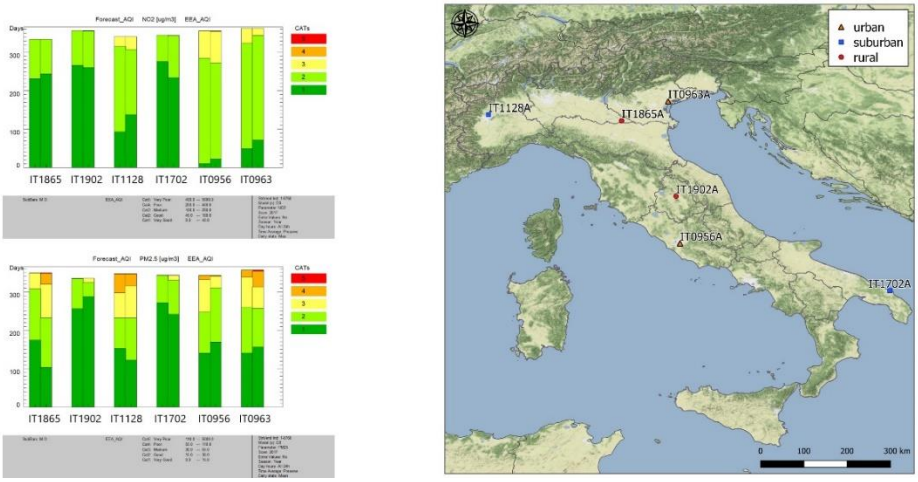


Figure 5. FORAIR-IT Forecast Air Quality Index Diagrams at six monitoring stations (right) for NO_2 daily maximum (upper left) and PM2.5 daily mean (lower left) concentrations

CONCLUSIONS

The FAIRMODE methodology for evaluating short-term forecasts of air quality, implemented in the DELTA Tool software, allows to detect 3 major capabilities which must be shown by a reliable forecast modelling system for a given application. One capability is to detect sudden changes of concentrations from day to day, which indicates that the model description is adequate to follow sharp changes of atmospheric variables. Another capability is to detect concentration threshold exceedances, which is the typical trigger of emergency measures applied by air quality managers for limiting emissions. The last capability is to reproduce multi-polluntant air quality indices, which are an effective way of presenting air quality to citizens. Both the methodology and the software are publicly available for testing and application, especially targeting European Member States and air quality forecasting services, like the Copernicus Atmospheric Monitoring Service.

REFERENCES

Adani, M., D'Isidoro, M., Mircea, M., Guarnieri, G., Vitali, L., D'Elia, I., Ciancarella, L., Gualtieri, M., Briganti, G., Cappelletti, A., Piersanti, A., Stracquadanio, M., Righini, G., Russo, F., Cremona, G., Villani, M.G., Zanini, G., 2022: Evaluation of air quality forecasting system FORAIR-IT over Europe and Italy at high resolution for year 2017. *Atmospheric Pollution Research*, **13**, 101456.

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

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.

Mittermaier, M. P., 2008: The Potential Impact of Using Persistence as a Reference Forecast on Perceived Forecast Skill. *Weather and Forecasting*, **23**, 1022-1031.

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**, 373–383.

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

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