https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

A new set of indicators for model evaluation complementing to FAIRMODE's MQO

Alexander de Meij[1], Cornelis Cuvelier[2♦], Philippe Thunis[2], Enrico Pisoni[2]

1MetClim, Varese, 21025, Italy
2 European Commission, Joint Research Centre (JRC), 21027, Ispra, Italy
♦retired with Active Senior Agreement

Correspondence to: Philippe Thunis (philippe.thunis@ec.europa.eu)

Abstract. In this study, we assess the relevance and utility of several performance indicators developed within the FAIRMODE framework by evaluating eight CAMS models and their ensemble in calculating concentrations of key air pollutants, specifically NO2, PM2.5, PM10, and O3. The models' outputs were compared with observations that were not assimilated into the models. For NO2, the results highlight difficulties in accurately modelling concentrations at traffic stations, with improved performance when these stations are excluded. While all models meet the established criteria for PM2.5, indicators such as bias and Winter-Summer gradients reveal underlying issues in air quality modelling, questioning the stringency of the current criteria for PM2.5. For PM10, the combination of MQI, bias, and spatial-temporal gradient indicators prove most effective in identifying model weaknesses, suggesting possible areas of improvement. O3 evaluation shows that temporal correlation and seasonal gradients are useful in assessing model performance. Overall, the indicators provide valuable insights into model limitations, yet there is a need to reconsider the strictness of some indicators for certain pollutants.

1. Introduction

Air Chemistry Transport Models (ACTMs) are used to calculate the complex physical and chemical processes that play a role in the formation and removal of gases and aerosols (e.g. NO2, O3, SOx, PM) from our atmosphere. Also, an ACTM is an instrument to assess the effects of future changes in aerosol (+ precursor) emissions, and models are therefore used to assist policy making in the design of effective reduction strategies to improve the air quality.

An air quality model requires a set of input data (e.g. emission and meteorology) and a description of (dynamical and chemical) processes to calculate gas and aerosol pollutants. The description of these processes in the model is associated with uncertainties. This may lead to large uncertainties in the estimated lifetimes of gases and aerosols in the atmosphere and the resulting air pollutant concentrations. Over the years, air quality modelling has improved as model's uncertainties have been reduced. Often classical statistical parameters are used to evaluate the air quality model's capability in calculating air pollutants. For example, bias (measure of overestimation or underestimation), standard deviation (a measure of the dispersion of the observed/calculated values around the mean), temporal correlation coefficient (linear relationship between model and observations), root mean square error (a measure of difference between the model and the observations; measure of accuracy) to name a few.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

These indicators are in general used to assess the model's performance against measurements. However, these indicators do not tell whether model results have reached a sufficient level of quality for a given application. For this reason, the Forum for Air quality Modelling (FAIRMODE) (https://fairmode.jrc.ec.europa.eu/home/index) developed several specific quality assurance and quality control (QA/QC) indicators and associated a threshold to each of them, that indicates the minimum level of quality to be reached by a model for policy use (Janssen and Thunis, 2022). Recent studies that have used these QA/QC indicators and associated thresholds to evaluate air quality model's performances are Kushta et al., (2018) and Thunis et al., (2021).

The goal of this study is to assess the relevance and usefulness of FAIRMODE's model quality assessment indicators and FAIRMODE's QA/QC Tools, by using as benchmark the Copernicus Atmospheric Monitoring Service (CAMS) air quality modelling and ensemble results over Europe.

More details on the model, methodology and emission inventories are given in Chapter 2. Followed by the analysis of the results in Chapter 3. In Chapter 4 the conclusions are provided.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

2. Methodology

CAMS produces annual air quality (interim) re-analysis for the European domain at a spatial resolution of 0.1 x 0.1 degrees (approx. 10km). A median ensemble is calculated from individual outputs, since ensemble products yield on average better performance than the individual model products. The spread between the eight models can be used to provide an estimate of the analysis uncertainty.

We assess the relevance and usefulness of FAIRMODE's model quality assessment indicators by means of evaluating calculated air pollutants (NO2, O3, PM2.5 and PM10) by the eight CAMS models for the year 2021, by comparing with observational data from the European Air quality database and assess the results against the indicator thresholds. The evaluation of the model's performance is based on the comparison with observations that are not used to assimilate simulated calculated concentrations. The eight CAMS models are:

CHIMERE (FR), DEHM (DK), EMEP (NO), FMIA-SILAM (FI), GEMAQ (PL), KNMA-LOTUS-EUROS (NL), MFM-MOCAGE (FR), RIU-EURAD-IM (DE) and Ensemble (ENSKCa). More details on the models are described in (https://confluence.ecmwf.int/display/CKB/Dataset+documentation). The data can be downloaded here: https://atmosphere.copernicus.eu/data

For the statistical analysis, the FAIRMODEs' benchmarking methodology is applied, that provides many different statistical parameters, which are described in FAIRMODE's Guidance document (Janssen et al., 2022)

In this work we focus on the following statistical parameters:

The Modelling Quality Indicator (MQI) is a statistical indicator of the accuracy of a specific modelling application 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:

$$MQI(i) = \frac{|O_i - M_i|}{\beta U(O_i)} \quad (1)$$

Where U(Oi) is the measurement uncertainty and β a coefficient of proportionality. The normalisation of the bias by the measurement uncertainty is motivated by the fact that both model and measurements are uncertain. We want to account for the fact that when measurement uncertainty is large, some flexibility on the model performance can be accepted, translating in accepting larger model-observed errors. With a current value of 2 proposed for β, the quality of a modelling application is said to be sufficient when the model-observation bias is less than twice the measurement uncertainty.

Applied to a complete time series, Equation (1) can be generalized to:

$$MQI = \frac{RMSE}{\beta RMS_U} \quad (2)$$

With this formulation, the RMSE between observed and modelled values (numerator) is compared to the root mean square sum of the measurement uncertainties (RMSU) which value is representative of the maximum allowed measurement uncertainty (denominator).

For yearly averaged pollutant concentrations, the MQI formula is adapted so that the mean bias between modelled and measured concentrations is normalised by the uncertainty of the mean measured concentration:

$$MQI = \frac{|\overline{O} - \overline{M}|}{\beta U(\overline{O})} \quad (3)$$

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

More details on formulation (1), (2) and (3) can be found in the MQO guidance document (Janssen et al., (2022).

For the statistical analysis of the four air pollutants, we use for NO2 the hourly values and for O3 the 8-hour running mean maximum values. Whilst for PM2.5 and PM10 the daily averages are used. These different time intervals are in compliance with the EU air quality standards as stated in the Directive 2008/50/EC (http://data.europa.eu/eli/dir/2008/50/2015-09-18). The time intervals are specific for each air pollutant, because the observed health impacts associated with the various pollutants occur over different exposure times.

The Modelling Quality Objective (MQO) is fulfilled when the MQI is less or equal to 1., for at least 90% of the available stations. The yearly MQI is in general more challenging to fulfil than the daily MQI (but this is not a rule), because of the smallest measurement uncertainties for yearly mean observed concentrations. The underlying reason for this is that the impact of random noise and periodic re-calibration on the daily observations lead to larger uncertainties, which are compensated for yearly averages.

The main drawback of the MQOs is that they provide a single summary pass/fail information for a modelling application. This simple test does not prevent a modelling application to pass for the wrong reason under certain circumstances. In addition, it does not provide any information on the capability of the model to reproduce hot spot areas (spatial variability) or on the timing of the pollution peaks (temporal variability).

For these reasons, additional indicators are proposed to assess the capacity of models to capture the temporal and spatial variability of the measurements. These indicators are based on temporal and spatial correlation and standard deviations that are normalised by the measurement uncertainty.

These indicators are constructed as follows:

For hourly frequency model output, values are first yearly averaged at each station. A temporal or spatial correlation and standard deviation indicator are then calculated for this set of values. The two indicators are normalised by the measurement uncertainty of the average concentrations:

$$RMS_{\bar{U}} = \sqrt{\frac{1}{N}\sum U(\bar{O})^2} \qquad (4)$$

The same approach applies for yearly frequency output.

These indicators are defined as:

	Model Performance Indicator (MPI)	**Model Performance Criteria (MPC)**
Correlation (5)	$MPI = \frac{1-R}{0.5\beta^2 \frac{RMS_{\bar{U}}{}^2}{\sigma_O \sigma_M}}$	$MPC: MPI \leq 1$
Standard deviation (6)	$MPI = \frac{\lvert\sigma_M - \sigma_O\rvert}{\beta RMS_{\bar{U}}}$	

Where the Model performance criteria is the criteria to be fulfilled in order to reach the quality objective of the modelling application.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

On top of these already agreed indicators included in FAIRMODE MQI system approach, we propose to complement them with incremental indicators, where relevant[1], to assess how concentration gradients between rural and urban or between traffic and urban stations are reproduced by the model. This is relevant in the context of the Ambient Air Quality Directive (AAQD), because the design of the monitoring network aims to capture existing gradients and differences occurring as a result of different pollution sources and different dispersion situations. These additional spatial indicators can be constructed similarly to other MQIs, i.e. normalised by the measurement uncertainty.

For example, the modelled incremental change between rural background (RB) and urban background (UB) locations is defined as:

$$INC_{UB-RB}^{model} = \bar{M}_{UB} - \bar{M}_{RB} \quad (7)$$

where M is the model value and similarly for the measured increment:

$$INC_{UB-RB}^{observed} = \bar{O}_{UB} - \bar{O}_{RB} \quad (8)$$

These indicators are then normalised by the measurement uncertainty.

	Model Performance Indicator (MPI)	**Model Performance Criteria (MPC)**
UB – RB (9)	$\mathrm{MPI} = 1/\beta * \frac{INC_{UB-RB}^{model} - INC_{UB-RB}^{observed}}{0.5 * (RMS_{U(UB)} + RMS_{U(RB)})}$	MPC: MPI ≤ 1
UT – UB (10)	$\mathrm{MPI} = 1/\beta * \frac{INC_{UB-UT}^{model} - INC_{UB-UT}^{observed}}{0.5 * (RMS_{U(UB)} + RMS_{U(UT)})}$	

where UT stands for “urban traffic”.

As mentioned earlier, the MQO generally applies to the average of a specific period, currently, one year. Consequently, it provides no information whether the modelling application manages to capture the temporal variability of the air quality situation. Since the AAQDs include also in the assessment the evaluation of exceedances for specific temporal indicators, the capability of the modelling application to reproduce the temporal variations becomes highly relevant in the context of air quality management.

For that reason, additional indicators to assess the temporal coherence of model results, at different frequencies are provided. These include seasonal, week/week-end or day/night indicators. Measurement and modelling results are then aggregated (all stations belonging to a certain type (urban – rural –traffic – industrial) together and checks are made through the following indicators:

		Model Performance Indicator (MPI)	**Model Perf. Criteria (MPC)**

[1] Indicators can only be applied with models that are designed to simulate the station types that are used in the indicators (e.g. urban-traffic incremental indicators cannot be applied to models that only simulate background levels).

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

Seasonal (12)	Industry	$\text{MPI} = \frac{SeasDiff_{Ind}^{mod} - SeasDiff_{Ind}^{obs}}{\beta RMS_{\bar{U}}}$	MPC: MPI ≤ 1
	Traffic	$\text{MPI} = \frac{SeasDiff_{traffic}^{mod} - SeasDiff_{traffic}^{obs}}{\beta RMS_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{SeasDiff_{bg}^{mod} - SeasDiff_{bg}^{obs}}{\beta RMS_{\bar{U}}}$	
Week / weekend (13)	Industry	$\text{MPI} = \frac{WeekDiff_{Ind}^{mod} - WeekDiff_{Ind}^{obs}}{\beta RMS_{\bar{U}}}$	
	Traffic	$\text{MPI} = \frac{WeekDiff_{traffic}^{mod} - WeekDiff_{traffic}^{obs}}{\beta RMS_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{WeekDiff_{bg}^{mod} - WeekDiff_{bg}^{obs}}{\beta RMS_{\bar{U}}}$	
Day/night (14)	Industry	$\text{MPI} = \frac{DayDiff_{Ind}^{mod} - DayDiff_{Ind}^{obs}}{\beta RMS_{\bar{U}}}$	
	Traffic	$\text{MPI} = \frac{DayDiff_{traffic}^{mod} - DayDiff_{traffic}^{obs}}{\beta RMS_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{DayDiff_{bg}^{mod} - DayDiff_{bg}^{obs}}{\beta RMS_{\bar{U}}}$	

155

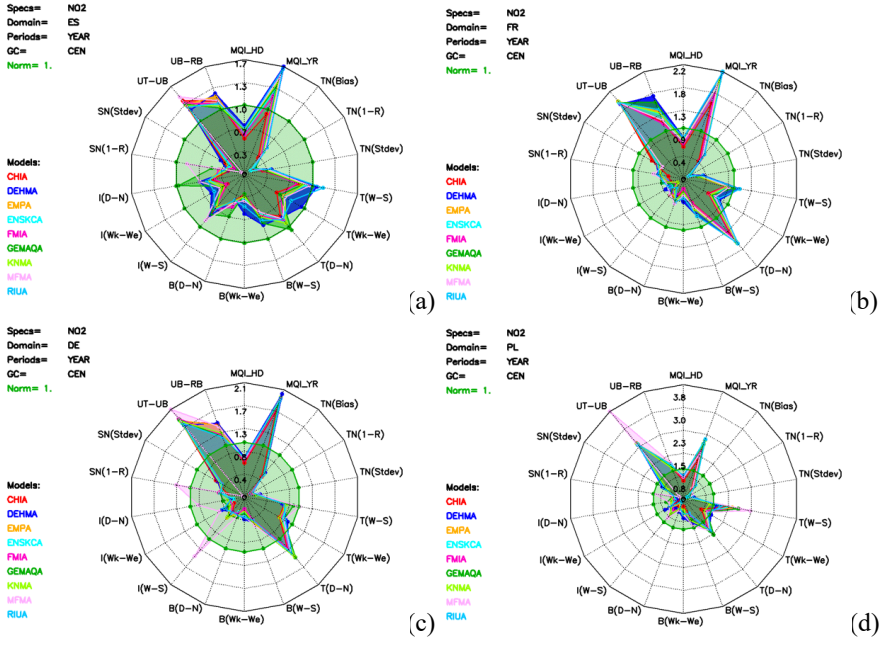
https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

3. Results

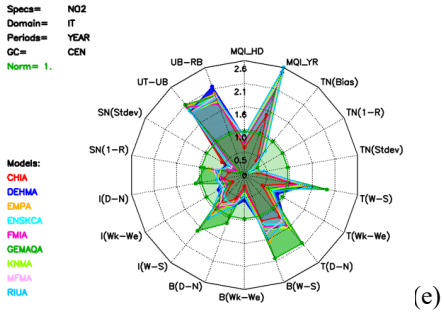
To best visualize all these indicators, we use a graphical representation in terms of radar plots. These plots help to assess the relevance and usefulness of the different statistical indicators by comparing all of them in a single diagram. We use this approach to assess models' performance for Spain, France, Germany, Poland and Italy. This allows us to see if (1) the MQI values fulfil the MQO. If this is not the case, the radar plots help to understand which of the other indicators are useful in determining the model's skill through analysing (2) the temporal and spatial indicators (1-R and Stdev), followed by (3) studying the models' capability in calculating the temporal variability i.e. seasonal (Winter-Summer [W-S]), week-weekend (Wk-We) and day-night (D-N) indicators and spatial indicators (e.g. urban background - rural background gradient).

3.1 NO2

In Fig. 1, the statistics for NO2 are shown for (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy by all models considering all stations (i.e. background (B), urban, traffic (T), industry (I)). The green circle represents the reference line, that is MQI is 1.0. Results for any statistical parameter that fall within the circle indicates that the MQO is achieved. Anything that falls outside the green circle indicates a poor agreement of the model results when compared to observations. The cyan solid contour in each radar plot represents the Ensemble Median. The other air quality models are presented by different colours.



https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025



(e)

Figure 1. Radar plots of the calculated air quality model indicators for NO2 for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are: MQI Hourly (MQI_HD), MQI Year (MQI_YR), Bias, 1-R (Time), Standard deviation (Time), gradients for Winter-Summer, Week-Weekend, Day-Night for Traffic, Industry, Background (T, I, B), 1-R Spatial, Standard Deviation spatial, Yearly Urban-Traffic vs Urban-Background (Year UT-UB), Yearly Urban-Background vs Rural-Background (Year UB-RB).

Fig. 1 shows that the yearly MQIs (MQI_YR) are generally higher than 1.5 for all models and all countries (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy, indicating that the MQOs are not achieved, while the short-term MQIs (MQI_HD) fulfil the MQOs. As mentioned earlier, the yearly MQI is more difficult to fulfil than the daily MQI, because of smaller measurement uncertainties for yearly mean observed concentrations. As a consequence, the MQI_YRs values are higher than MQI_HD, indicating that each model has difficulties capturing well the observed yearly concentrations for NO2.

As mentioned earlier, the MQOs tells if the model fails or passes the MQI, but with limited information on the model's capability to calculate the temporal and spatial variability of the air pollutant concentrations. This is why we introduced additional indicators, see (Equations 4 – 6), which present the bias and temporal- and spatial correlation.

A more stringent source of information to the additional indicators in Equations 4 – 6 are presented in Equations 7 - 10. We see that for example these indicators describe the differences between biases for Day versus Night values for Background [B(D-N] and Industry [I(D-N] stations are smaller than 1.0, except for Italy by GEMAQA (see Annex). Therefore, one would expect that the models are, in general, capable of calculating well the NO2 concentrations. But when the spatial indicators are considered, this is clearly not the case. For example, the spatial concentration gradient around a Traffic station considering the Urban Background stations (Year UT-UB) and UB-RB (concentration gradient around a Background station considering Rural Background stations), exceeds the reference line (1.0) indicating that the model's capability in calculating the spatial gradient is poor when compared to the observations and therefore doesn't fulfil the MQO.

This can be explained by the fact that the model resolution (0.1 x 0.1) is too coarse to capture the emissions from the road transport sector. This is illustrated in Fig. 2, which shows the difference between observations and calculated yearly mean NO2 concentrations for Traffic, Industry, All and Background stations for Germany. The calculated NO2 concentrations for Traffic and All stations remain flat, i.e. the concentrations are very similar around 13 µg/m3. While the difference in observed concentrations (grey bar) between Traffic stations and All stations is around 7 µg/m3 (27 for Traffic and 20 µg/m3 for All stations).

Also, the Bias for Traffic stations is much larger (up to -14 µg/m3), while the Bias for all stations is smaller (up to -9 µg/m3), see Fig. 3. This indicates that the models have difficulties in calculating the NO2 concentrations for Traffic

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

stations as mentioned earlier. Once again this is expected, given the resolution of the models, but it shows the relevance of the indicators and associated thresholds to detect it.

The mean calculated NO2 concentrations by the models for Industry and Background stations agrees well with the observations. This reflects into low bias for Industry and Background stations (< 3 µg/m3).

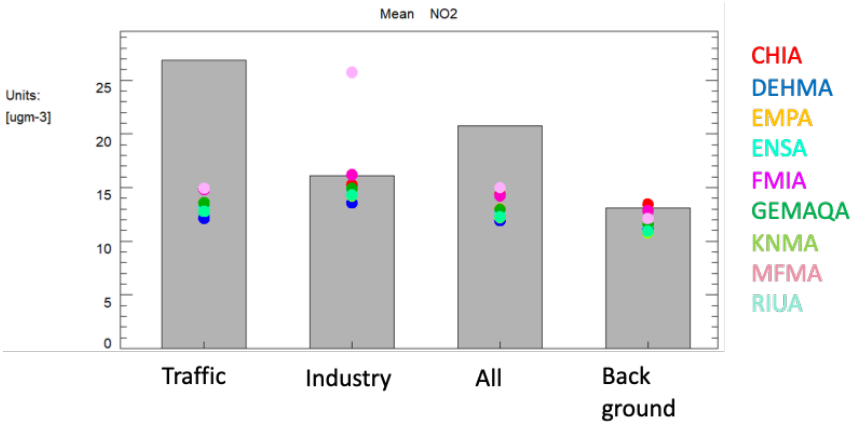


Figure 2. Yearly mean observed (grey bar) and calculated (coloured dots) NO2 concentrations for Germany for Traffic, Industry, All and Background stations.

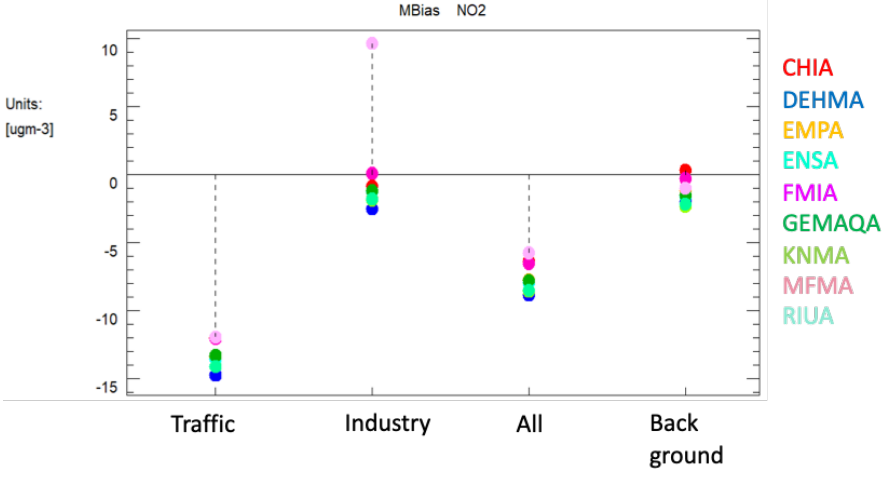


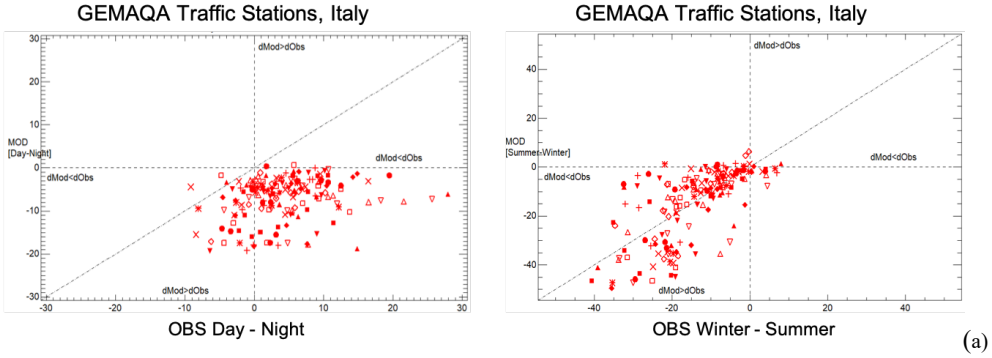
Figure 3. Yearly mean bias for NO2 for Traffic, Industry, All and Background stations for the different models (coloured dots) for Germany stations.

Looking in more details we show in Fig. 4 the comparison between the model versus Day - Night and Winter - Summer mean observations for Traffic and Background stations in Italy. Well behaving results should lie along the 1 to 1 line. Results located in the lower right and upper left parts of the graphs are poor.

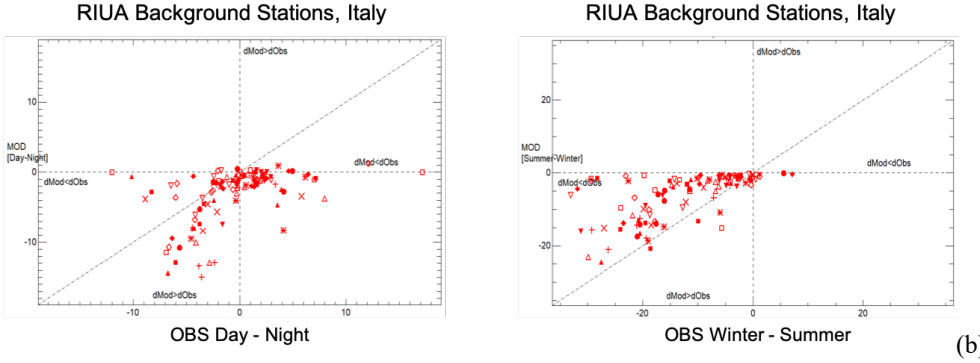
Like the other models, GEMAQA (Fig. 4a) shows a poor agreement for the traffic stations to capture the Day - Night and Winter - Summer profiles for Italy. A similar behaviour is found for the Background stations as shown in Fig. 4b for

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

RIUA. Note that for the other countries the Day - Night and Winter - Summer profiles are satisfactory for Background stations, but not for Traffic stations.



(a)



(b)

Figure 4. NO2 scatter plots of modelled versus observed day-night and summer-winter mean differences for traffic stations by (a) GEMAQA and background stations by (b) RIUA model.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

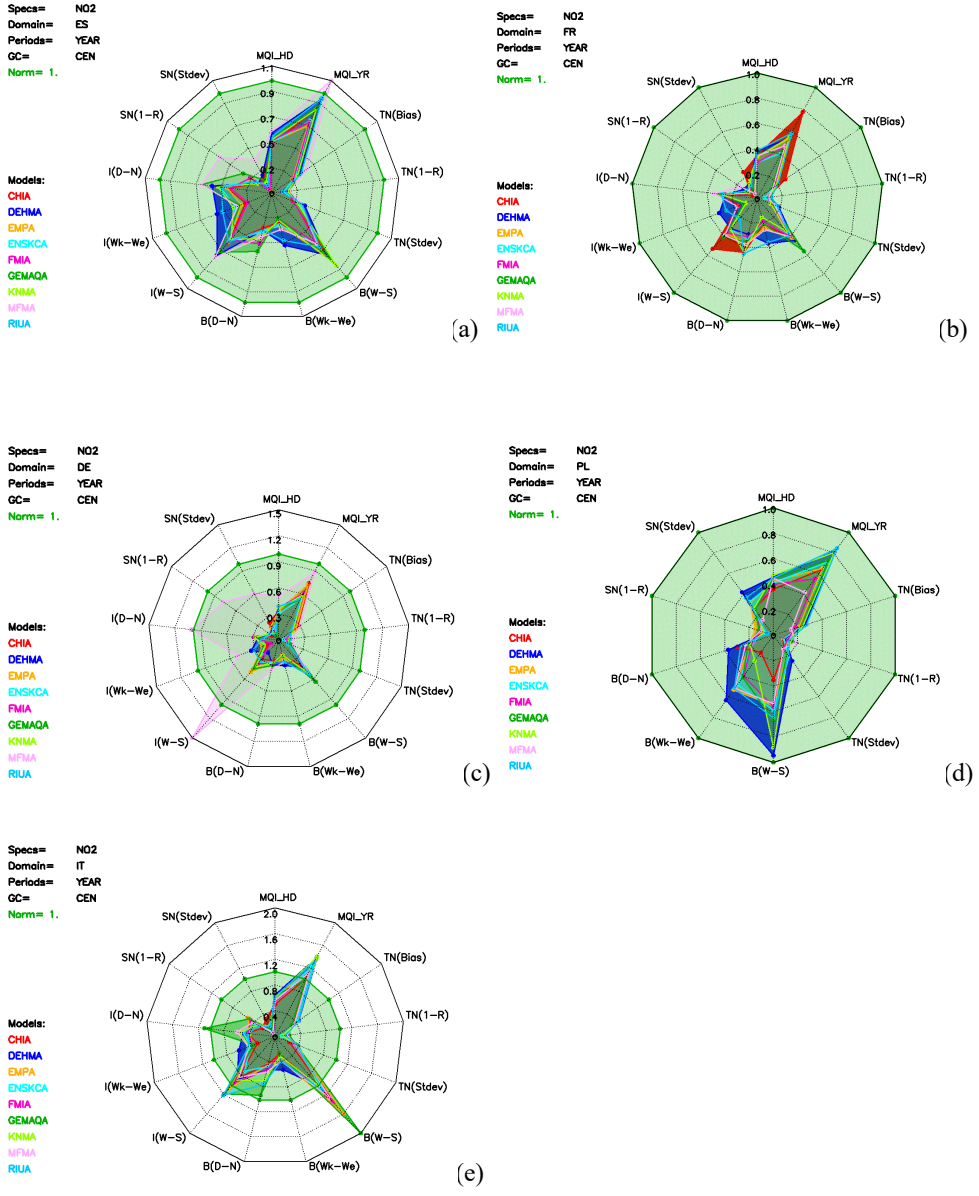


Fig. 5 Radar plots of the calculated air quality model indicators for NO2 for different countries excluding the Traffic stations: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy.

When Traffic stations are excluded from the analysis (Fig. 5), we see that the yearly MQI are much lower for the five countries and even fulfil the MQO for France, Germany and Poland.

This confirms that the models have difficulties in calculating the NO2 concentrations for Traffic stations. The reason for this is that the model resolution is not fine enough to capture the traffic emissions and as a result the short lifetime of NO2 (about one hour) and consequently the non-linear production and loss of NO2 concentrations.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

As indicated, this result was expected and demonstrates that the level of stringency of the QA/QC indicators is relevant. Apart this expected result for traffic stations, these indicators also flag some aspects that need to be improved for NO2, such as the spatial concentration gradient.

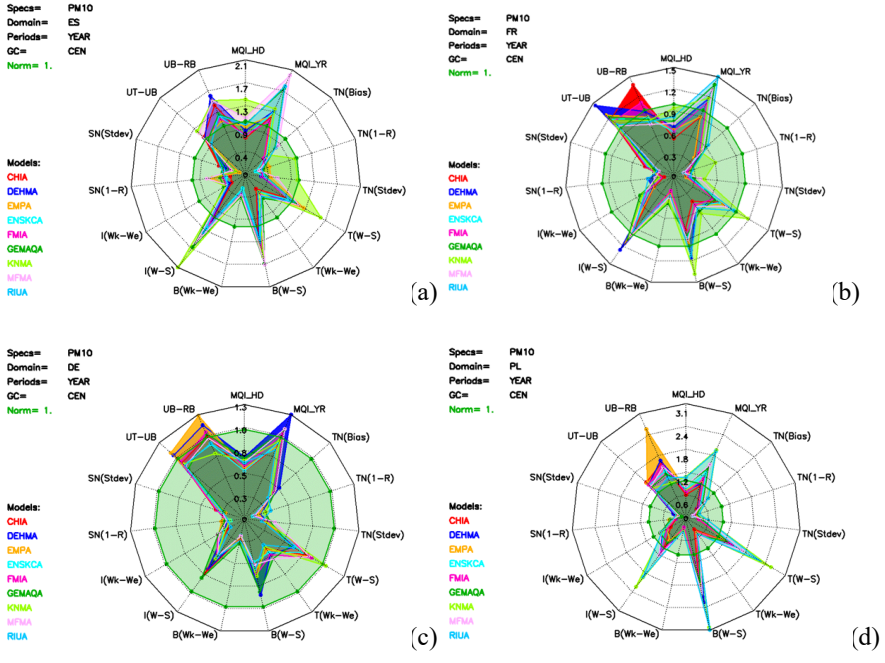
All the results of the statistical analysis for NO2 (and other air pollutants) are provided in Table S1 of the Supplement material.

3.2 PM10

The MQI_YRs for PM10 concentrations are higher than the MQI_HDs (Fig. 6), which can be explained by the smaller measurement uncertainties for yearly PM10 observations as mentioned before. For Germany the Ensemble MQI_YR is close to unity, i.e. 1.00 (± 0.14).

Looking at the different statistical indicators in the radar plots, we see that all the models show similar shapes in the radar plots, indicating that the models show the same strengths and weaknesses. The temporal correlation coefficient (1-R) and standard deviation for all the models and countries are lower than 1.0. This means that the models are good for these indicators or that the level of stringency is too low. This implies that that other indicators are required to perform a more stringent evaluation of the air quality model.

The radar plots show that the models have in general difficulties in calculating the spatial profiles (Year UT-UB, UB-RB) and temporal profiles (Winter - Summer gradient for Traffic, Background and Industry) for Spain, France, Poland and Italy. While for Germany all indicators are below unity by the different models, apart from UT-UB and UB-RB by DEHMa and EMPa, and MQI_YRs by DEHMa, GEMAQa and MFMa.



https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

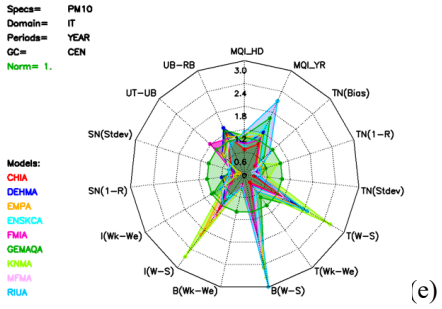


Figure 6. Radar plots of the calculated air quality model indicators for PM10 for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are: MQI Hourly (MQI_HD), MQI Year (MQI_YR), Bias, 1-R (Time), Standard deviation (Time), gradients for Winter-Summer, Week-Weekend, Day-Night for Traffic, Industry, Background (T, I, B), 1-R Spatial, Standard Deviation spatial, Yearly Urban-Traffic vs Urban-Background (Year UT-UB), Yearly Urban-Background vs Rural-Background (Year UB-RB).

The poor skill for Spain and Poland is illustrated in Fig. 7, which shows the large differences between the models in calculating the average PM10 concentrations for the different station types. Only DEHMA shows a small positive bias (~1 ug/m3) for all the station types for Spain, while most of the models underestimate on average the observed PM10 concentrations.

For Poland, all the models underestimate the observed PM10 concentrations for the different station types (Fig.7). The highest PM10 concentrations are observed for Traffic stations for Poland. It is for these stations that the models' capability in calculating elevated PM10 concentrations for Traffic stations is poor, which is shown in the largest bias found for these stations. Excluding the traffic stations from the comparison results in an MQI of 0.99, while with traffic stations MQI is 1.32.

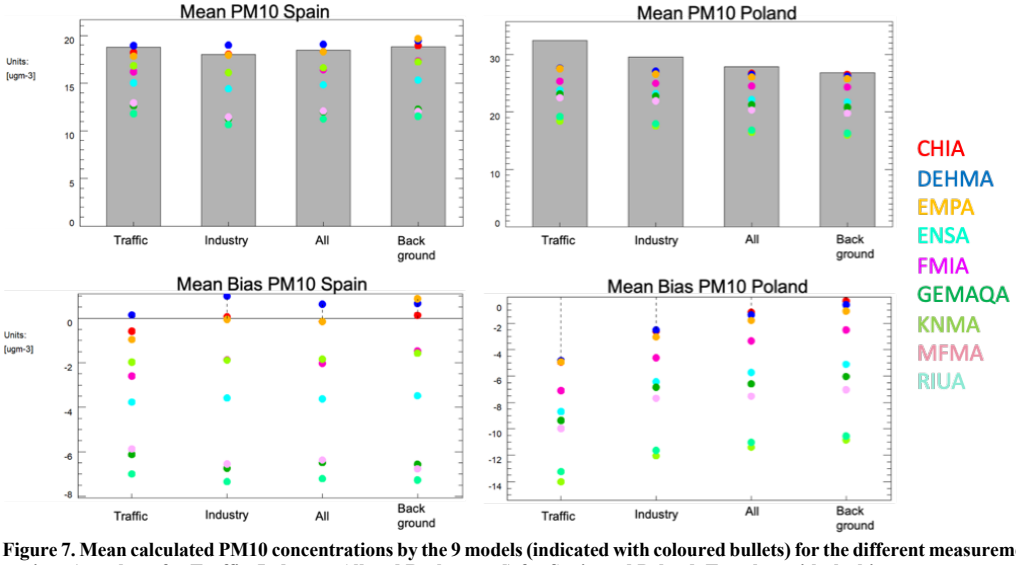


Figure 7. Mean calculated PM10 concentrations by the 9 models (indicated with coloured bullets) for the different measurement stations (grey bars for Traffic, Industry, All and Background) for Spain and Poland. Together with the bias.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

The radar plots show that the Winter – Summer gradient are larger than 1.0 for the different countries. For that reason, we analyse in more details the PM10 concentrations for Poland during different seasons that will help to understand the reason for the higher bias for traffic. The mean bias during the summer period (Fig. 8, left panel) is the highest for Traffic stations (up to ~-10 ug/m3) with a small positive bias for a few models when All and Background stations are considered. For the winter period (right panel), the mean bias is a factor ~2 higher than for the summer, with RIUA and KNMA showing the highest bias (up to ~-20 ug/m3) for the four different station types. This indicates that the models underestimate the PM10 concentrations for the whole country, especially during winter time, even though the model concentrations are assimilated.

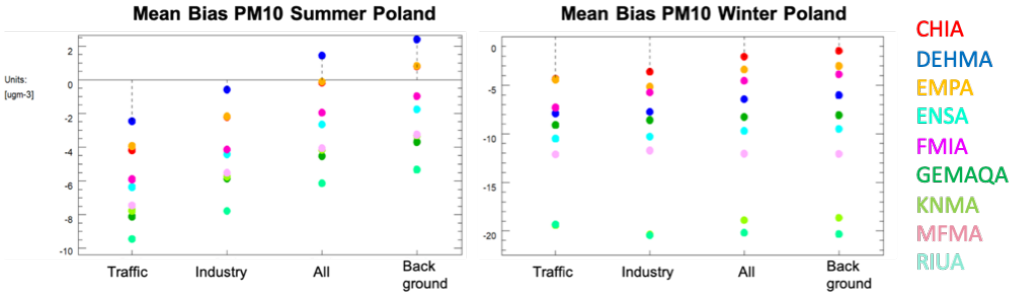


Figure 8. Mean Bias PM10 for Summer (JJA) and Winter (DJF) for Poland by all the models for the different station types (Traffic, Industry, All and Background).

When traffic stations are excluded in the analysis, it appears that only for Germany, Poland, and Italy the Ensemble's MQI_YR is lower (e.g. for Poland ~1.4 versus ~1.0 without traffic stations). As mentioned earlier the Winter – Summer profiles for Industry, Background (and to some extend traffic) stations hampers the overall model's performance in calculation the PM10 concentrations (indices are well above the reference criteria of 1.0). For example, the Winter-Summer gradients for Spain (Fig. 9) are scattered around the 1:1 line, while the Week-Weekend profiles are closer to the 1:1 line. The latter corroborates the indicator values below the criteria.

This tells us that in addition to the MQI, the bias and spatial gradient indicators are relevant and useful to highlight the potential model weaknesses in calculating PM10 concentrations.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

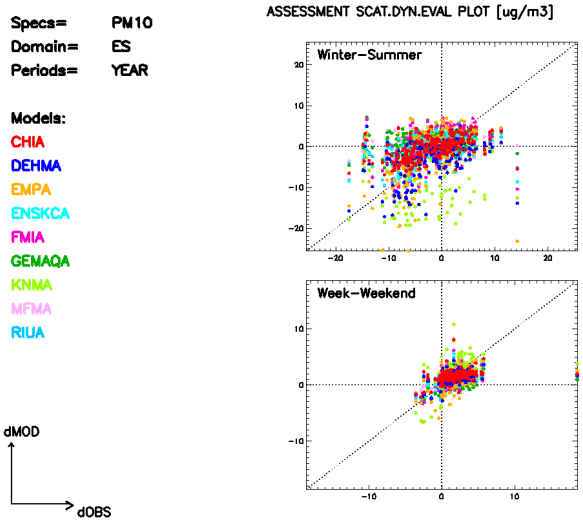


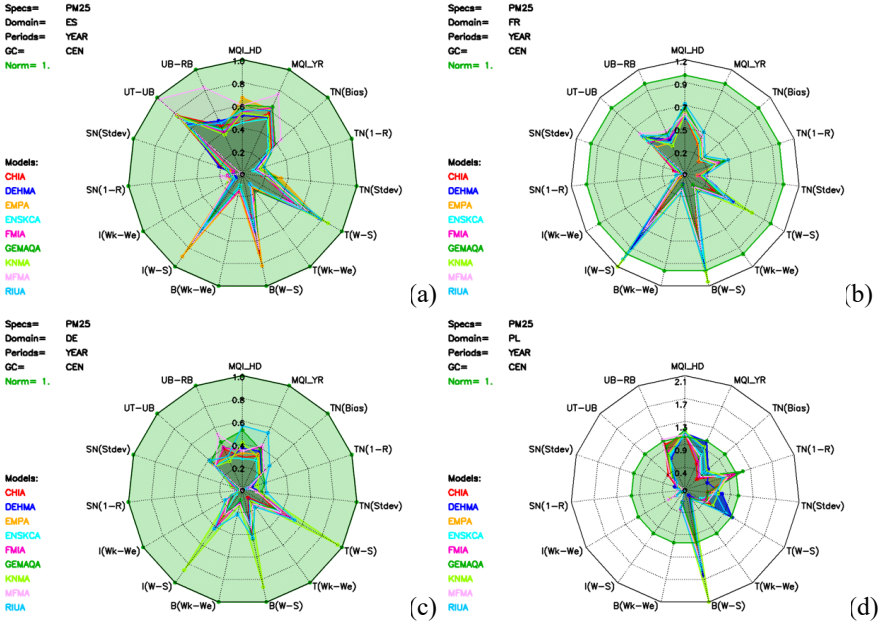
Figure 9 PM10 Scatter plots of modelled versus observed Winter-Summer and Week-Weekend mean differences for Spain for all the models.

340

3.3 PM2.5

Yearly MQIs for PM2.5 fulfil the MQOs for all models and countries. Also, the MQIs are in general lower than for PM10 (Fig. 10). This can be explained by the higher measurement uncertainty assumed for PM2.5 than for PM10 in the MQI Equations, allowing less stringency on the model results when calculating the MQI for PM2.5 (Thunis et al., 2021).

345



https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

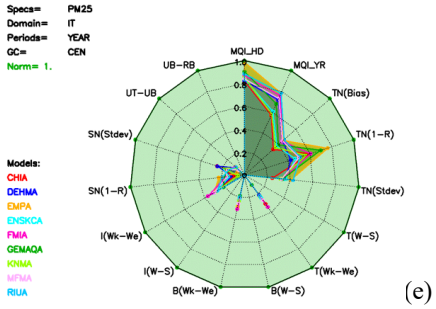


Figure 10. Radar plots of the calculated air quality model indicators for PM25 for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are: MQI Hourly (MQI_HD), MQI Year (MQI_YR), Bias, 1-R (Time), Standard deviation (Time), gradients for Winter-Summer, Week-Weekend, Day-Night for Traffic, Industry, Background (T, I, B), 1-R Spatial, Standard Deviation spatial, Yearly Urban-Traffic vs Urban-Background (Year UT-UB), Yearly Urban-Background vs Rural-Background (Year UB-RB).

For Poland where coal combustion in households is still an important contributor to PM (De Meij et al., 2024) larger biases are found for the winter period (up to -13 ug/m3) than for the summer (up to -3 ug/m3), see Fig. 11. Our analysis further showed that for PM2.5 Daily and Yearly MQI values for Poland are on average a factor ~2 higher during winter (1.23 and 1.02 respectively) than summer (0.60 and 0.48 respectively). The absence of condensables in the emission inventories (or possibly other seasonal dependent emissions, such as emissions released by forest fires) may lead to much higher biases during the peak season and as a consequence potentially result in higher daily than yearly MQI values.

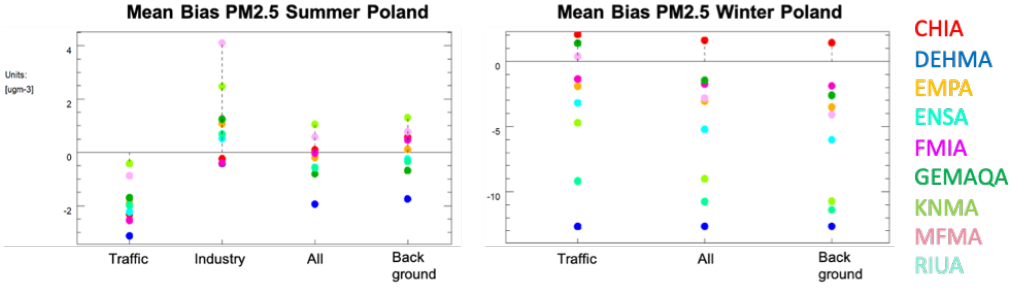


Figure 11. Mean Bias PM2.5 for Summer (JJA) and Winter (DJF) for Poland by the models for the different station types (Traffic, Industry, All and Background). Note that for Winter, there's only one Industry station, therefore the bias for this station type is not shown.

As we have seen before, considering only the MQI for the model evaluation doesn't provide enough information of the model's skill in calculating the temporal and spatial variability of the pollutant. The radar plots that include additional temporal and spatial indicators show that for Spain, France and Germany all the models show a similar behaviour, i.e. elevated values for the Winter – Summer indicators for Industry and Background, but still below unity. Just like for Poland, the Winter – Summer profiles for Background, Traffic and Industry stations are higher than 1.0 for DEHMA, KNMA and RIUA. While GEMAQA has difficulties in capturing the temporal correlation.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

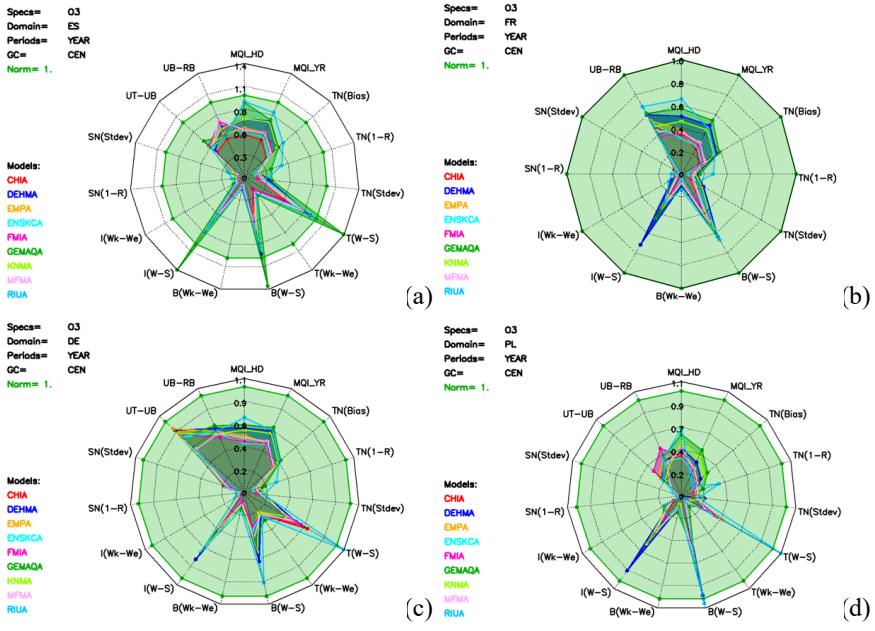
The bias and the Winter – Summer indicators reveal potential problems in air quality modelling for PM2.5 and for that reason are very useful.

3.4 O3

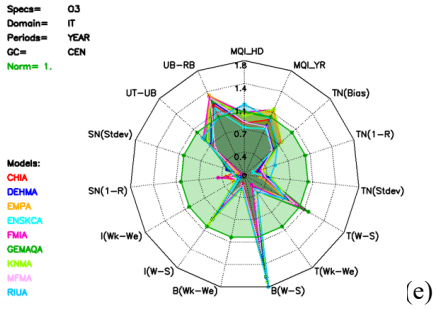
For O3, all indicators are lower than unity for France, indicating that the models capture well the 8-hour maximum O3 values (Fig. 12). Except by GEMAQA for Spain, i.e. the Winter-Summer Traffic, Background and Industry indicators are larger than 1.0. This is also true for the Winter-Summer Traffic indicator by RIUa.

Only for Poland, the RIUa model fails to capture the temporal profiles for Winter - Summer for the Traffic and Background stations. Looking in more details at the temporal correlation coefficient (R) for RIUa for all the available stations (35 stations in total), we see that R varies between 0.06 and 0.81 (on average R is 0.63), while for ENSKCa R varies between 0.42 and 0.98 (on average 0.90). This indicates that RIUa has more difficulties to capture the temporal profile for some stations when compared to the other models.

For Italy, MQI_YR is higher than 1.0 by EMPa, FMIa and RIUa, and all the models have difficulties to capture the temporal profile for Winter - Summer Background stations, i.e. the results are scattered around the 1:1 line (not shown). Also, the spatial gradients for UB-RB are higher than 1.0 by GEMAQa and EMPa.



https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025



(e)

Figure 12. Radar plots of the calculated air quality model indicators for 8-hour maximum O3 values for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are: MQI Hourly (MQI_HD), MQI Year (MQI_YR), Bias, 1-R (Time), Standard deviation (Time), gradients for Winter-Summer, Week-Weekend, Day-Night for Traffic, Industry, Background (T, I, B), 1-R Spatial, Standard Deviation spatial, Yearly Urban-Traffic vs Urban-Background (Year UT-UB), Yearly Urban-Background vs Rural-Background (Year UB-RB).

Even though the daily and yearly MQI for 8-hour maximum O3 values are in general below 1.0, the temporal correlation coefficient, together with the Winter-Summer gradients appear to be useful indicators to highlight potential problems for O3 concentrations modelling.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

4. Conclusion remarks

In this work, we examine the relevance and usefulness of assessment indicators within the FAIRMODE framework by evaluating the performance of eight CAMS models and their ensemble in calculating air pollutants. The evaluation is based on comparisons with observations that were not used to assimilate the modelled concentrations.

For nitrogen dioxide (NO2), we found that the yearly Model Quality Indicators (MQI), as well as the Winter-Summer and spatial gradient indicators, clearly show the challenges the models face in accurately calculating NO2 concentrations at traffic stations. This highlights the value of these indicators in assessing model performance. As expected, the exclusion of traffic stations from the analysis improves the models' performance, confirming that the indicators are effectively capturing the models' difficulties. For background stations, all indicator values fall below the threshold of 1.0, except for the GEMAQ model in Italy, suggesting better model performance in less complex environments.

When analysing fine particulate matter (PM2.5), we observed that the yearly and daily MQI for all models meet the established criteria. This, however, raises questions about the stringency of the indicators, as passing the criteria does not necessarily indicate flawless performance. Our analysis demonstrated that other indicators, such as bias and Winter-Summer gradients, are crucial for identifying the underlying issues in air quality modelling for PM2.5, making these indicators highly valuable.

For PM10, the yearly MQI, Winter-Summer indicators, and spatial gradients were not always met by the models. This suggests that, in addition to MQI, bias and both temporal and spatial gradient indicators are particularly important for identifying weaknesses in the models' ability to calculate PM10 concentrations. On the other hand, temporal correlation and standard deviation indicators seem to be less useful for evaluating model performance in this context.

Regarding ozone (O3), although the daily and yearly MQI for the 8-hour maximum O3 values generally fall below the threshold of 1.0, additional indicators such as the temporal correlation coefficient and Winter-Summer gradients prove useful for identifying potential model issues in calculating O3 concentrations.

Overall, the various indicators effectively served their purpose of revealing the specific limitations in the model applications, and assisting the modelling community in understanding where improvements are needed. However, there is ongoing debate about the appropriate level of stringency for certain indicators and pollutants, suggesting that there is room for refinement in the evaluation process.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

440

Data availability

The CAMS data is available from the Copernicus CAMS website, via https://atmosphere.copernicus.eu/data .

445

Author contribution

ADM performed the data analysis and wrote the draft of the manuscript. CC provided the research tool for the evaluation. CC and PT designed the study and helped with the data analysis. EP collected the data. All co-authors helped in editing suggestions to the manuscript.

450

Competing interests

The authors declare that they have no conflict of interest.

455

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

References

De Meij, A., Cuvelier, C., Thunis, P., Pisoni, E., and Bessagnet, B.: Sensitivity of air quality model responses to emission changes: comparison of results based on four EU inventories through FAIRMODE benchmarking methodology, Geosci. Model Dev., 17, 587–606, https://doi.org/10.5194/gmd-17-587-2024, 2024.

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, ISBN 978-92-76-52425-0, doi:10.2760/41988, JRC129254, 2022; https://fairmode.jrc.ec.europa.eu/document/fairmode/WG1/Guidance_MQO_Bench_vs3.3_20220519.pdf, last access: 26-01-2024.

Kuenen, J., Dellaert, S., Visschedijk, A., Jalkanen, J.-P., Super, I., Denier van der Gon, H.: CAMS-REG-v4: a state-of-the-art high-resolution European emission inventory for air quality modelling, Earth Syst. Sci. Data, 14, 491–515, https://doi.org/10.5194/essd-14-491-2022, 2022.

Kushta, J., Georgiou, G.K., Proestos, Y. et al. Evaluation of EU air quality standards through modeling and the FAIRMODE benchmarking methodology. Air Qual Atmos Health 12, 73–86 (2019). https://doi.org/10.1007/s11869-018-0631-z

Philippe Thunis, Monica Crippa, Cornelis Cuvelier, Diego Guizzardi, Alexander de Meij, Gabriel Oreggioni, Enrico Pisoni, Sensitivity of air quality modelling to different emission inventories: A case study over Europe, Atm. Env. X, Volume 10, 2021, 100111, ISSN 2590-1621, https://doi.org/10.1016/j.aeaoa.2021.100111.

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

Citing the models (https://confluence.ecmwf.int/display/CKB/CAMS+Regional%3A+European+air+quality+reanalyses+data+documentation)

For ENSEMBLE

Institut national de l'environnement industriel et des risques (Ineris), Aarhus University, Norwegian Meteorological Institute (MET Norway), Jülich Institut für Energie- und Klimaforschung (IEK), Institute of Environmental Protection – National Research Institute (IEP-NRI), Koninklijk Nederlands Meteorologisch Instituut (KNMI), METEO FRANCE, Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO), Swedish Meteorological and Hydrological Institute (SMHI), Finnish Meteorological Institute (FMI), Italian National Agency for New Technologies, Energy and Sustainable Economic Development (ENEA) and Barcelona Supercomputing Center (BSC) (2022): CAMS European air quality forecasts, ENSEMBLE data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For CHIMERE

Institut national de l'environnement industriel et des risques (Ineris) (2020): CAMS European air quality forecasts, CHIMERE model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For DEHM

Aarhus University (2020): CAMS European air quality forecasts, DEHM model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For EMEP

Norwegian Meteorological Institute (MET Norway) (2020): CAMS European air quality forecasts, EMEP model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For EURAD-IM

Jülich Institut für Energie- und Klimaforschung (IEK) (2020): CAMS European air quality forecasts, EURAD-IM model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For GEM-AQ

Institute of Environmental Protection – National Research Institute (IEP-NRI) (2020): CAMS European air quality forecasts, GEM-AQ model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS).

https://doi.org/10.5194/egusphere-2024-3690
Preprint. Discussion started: 3 January 2025

(Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For LOTOS-EUROS

Koninklijk Nederlands Meteorologisch Instituut (KNMI) and Nederlandse Organisatie voor toegepast-natuurwetenschappelijk onderzoek (TNO) (2020): CAMS European air quality forecasts, LOTOS-EUROS model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For MOCAGE

METEO-FRANCE (2020): CAMS European air quality forecasts, MOCAGE model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview

For SILAM

Finnish Meteorological Institute (FMI) (2020): CAMS European air quality forecasts, SILAM model data. Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store (ADS). (Accessed on <DD-MMM-YYYY>), https://ads.atmosphere.copernicus.eu/cdsapp#!/dataset/cams-europe-air-quality-reanalyses?tab=overview