



A new set of indicators for model evaluation complementing FAIRMODE's modelling quality objective (MQO)

Alexander de Meij¹, Cornelis Cuvelier^{2,☆}, Philippe Thunis², and Enrico Pisoni²

¹MetClim, 21025 Varese, Italy

²European Commission, Joint Research Centre (JRC), 21027 Ispra, Italy

☆retired with Active Senior Agreement

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

Received: 26 November 2024 – Discussion started: 3 January 2025

Revised: 9 April 2025 – Accepted: 10 April 2025 – Published: 15 July 2025

Abstract. In this study, we assess the relevance and utility of several performance indicators (model quality (bias) and model performance (temporal and spatial) indicators), developed within the FAIRMODE framework by evaluating eight Copernicus Atmospheric Monitoring Service (CAMS) models and their ensemble in calculating concentrations of key air pollutants, specifically NO₂, PM_{2.5}, 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 PM_{2.5}, indicators such as bias and winter–summer gradients reveal underlying issues in air quality modelling, questioning the stringency of the current criteria for PM_{2.5}. For PM₁₀, the combination of model quality indicators, 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.

(e.g. NO₂, O₃, SO_x, PM) from our atmosphere. Also, an ACTM is an instrument used 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 air quality.

An ACTM requires a set of input data (e.g. emissions 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. Model performance depends on the quality of the input data and on the way we represent the dynamical and chemical processes leading to gas and aerosol concentrations. Many approaches exist to manage these two points, leading to some variability among model results. This variability can be understood as the modelling uncertainty. Previous studies investigated the uncertainties associated with certain processes when air chemistry transport models are used, through model ensemble approaches such as those described in Vautard et al. (2006, 2009) and Van Loon et al. (2007). Other studies investigated the uncertainties associated with model resolution (De Meij et al., 2007; Wang et al., 2015; Huang et al., 2021), chemistry (Textor et al., 2006; Thunis et al., 2021b; Clappier et al., 2021), meteorology (De Meij et al., 2009; Gilliam et al., 2015) and emission inventories (Thunis et al., 2021c; Colette et al., 2017). Over the years, air quality modelling has improved as model uncertainties have been reduced. Often, classical statistical parameters are used to evaluate the ACTM's capability in calculating air pollutants. For example, bias (a measure of overestimation or underestimation), standard deviation (a measure of the dispersion of the observed/calculated values around

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

the mean), temporal correlation coefficient (the linear relationship between model and observations) and root mean square error (a measure of difference between the model and the observations; measure of accuracy) to name a few. In the United States of America, modelling guidance and performing evaluation was first introduced by the US Environmental Protection Agency (EPA) in 1991 (EPA, 1991). This was followed by the introduction of the concepts of “goals” (i.e. model accuracy) and “criteria” (i.e. threshold of model performance) in studies by Boylan and Russell (2006) and Emery et al. (2017). In the USA, air quality models are evaluated based on several model performance indicators to ensure their accuracy and reliability. These indicators are mean bias (MB), mean absolute error (MAE), root mean square error (RMSE), fractional bias (FB), normalised mean bias (NMB), normalised mean error (NME), Pearson correlation coefficient (R or R^2) and index of agreement (IOA). For operational air quality performance, additional indicators are used: Prediction Accuracy, Hit Rate & False Alarm Rate, and Skill Scores. The EPA has specific regulatory performance criteria for key pollutants such as $\text{PM}_{2.5}$, NO_2 and O_3 .

For O_3 modelling, a model is considered acceptable if

- NMB is within $\pm 15\%$
- NME is $\leq 25\%$.

For $\text{PM}_{2.5}$ the performance goals are

- NMB within $\pm 30\%$
- NME $\leq 50\%$.

Also, the EPA's Support Center for Regulatory Atmospheric Modeling (SCRAM) provides resources and guidance on air quality models and their evaluation.

In China, Huang et al. (2021) proposes benchmarks for MB, MAE, RMSE, IOA, R and FB for air quality model applications since there are no unified guidelines or benchmarks developed for ACTM applications in China. Huang et al.'s (2021) methodology is based on Emery et al. (2017), applying goals and criteria for NMB, NME, FB, FE, IOA and R . Note that the model criteria are fixed in Huang et al. (2021) and Emery et al. (2017), while in our work the criteria depend on the observation uncertainties, which is different for each pollutant.

These indicators are, in general, used to assess model performance against measurements. However, these indicators do not tell us whether model results have reached a sufficient level of quality for a given application. In Huang et al. (2021), recommendations are given to provide a better overview of model performance. For example, for $\text{PM}_{2.5}$ the NMB should be between 10 % and 20 % and R should lie between 0.6 and 0.7 for hourly and daily $\text{PM}_{2.5}$ and between 0.70 and 0.90 for monthly $\text{PM}_{2.5}$ concentration values. Different temporal resolutions for $\text{PM}_{2.5}$ calculated values are introduced. Furthermore, benchmarks for speciated

PM components (elemental/organic carbon, nitrate, sulfate and ammonium) were recommended.

Along the same lines, the Forum for Air quality Modelling (FAIRMODE) (<https://fairmode.jrc.ec.europa.eu/home/index>, last access: 7 July 2025) developed several specific quality assurance and quality control (QA/QC) indicators and associated a threshold to each of them; these indicate 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 ACTM performances are Kushta et al. (2019) and Thunis et al. (2021a).

Note that the goals and criteria proposed in the US and China remain independent of the concentration level. In this work, we define a threshold on the maximum accepted modelling uncertainty. Because we do not know the modelling uncertainty in practice, we set it to be proportional to the measurement uncertainty. With this definition, the more uncertain the measurement is (e.g. relative uncertainties become larger in the lower concentration range), the more flexibility we allow to the modelling results, i.e. a higher threshold value (and vice versa).

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 models, methodology and emission inventories are given in Sect. 2, followed by the analysis of the results in Sect. 3. In Sect. 4 the conclusions are provided.

2 Methodology

CAMS produces annual air quality (interim) reanalysis for the European domain at a spatial resolution of $0.1 \times 0.1^\circ$. 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 (Marécal et al., 2015; CAMS, 2020).

We assess the relevance and usefulness of FAIRMODE's model quality assessment indicators by means of evaluating simulated air pollutants (NO_2 , O_3 , $\text{PM}_{2.5}$ and PM_{10}) by the eight CAMS models for the year 2021 by comparing the results with observational data from the European air quality database and assessing the results against specific performance indicators. The evaluation of the model's performance is based on the comparison with observations that are not used to assimilate calculated concentrations. The eight CAMS models are CHIMERE (FR; CHIA), DEHM (DK; DEHMA), EMEP (NO; EMPA), FMIA-SILAM (FI; FMIA), GEMAQ (PL; GEMAQA), KNMA-LOTUS-EUROS (NL; KNMA), MFM-MOCAGE (FR; MFMA), RIU-EURAD-IM (DE; RIUA) and Ensemble (ENSKCa). The CAMS regional

air quality models generate reanalysis, detailing the concentrations of major atmospheric pollutants in the lowest layers of the atmosphere across the European domain (ranging from 25.0° W to 45.0° E and 30.0° N to 72.0° N). The horizontal resolution is approximately 0.1°, varying from around 3 km at 72.0° N to 10 km at 30.0° N.

For that reason, an overview of the type of assimilation methodology, which species are assimilated, together with gas and aerosol schemes are given in Table S1 of the Supplement. More details of the different models are described in CAMS (2020) (<https://confluence.ecmwf.int/display/CKB/CAMS+Regional:+European+air+quality+reanalyses+data+documentation>, last access: 7 July 2025). The data can be downloaded here: <https://atmosphere.copernicus.eu/data> (last access: 7 July 2025).

For the statistical analysis, the FAIRMODE benchmarking methodology is applied; this methodology provides many different statistical parameters, which are described in FAIRMODE's guidance document (Janssen and Thunis, 2022).

The indicators and modelling criteria described in this study were defined in the context of FAIRMODE to support the application of modelling in the context of the Air Quality Directive.

Initially, FAIRMODE developed a single model performance indicator: the Modelling Quality Indicator (MQI). While this indicator provides a relevant pass/fail test, passing the test does not ensure that modelling results are fit for purpose. This is why additional indicators have progressively been added, in particular to assess how models capture temporal and spatial aspects. The 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

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

where $U(O_i)$ 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, Eq. (1) can be generalised to

$$\text{MQI} = \frac{\text{RMSE}}{\beta \text{RMS}_U}. \quad (2)$$

A complete time series entails 75 % data availability over the selected time period. Note that this number is less than the

one requested in the European Commission's Ambient Air Quality Directive (AAQD, 2024) (i.e. 90 %) to increase the available number of measurement stations for validation. We, however, impose that available data are representative of the full year.

With this formulation, the RMSE between observed and modelled values (numerator) is compared to the root mean square sum of the measurement uncertainties (RMS_U), the value of which 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 ($U(\bar{O})$):

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

More details on Eqs. (1)–(3) can be found in the modelling quality objective (MQO) guidance document (Janssen and Thunis, 2022).

For the statistical analysis of the four air pollutants, we use for NO₂ the hourly values and for O₃ the 8 h running mean maximum values, while for PM_{2.5} and PM₁₀ the daily averages are used. These different time intervals are in compliance with the EU air quality standards as stated in the AAQD. 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 MQO is fulfilled when the MQI is less than or equal to 1.0 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 recalibration on the daily observations lead to larger uncertainties, which are compensated for yearly averages.

The main drawback of the MQOs is that they provide single summary pass/fail information for a modelling application. This simple test does not prevent a modelling application passing 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 or 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:

$$\text{RMS}_{\bar{U}} = \sqrt{\frac{1}{N} \sum U(\bar{O})^2}. \quad (4)$$

The same approach applies for yearly frequency output. These indicators are defined in Table 1.

On top of these already agreed indicators included in the 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 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

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

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

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

These indicators are then normalised by the measurement uncertainty, see Table 2.

As mentioned earlier, the MQO generally applies to the average of a specific period, currently 1 year. Consequently, it provides no information on whether the modelling application manages to capture the temporal variability of the air quality situation. Since the AAQDs include 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 (Table 3). These include seasonal, week/weekend 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. The AAQD of the European Commission provides definitions for different types of air quality monitoring stations based on their location and the pollution sources they are exposed to. These station types ensure a comprehensive assessment of air quality across different environments, helping policymakers and researchers analyse pollution trends and enforce regulatory limits.

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).

The key definitions are as follows:

- *Traffic stations*: near major roads or intersections (at least 25 m from major intersections, but no more than 10 m from the road), dominated by vehicle emissions (NO_2 , PM_{10} , $\text{PM}_{2.5}$), reflecting population exposure to road transport pollution.
- *Urban stations*: in residential or commercial areas (more than 50 m away from major roads and more than 4 km away from industrial sources), measuring background pollution levels affecting the general urban population.
- *Industrial stations*: near factories or power plants, monitoring emissions like SO_2 , NO_2 , heavy metals and volatile organic compounds (VOCs).
- *Rural stations*: in the countryside or suburban areas (at least 20 km from urban areas and 5 km from industrial sources), assessing regional and long-range pollution transport.

A more detailed description of the station types can be found in Annex III (“Assessment of Air Quality and Location of Sampling Points”) of the Air Quality Directive (2008/50/EC).

3 Results

To best visualise 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 model 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 SD), followed by (3) studying model 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 Model performance analysis for NO_2

In Fig. 1, the statistics for NO_2 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

Table 1. Model performance indicators for temporal and spatial correlation.

	Model performance indicator (MPI)	Model performance criterion (MPC)
Correlation (Eq. 5)	$\text{MPI} = \frac{1-R}{0.5\beta^2 \frac{\text{RMS}_{\overline{U}}^2}{\sigma_O \sigma_M}}$ (bias = 0, $\sigma_O = \sigma_M$)	MPC: $\text{MPI} \leq 1$
Standard deviation (Eq. 6)	$\text{MPI} = \frac{ \sigma_M - \sigma_O }{\beta \text{RMS}_{\overline{U}}}$ (bias = 0, $R = 1$)	

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

Table 2. Model performance indicators that describe the incremental change between rural background (RB) and urban background (UB) locations.

	Model performance indicator (MPI)	Model performance criterion (MPC)
UB-RB (Eq. 9)	$\text{MPI} = 1/\beta \cdot \frac{\text{INC}_{\text{UB-RB}}^{\text{model}} - \text{INC}_{\text{UB-RB}}^{\text{observed}}}{0.5 \cdot (\text{RMS}_{\overline{U(\text{UB})}} + \text{RMS}_{\overline{U(\text{RB})}})}$	MPC: $\text{MPI} \leq 1$
UT-UB (Eq. 10)	$\text{MPI} = 1/\beta \cdot \frac{\text{INC}_{\text{UB-UT}}^{\text{model}} - \text{INC}_{\text{UB-UT}}^{\text{observed}}}{0.5 \cdot (\text{RMS}_{\overline{U(\text{UB})}} + \text{RMS}_{\overline{U(\text{UT})}})}$	

Where UT stands for “urban traffic”.

Ensemble median. The other ACTMs are presented in different colours.

Figure 1 shows that the yearly MQIs (MQI_YR) are generally higher than 1.5 for all models and all countries, 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_YR values are higher than those for MQI_HD, indicating that each model has difficulties in capturing well the observed yearly concentrations for NO₂.

As mentioned earlier, the MQOs tell us whether 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 Eqs. (4)–(6), which present the bias and temporal and spatial correlation.

A more stringent source of information to the additional indicators in Eqs. (4)–(6) are presented in Eqs. (7)–(10). We see that for example these indicators describe the differences between biases for day vs. 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 (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 does not fulfil the MQO. This highlights the value of these indicators in assessing model performance.

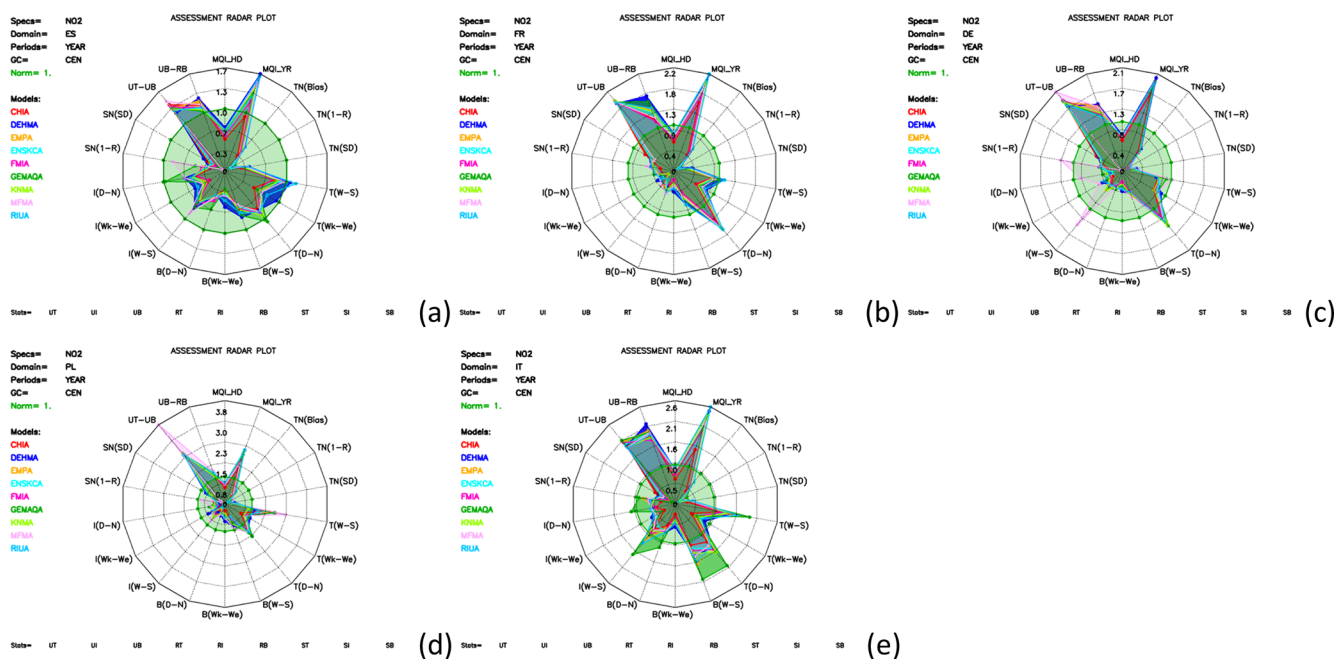
This can be explained by the fact that the model resolution (0.1×0.1) is too coarse to capture the emissions from the road transport sector. This is illustrated in Fig. 2, which shows the difference between observations and calculated yearly mean NO₂ concentrations for traffic, industry, all and background stations for Germany. The calculated NO₂ concentrations for traffic and all stations remain flat, i.e. the concentrations are very similar around $13 \mu\text{g m}^{-3}$. While the difference in observed concentrations (grey bar) between traffic stations and all stations is around $7 \mu\text{g m}^{-3}$ (27 for traffic and $20 \mu\text{g m}^{-3}$ for all stations).

Also, the bias for traffic stations is much larger (up to $-14 \mu\text{g m}^{-3}$), while the bias for all stations is smaller (up to $-9 \mu\text{g m}^{-3}$), see Fig. 3. This indicates that the models have difficulties in calculating the NO₂ concentrations for traffic 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 agree well with the ob-

Table 3. Model quality indicators at different frequencies: seasonal, week/weekend or day/night.

		Model performance indicator (MPI)	Model perf. criteria (MPC)
Seasonal (Eq. 11)	Industry	$\text{MPI} = \frac{\text{SeasDiff}_{\text{Ind}}^{\text{mod}} - \text{SeasDiff}_{\text{Ind}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	MPC: $\text{MPI} \leq 1$
	Traffic	$\text{MPI} = \frac{\text{SeasDiff}_{\text{traffic}}^{\text{mod}} - \text{SeasDiff}_{\text{traffic}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{\text{SeasDiff}_{\text{bg}}^{\text{mod}} - \text{SeasDiff}_{\text{bg}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
Week/ weekend (Eq. 12)	Industry	$\text{MPI} = \frac{\text{WeekDiff}_{\text{Ind}}^{\text{mod}} - \text{WeekDiff}_{\text{Ind}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
	Traffic	$\text{MPI} = \frac{\text{WeekDiff}_{\text{traffic}}^{\text{mod}} - \text{WeekDiff}_{\text{traffic}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{\text{WeekDiff}_{\text{bg}}^{\text{mod}} - \text{WeekDiff}_{\text{bg}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
Day/night (Eq. 13)	Industry	$\text{MPI} = \frac{\text{DayDiff}_{\text{Ind}}^{\text{mod}} - \text{DayDiff}_{\text{Ind}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
	Traffic	$\text{MPI} = \frac{\text{DayDiff}_{\text{traffic}}^{\text{mod}} - \text{DayDiff}_{\text{traffic}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	
	Background	$\text{MPI} = \frac{\text{DayDiff}_{\text{bg}}^{\text{mod}} - \text{DayDiff}_{\text{bg}}^{\text{obs}}}{\beta \text{RMS}_{\bar{U}}}$	

**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 hourly MQI (MQI_{HD}), yearly MQI (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 (UT–UB), yearly urban background vs rural background (UB–RB).

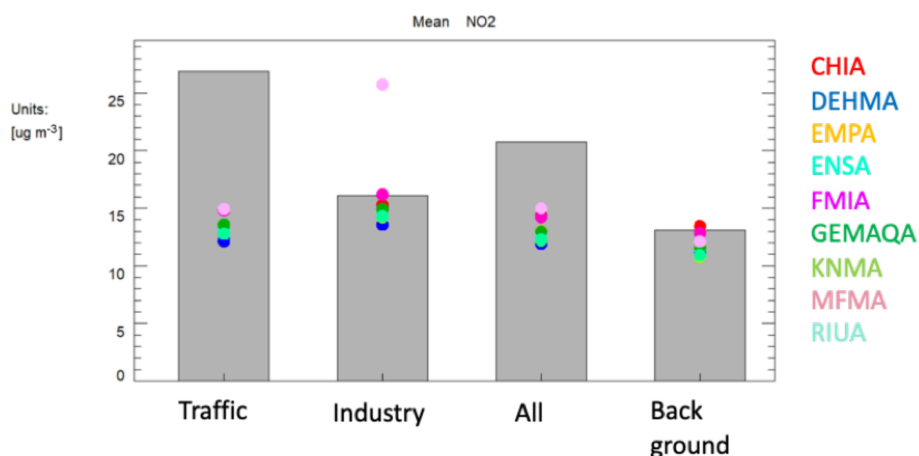


Figure 2. Yearly mean observed (grey bar) and calculated (coloured dots) NO₂ concentrations for Germany for traffic, industry, all and background stations.

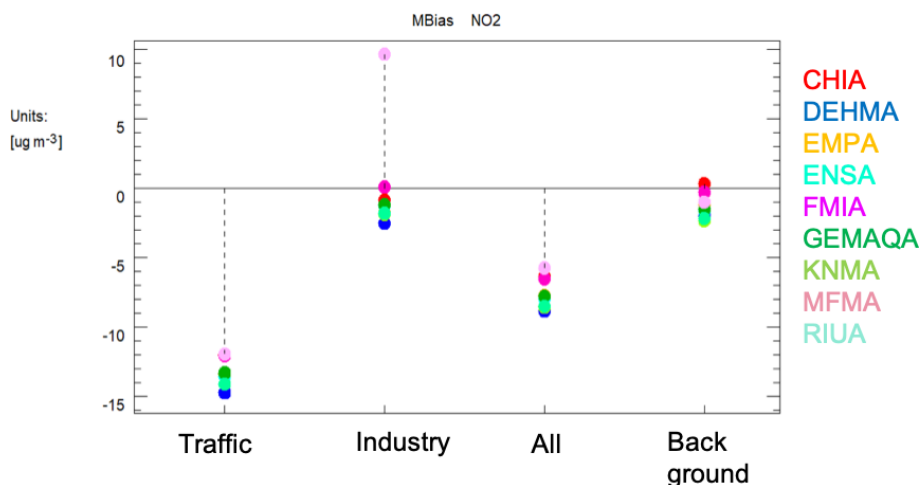


Figure 3. Yearly mean bias for NO₂ for traffic, industry, all and background stations for the different models (coloured dots) for Germany stations.

servations. This reflects into low bias for industry and background stations ($< 3 \mu\text{g m}^{-3}$).

Looking in more detail, we show in Fig. 4 the comparison between the model vs. day–night and winter–summer mean observations for traffic and background stations in Italy. Well-behaving results should lie along the 1 : 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 RIUA. Note that for the other countries the day–night and winter–summer profiles are satisfactory for background stations, but not for traffic stations. In general, for background stations, all indicator values remain below the threshold of 1.0, except for the GEMAQA model in Italy. This sug-

gests that the models perform better in less complex environments and that these indicators may be less effective for assessing model performance in this context.

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. The short lifetime of NO₂ (about 1 h) requires high model resolution to capture well the non-linear production and loss of NO₂ concentrations.

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 in-

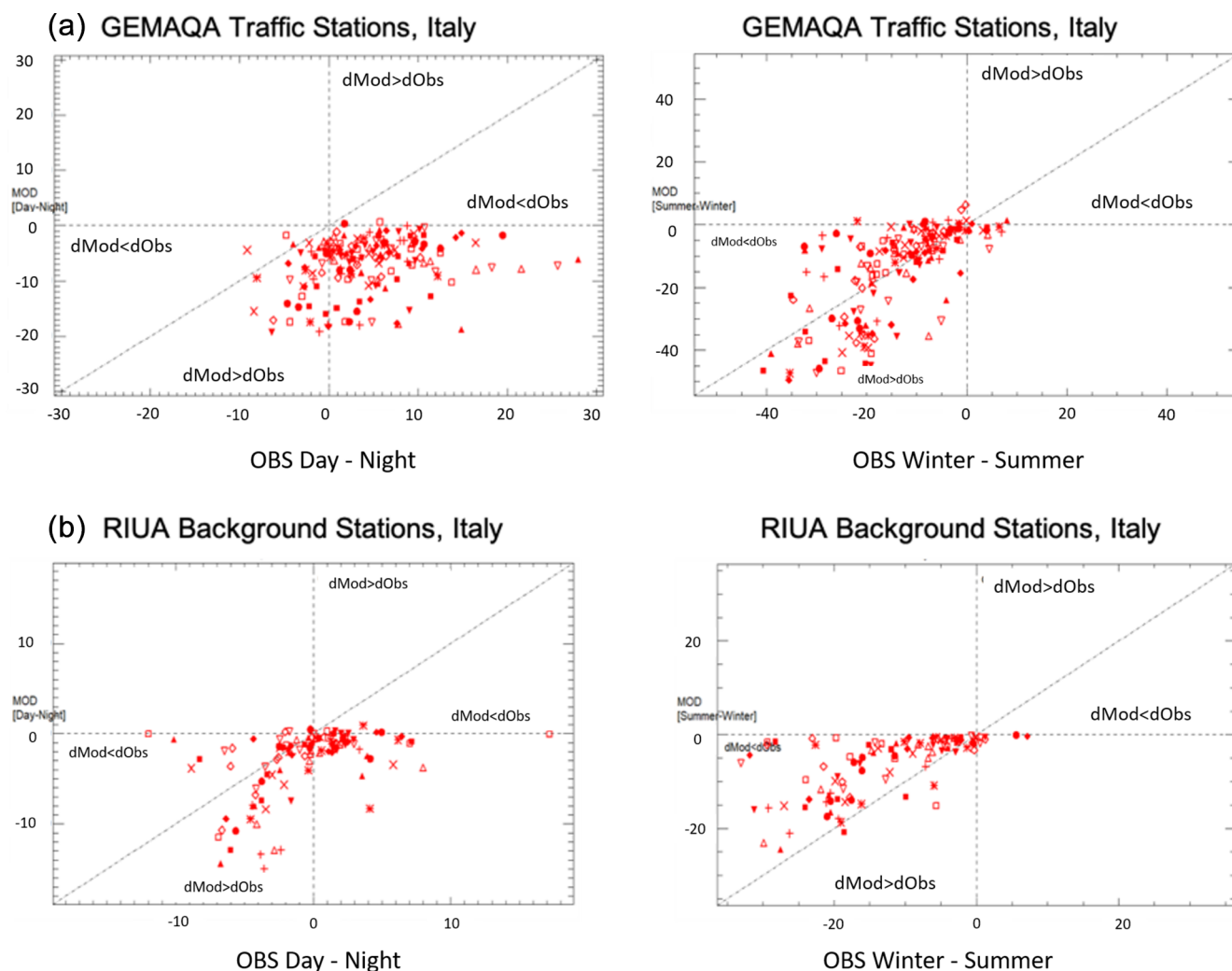


Figure 4. NO₂ scatter plots of modelled vs observed day–night and summer–winter mean differences for (a) traffic stations by GEMAQA and (b) background stations by RIUA model.

dicators 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 S2.

3.2 Model performance analysis for 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 already mentioned. 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 normalised temporal correlation coefficient is expressed in terms of $1 - R$; the threshold for this indicator

remains 1 as for all indicators, meaning that values below 1 fulfil the objective. Values closer to zero indicate even better performances. This implies that other indicators are required to perform a more stringent evaluation of the ACTM.

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. For Germany, all indicators are below unity for the different models, apart from UT-UB and UB-RB by DEHMA and EMPA, and MQI_YRs by DEHMA, GEMAQA and MFMA.

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 ($\sim 1 \mu\text{g m}^{-3}$) for all the station types for Spain, while most

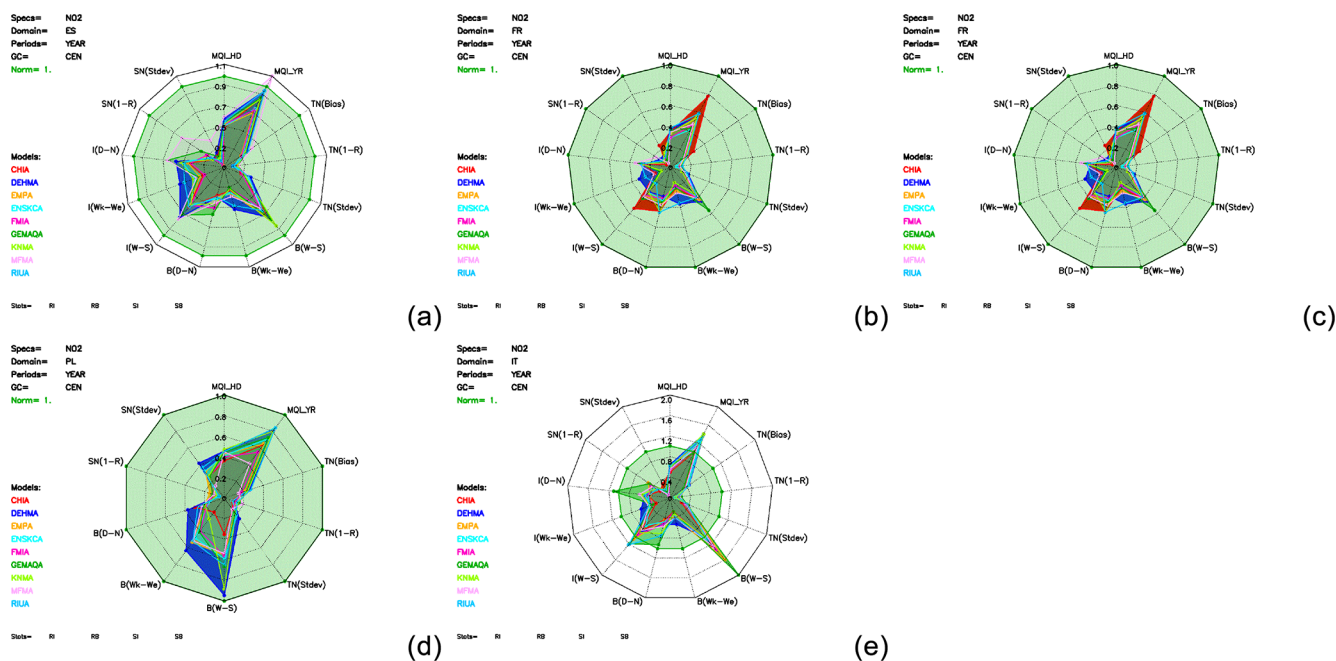


Figure 5. Radar plots of the calculated air quality model indicators for NO₂ for different countries excluding the traffic stations: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy.

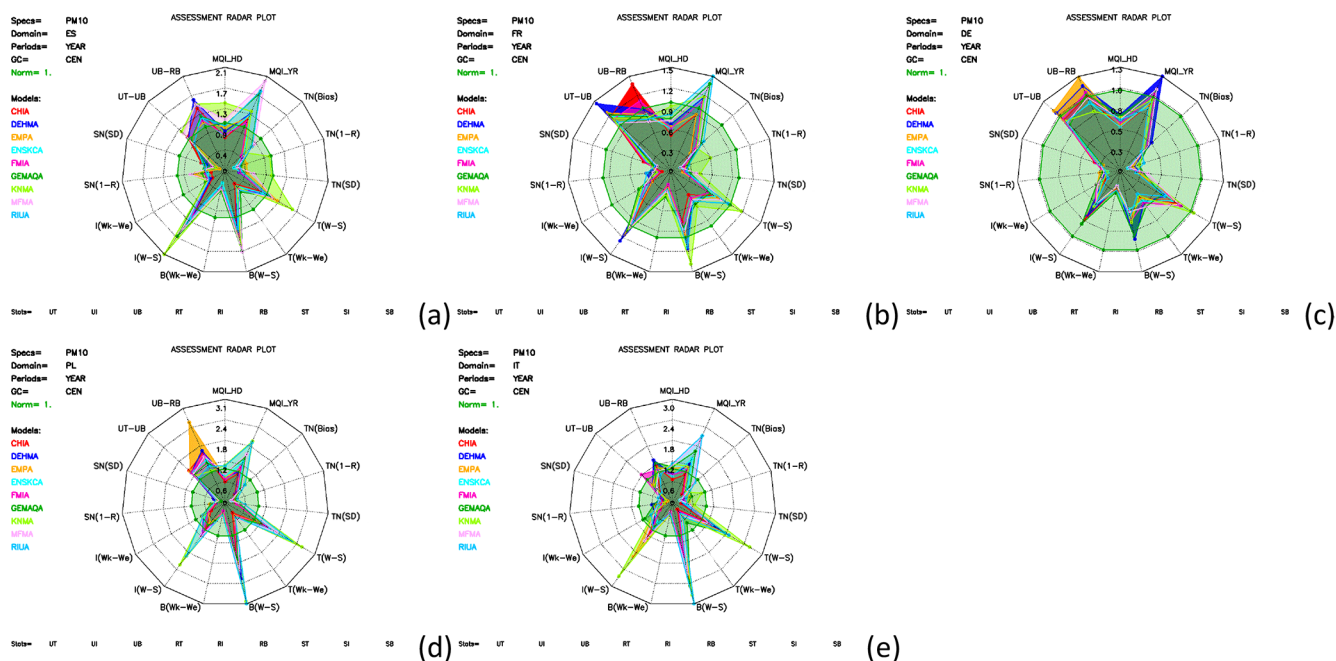


Figure 6. Radar plots of the calculated air quality model indicators for PM₁₀ for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are hourly MQI (MQI_{HD}), yearly MQI (MQI_{YR}), bias, 1 – *R* (time), standard deviation (time), gradients for winter–summer, 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).

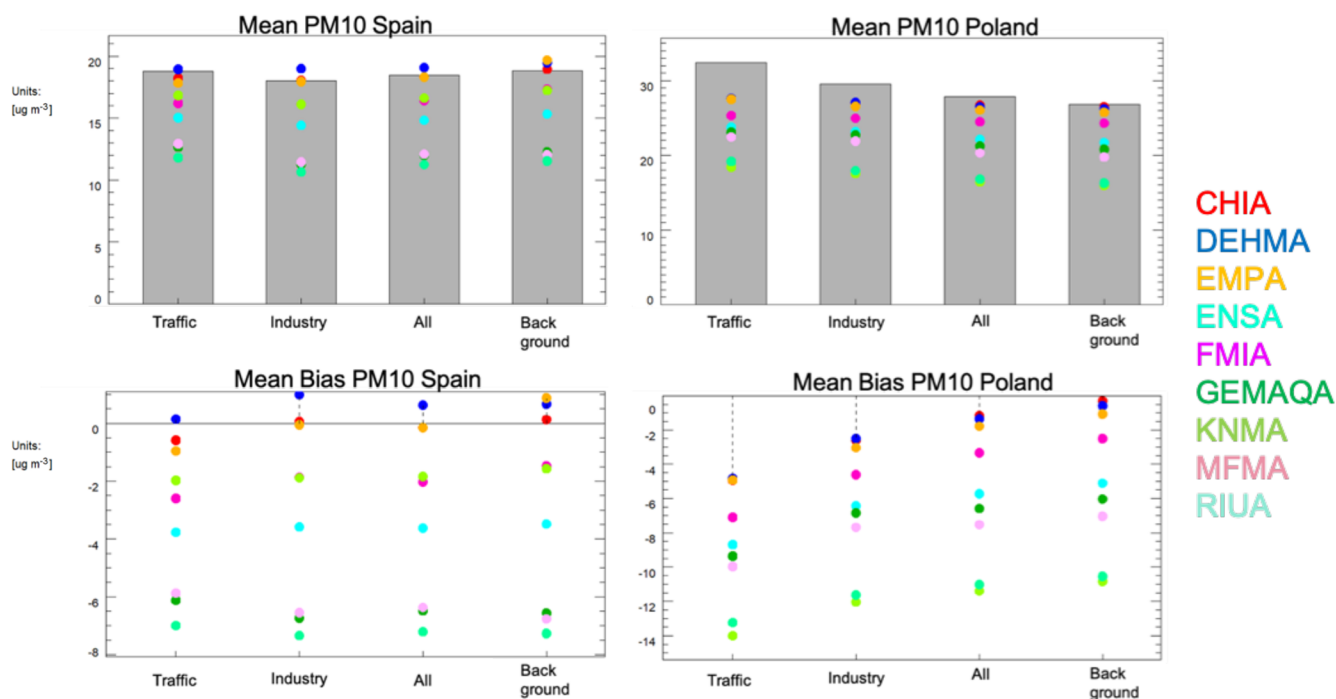


Figure 7. Mean calculated PM_{10} concentrations by the nine models (indicated by coloured bullets) for the different measurement stations (grey bars for traffic, industry, all and background) for Spain and Poland, together with the bias.

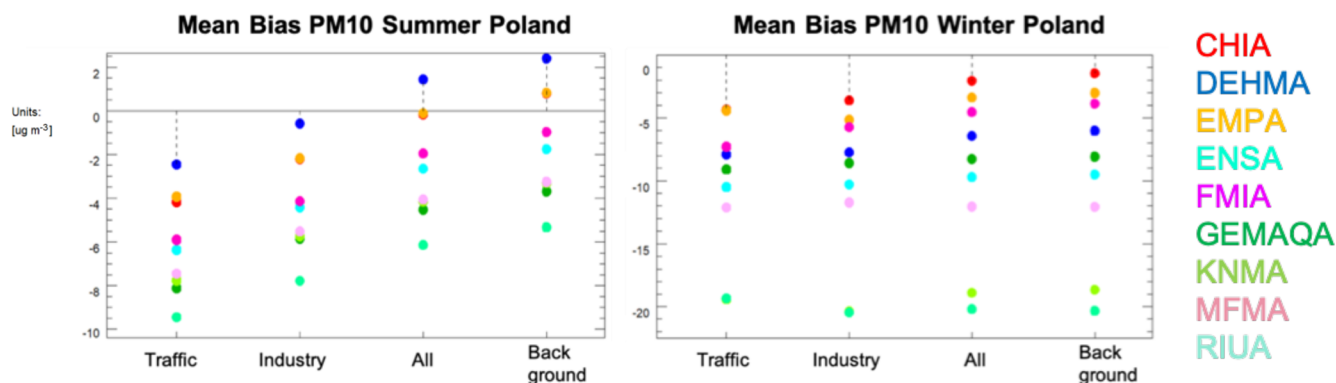


Figure 8. Mean bias PM_{10} for summer (JJA) and winter (DJF) for Poland by all the models for the different station types (traffic, industry, all and background).

of the models underestimate on average the observed PM_{10} concentrations.

For Poland, all the models underestimate the observed PM_{10} concentrations for the different station types (Fig. 7). The highest PM_{10} concentrations are observed for traffic stations for Poland. It is for these stations that the model capability in calculating elevated PM_{10} 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.

The radar plots show that the winter–summer gradients are larger than 1.0 for the different countries. For that reason, we analyse in more detail the PM_{10} 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 $\sim -10 \mu\text{g m}^{-3}$) with a small positive bias for a few models when all and background stations are considered. For the winter period (Fig. 8, right panel), the mean bias is a factor ~ 2 higher than for the summer, with RIUA and KNMA showing the highest bias (up to $\sim -20 \mu\text{g m}^{-3}$) for the four different station types. This indicates that the models

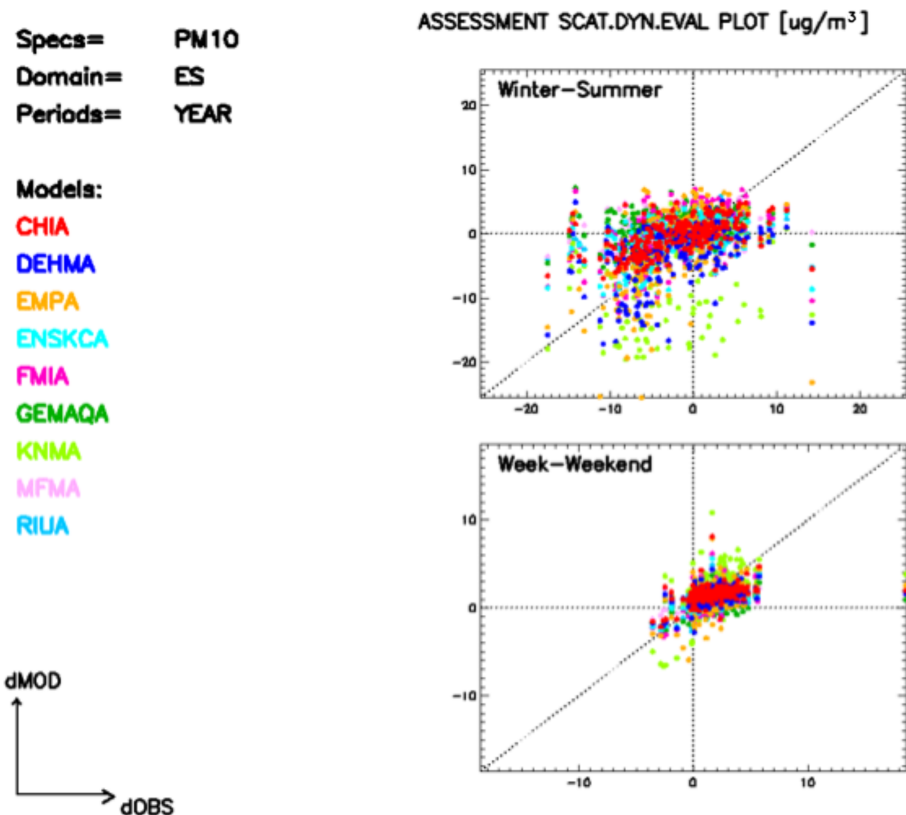


Figure 9. PM₁₀ Scatter plots of modelled vs observed winter–summer and week–weekend mean differences for Spain for all the models.

underestimate the PM₁₀ concentrations for the whole country, especially during wintertime, even though the model concentrations are assimilated.

When traffic stations are excluded from the analysis, it appears that only for Germany, Poland and Italy the Ensemble's MQI_YR is lower (e.g. for Poland ~ 1.4 vs ~ 1.0 without traffic stations). As mentioned earlier, the winter–summer profiles for industry, background (and to some extent traffic) stations hampers the overall model performance in calculating the PM₁₀ concentrations (indices are well above the reference criterion 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.

The analysis above 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. On the other hand, temporal correlation and standard deviation indicators seem to be less useful for evaluating model performance in this context.

3.3 Model performance analysis for PM_{2.5}

Yearly MQIs for PM_{2.5} 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 PM_{2.5} than for PM₁₀ in the MQI equations, allowing less stringency on the model results when calculating the MQI for PM_{2.5} (Thunis et al., 2021a).

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 $-13 \mu\text{g m}^{-3}$) than for the summer (up to $-3 \mu\text{g m}^{-3}$), see Fig. 11. Our analysis further showed that for PM_{2.5} 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.

As we have seen before, considering only the MQI for the model evaluation does not provide enough information on 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,

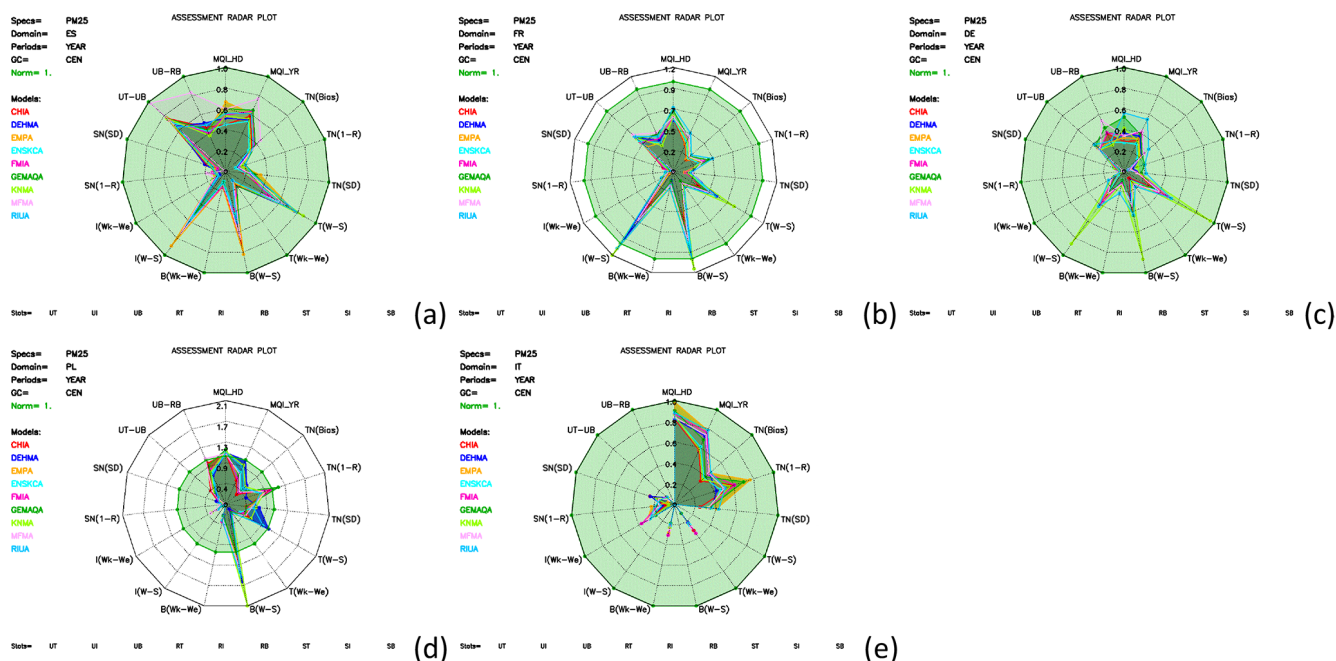


Figure 10. Radar plots of the calculated air quality model indicators for $\text{PM}_{2.5}$ for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are hourly MQI (MQI_{HD}), yearly MQI (MQI_{YR}), bias, $1 - R$ (time), standard deviation (time), gradients for 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).

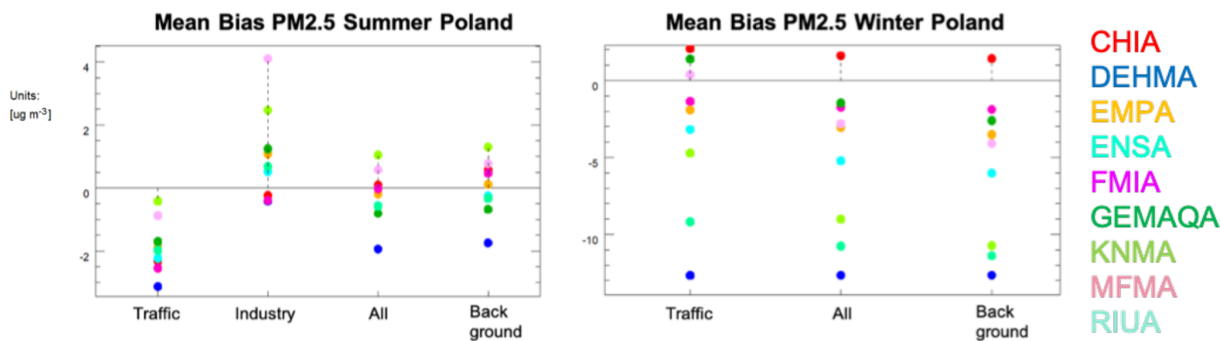


Figure 11. Mean bias $\text{PM}_{2.5}$ for summer (JJA) and winter (DJF) for Poland by the models for the different station types (traffic, industry, all and background). Note that for winter, there is only one industry station; therefore the bias for this station type is not shown.

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.

The analysis raises questions about the stringency of the indicators for $\text{PM}_{2.5}$, as passing the criteria does not necessarily indicate flawless performance. The bias and the winter–summer indicators reveal potential problems in air quality modelling for $\text{PM}_{2.5}$ and for that reason are very useful.

3.4 Model performance analysis for O_3

For O_3 , all indicators are lower than unity for France, indicating that the models capture well the 8 h maximum O_3 values (Fig. 12). Except for 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

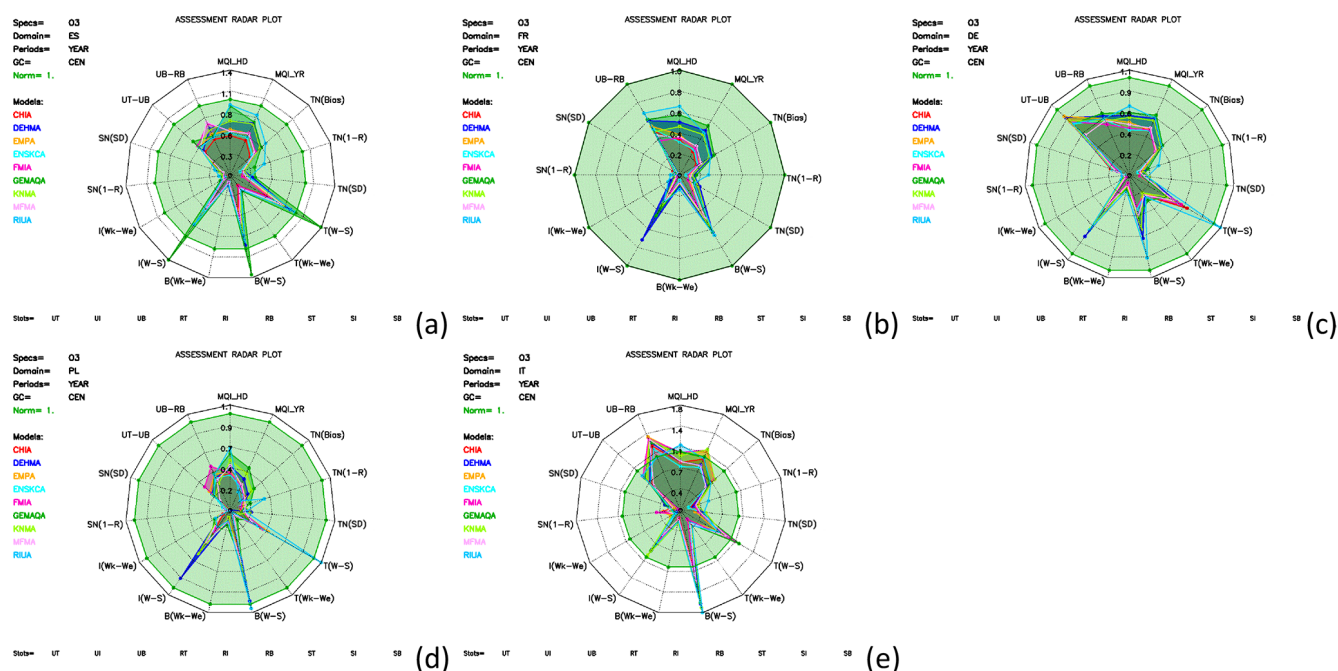


Figure 12. Radar plots of the calculated air quality model indicators for 8 h maximum O₃ values for different countries: (a) Spain, (b) France, (c) Germany, (d) Poland and (e) Italy. Indicators are hourly MQI (MQI_{HD}), yearly MQI (MQI_{YR}), bias, 1 – *R* (time), standard deviation (time), gradients 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).

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 difficulty capturing 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 difficulty capturing 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 GEMQA and EMPA.

Even though the daily and yearly MQI for 8 h 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 concentration modelling.

4 Conclusion

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 MQIs, as well as the winter–summer and spatial gradient indicators, clearly show the challenges the models face in accu-

ately 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 model performance, confirming that the indicators are effectively capturing the model difficulties. For background stations, all indicator values fall below the threshold of 1.0, except for the GEMQA model in Italy, suggesting better model performance in less complex environments.

When analysing fine particulate matter (PM_{2.5}), 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 PM_{2.5}, 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 O₃, although the daily and yearly MQI for the 8 h 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.

Code availability. The IDL® (8.8.3) source code (<https://www.nv5geospatialsoftware.com/docs/idl-install.html>, NV5 Geospatial Solutions Inc., 2025) of the DeltaToolLight (version 1.4) of the screening method of the statistical analysis can be found here: <https://doi.org/10.5281/zenodo.14870503> (Cuvelier et al., 2025).

Data availability. The CAMS data are available from the Copernicus CAMS website, via <https://doi.org/10.24381/7cc0465a>; generated using Copernicus Atmosphere Monitoring Service (2021). The Copernicus Atmosphere Monitoring Service is operated by the European Centre for Medium-Range Weather Forecasts on behalf of the European Commission as part of the Copernicus Programme (<http://copernicus.eu>, last access: 7 July 2025). Also, the observation data and modelling data (at station location) can be found here: <https://doi.org/10.5281/zenodo.14870503> (Cuvelier et al., 2025).

Supplement. The supplement related to this article is available online at <https://doi.org/10.5194/gmd-18-4231-2025-supplement>.

Author contributions. 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.

Competing interests. The contact author has declared that none of the authors has any competing interests.

Disclaimer. Publisher's note: Copernicus Publications remains neutral with regard to jurisdictional claims made in the text, published maps, institutional affiliations, or any other geographical representation in this paper. While Copernicus Publications makes every effort to include appropriate place names, the final responsibility lies with the authors.

Review statement. This paper was edited by Slimane Bekki and reviewed by three anonymous referees.

References

- Ambient Air Quality Directive: Directive (EU) 2024/2881 of the European Parliament and of the Council of 23 October 2024 on ambient air quality and cleaner air for Europe (recast), <https://eur-lex.europa.eu/eli/dir/2024/2881/oj> (last access: 11 July 2025), 2024.
- Boylan, J. W. and Russell, A. G.: PM and light extinction model performance metrics, goals, and criteria for three-dimensional air quality models, *Atmos. Environ.*, 40, 4946–4959, <https://doi.org/10.1016/j.atmosenv.2005.09.087>, 2006.
- Clappier, A., Thunis, P., Beekmann, M., Putaud, J. P., and De Meij, A.: Impact of SO_x, NO_x and NH₃ emission reductions on PM_{2.5} concentrations across Europe: Hints for future measure development, *Environ. Int.*, 156, 106699, <https://doi.org/10.1016/j.envint.2021.106699>, 2021.
- Colette, A., Andersson, C., Manders, A., Mar, K., Mircea, M., Pay, M.-T., Raffort, V., Tsyro, S., Cuvelier, C., Adani, M., Bessagnet, B., Bergström, R., Briganti, G., Butler, T., Cappelletti, A., Couvidat, F., D'Isidoro, M., Doumbia, T., Fagerli, H., Granier, C., Heyes, C., Klimont, Z., Ojha, N., Otero, N., Schaap, M., Sindelarova, K., Stegehuis, A. I., Roustan, Y., Vautard, R., van Meijgaard, E., Vivanco, M. G., and Wind, P.: EURODELTA-Trends, a multi-model experiment of air quality hindcast in Europe over 1990–2010, *Geosci. Model Dev.*, 10, 3255–3276, <https://doi.org/10.5194/gmd-10-3255-2017>, 2017.
- CAMS – Copernicus Atmospheric Monitoring Service: Regional Production, Updated documentation covering all Regional operational systems and the ENSEMBLE, Following U3 upgrade, November 2020, <https://confluence.ecmwf.int/display/CKB/CAMS+Regional:+European+air+quality+reanalyses+data+documentation> (last access: 10 July 2025), 2020.
- Copernicus Atmosphere Monitoring Service (CAMS): CAMS European air quality reanalyses, Copernicus Atmosphere Monitoring Service (CAMS) Atmosphere Data Store [data set], <https://doi.org/10.24381/7cc0465a>, 2021.
- Cuvelier, C., Thunis, P., Pisoni, E., and De Meij, A.: Source code and data to: “A new set of indicators for model evaluation complementing to FAIRMODE's MQO”, Zenodo [code and data set], <https://doi.org/10.5281/zenodo.14870503>, 2025.
- De Meij, A., Wagner, S., Gobron, N., Thunis, P., Cuvelier, C., Dentener, F., and Schaap, M.: Model evaluation and scale issues in chemical and optical aerosol properties over the greater Milan area (Italy), for June 2001, *Atmos. Res.*, 85, 243–267, 2007.
- De Meij, A., Gzella, A., Cuvelier, C., Thunis, P., Bessagnet, B., Vinuesa, J. F., Menut, L., and Kelder, H. M.: The impact of MM5 and WRF meteorology over complex terrain on CHIMERE model calculations, *Atmos. Chem. Phys.*, 9, 6611–6632, <https://doi.org/10.5194/acp-9-6611-2009>, 2009.
- 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.

- Emery, C., Liu, Z., Russell, A. G., Odman, M. T., Yarwood, G., and Kumar, N.: Recommendations on statistics and benchmarks to assess photochemical model performance, *JAPCA J. Air. Waste Manage.*, 67, 582–598, <https://doi.org/10.1080/10962247.2016.1265027>, 2017.
- EPA: Guideline for regulatory application of the Urban Airshed Model (No. PB-92-108760/XAB), Environmental Protection Agency, Research Triangle Park, NC, USA, 1991.
- Gilliam, R. C., Hogrefe, C., Godowitch, J. M., Napelenok, S., Mathur, R., and Rao, S. T.: Impact of Inherent Meteorology Uncertainty on Air Quality Model Predictions, *J. Geophys. Res.-Atmos.*, 120, 12259–12280, 2015.
- Huang, L., Zhu, Y., Zhai, H., Xue, S., Zhu, T., Shao, Y., Liu, Z., Emery, C., Yarwood, G., Wang, Y., Fu, J., Zhang, K., and Li, L.: Recommendations on benchmarks for numerical air quality model applications in China – Part 1: PM_{2.5} and chemical species, *Atmos. Chem. Phys.*, 21, 2725–2743, <https://doi.org/10.5194/acp-21-2725-2021>, 2021.
- Janssen, S. and 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, <https://doi.org/10.2760/41988>, 2022.
- Kushta, J., Georgiou, G. K., Proestos, Y., Christoudias, T., Thunis, P., Savvides, C., Papadopoulos, C., and Lelieveld, J.: Evaluation of EU air quality standards through modeling and the FAIRMODE benchmarking methodology, *Air Qual. Atmos. Health*, 12, 73–86, <https://doi.org/10.1007/s11869-018-0631-z>, 2019.
- Marécal, V., Peuch, V.-H., Andersson, C., Andersson, S., Arteta, J., Beekmann, M., Benedictow, A., Bergström, R., Bessagnet, B., Cansado, A., Chéroux, F., Colette, A., Coman, A., Curier, R. L., Denier van der Gon, H. A. C., Drouin, A., Elbern, H., Emili, E., Engelen, R. J., Eskes, H. J., Foret, G., Friese, E., Gauss, M., Giannaros, C., Guth, J., Joly, M., Jaumouillé, E., Josse, B., Kadyrov, N., Kaiser, J. W., Krajsek, K., Kuenen, J., Kumar, U., Liora, N., Lopez, E., Malherbe, L., Martinez, I., Melas, D., Meleux, F., Menut, L., Moinat, P., Morales, T., Parmentier, J., Piacentini, A., Plu, M., Poupkou, A., Queguiner, S., Robertson, L., Rouil, L., Schaap, M., Segers, A., Sofiev, M., Tarasson, L., Thomas, M., Timmermans, R., Valdebenito, Á., van Velthoven, P., van Versendaal, R., Vira, J., and Ung, A.: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production, *Geosci. Model Dev.*, 8, 2777–2813, <https://doi.org/10.5194/gmd-8-2777-2015>, 2015.
- NV5 Geospatial Solutions Inc.: IDL® 8.8.3, <https://www.nv5geospatialsoftware.com/docs/idl-install.html>, last access: 7 July 2025.
- Textor, C., Schulz, M., Guibert, S., Kinne, S., Balkanski, Y., Bauer, S., Bernsten, T., Berglen, T., Boucher, O., Chin, M., Dentener, F., Diehl, T., Easter, R., Feichter, H., Fillmore, D., Ghan, S., Ginoux, P., Gong, S., Grini, A., Hendricks, J., Horowitz, L., Huang, P., Isaksen, I., Iversen, I., Kloster, S., Koch, D., Kirkevåg, A., Kristjansson, J. E., Krol, M., Lauer, A., Lamarque, J. F., Liu, X., Montanaro, V., Myhre, G., Penner, J., Pitari, G., Reddy, S., Seland, Ø., Stier, P., Takemura, T., and Tie, X.: Analysis and quantification of the diversities of aerosol life cycles within AeroCom, *Atmos. Chem. Phys.*, 6, 1777–1813, <https://doi.org/10.5194/acp-6-1777-2006>, 2006.
- Thunis, P., Crippa, M., Cuvelier, C., Guizzardi, D., De Meij, A., Oreggioni, G., and Pisoni, E.: Sensitivity of air quality modelling to different emission inventories: A case study over Europe, *Atm. Environ.*, 10, 100111, <https://doi.org/10.1016/j.aeaoa.2021.100111>, 2021a.
- Thunis, P., Clappier, A., Beekmann, M., Putaud, J. P., Cuvelier, C., Madrazo, J., and De Meij, A.: Non-linear response of PM_{2.5} to changes in NO_x and NH₃ emissions in the Po basin (Italy): consequences for air quality plans, *Atmos. Chem. Phys.*, 21, 9309–9327, <https://doi.org/10.5194/acp-21-9309-2021>, 2021b.
- Thunis, P., Crippa, M., Cuvelier, C., Guizzardi, D., De Meij, A., Oreggioni, G., and Pisoni, E.: Sensitivity of air quality modelling to different emission inventories: A case study over Europe, *Atmos. Environ.*, 10, 100111, <https://doi.org/10.1016/j.aeaoa.2021.100111>, 2021c.
- Van Loon, M., Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., and Krol, M. C.: Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble, *Atmos Environ.*, 41, 2083–2097, 2007.
- Vautard, R., Van Loon, M., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, K., Graf, A., Jonson, J. E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrasón, L., Thunis, P., Vignati, E., White, L., and Wind, P.: Is regional air quality model diversity representative of uncertainty for ozone simulation?, *Geophys. Res. Lett.*, 33, L24818, <https://doi.org/10.1029/2006GL027610>, 2006.
- Vautard, R., M., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, C., Foltescu, V., Graff, A., Kerschbaumer, A., Krol, M., Roberts, P., Rouil, L., Stern, R., Tarrason, L., Thunis, P., Vignati, E., and Wind, P.: Skill and uncertainty of a regional air quality model ensemble, *Atmos. Environ.*, 43, 4822–4832, <https://doi.org/10.1016/j.atmosenv.2008.09.083>, 2009.
- Wang, L., Wei, Z., Wei, W., Fu, J. S., Meng, C., and Ma, S.: Source apportionment of PM_{2.5} in top polluted cities in Hebei, China using the CMAQ model, *Atmos. Environ.*, 122, 723–736, <https://doi.org/10.1016/j.atmosenv.2015.10.041>, 2015.