We thank the reviewers for the constructive comments which have been very helpful in improving the manuscript. Please find below a point-to-point reply to the comments.

Reviewer 2

This paper assesses the relevance and usefulness of the model performance indicators developed within the FAIRMODE framework by evaluating 8 CAMS models and their ensemble results for predicting four major air pollutants (NO2, O3, PM2.5 and PM10) across Europe. The study compares the model predicted air pollutant concentrations with observations, and highlights the limitations of the current MQOs and the need to reconsider the strictness of some indicators for certain pollutants. The major limitation of the current MQOs is that they provide a single pass/fail summary for a modelling application, which allows a modelling test to pass for the wrong reason under certain circumstances.
Additionally, it does not provide any information on the capability of the model to reproduce spatial variability or on the timing of the pollution peaks. With these in mind, the authors propose a new set of indicators to assess the capacity of models to capture the temporal and spatial variability, complementing the current FAIRMODE MQOs. While the manuscript makes a valuable contribution to model performance evaluation by proposing more comprehensive indicators, I have several concerns for the authors to address before the manuscript can be considered for publication.

Major concerns:
1. The methodology section requires more detailed information. Key aspects such as the emissions inventory, meteorological simulations, modelling time period (winter? Summer? 2021 whole year?), modelling domain, and parameterizations of the models should be described. These details should at least be included in the supplementary information and briefly mentioned in the main text to help readers understand the origins of uncertainties. It would be helpful to include a brief discussion of the assumptions made during model construction and any limitations of the current approach.

We've added the following text to the manuscript in section 2 and put the table below in the Supplement material.
"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. Uncertainties in the representation of dynamical and chemical processes, emission inventories and meteorological input data typically limit the accuracy of calculated gas and aerosol concentrations (De Meij et al., 2012 and references therein). 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 material. More details of the different models are described in (https://confluence.ecmwf.int/display/CKB/CAMS+Regional%3A+European+air+quality+reanalyses+data+documentation)."

Supplement material
Table S1 Overview model characteristics

Model	Meteorological driver	Emissions	Boundary Conditions	Gas phase chemistry / Inorganic aerosols	Assimilated surface pollutants	Assimilation
Chimere	IFS, 3	CAMS-REG-AP	CAMS-	MELCHIOR 2 /	NO2, O3,	Kriging-

	hourly		Global IFS	ISORROPIA 2.1	PM2.5, PM10	based
DEHM	IFS, 3 hourly	CAMS-REG-AP	CAMS-Global IFS	Modified Strand and Hov (1994) / Frohn (2004)	NO2, CO, SO2 O3, PM2.5, PM10	Intermittent 3D-Var
EMEP	IFS, 3 hourly	CAMS-REG-AP	CAMS-Global IFS	EmChem19a / MARS (Binkowski and Shankar, 1995)	NO2, CO, SO2, O3, PM2.5, PM10	Intermittent 3D-Var
EURAD	IFS	CAMS-REG-AP	CAMS-Global IFS	RACM-MM/ Thermodynamic equilibrium (Friese and Ebel, 2010)	NO2, CO, SO2 O3, PM2.5, PM10	Intermittent 3D-Var
GEMAQ	IFS, 3 hourly	CAMS-REG-AP	CAMS-Global IFS	Modified ADOM IIB mechanism / Gong et al., (2003)	NO2, O3, PM2.5, PM10	Optimal Interpolatio n
Lotos Euros	IFS, 3 hourly	CAMS-REG-AP	CAMS-Global IFS	Modified CBM-IV / ISORROPIA-2	NO2, O3, PM2.5, PM10	Zhang (2001)
MOCAGE	IFS, 1 hourly	CAMS-REG-AP	CAMS-Global IFS + MOCAGE	RACM / ISORROPIA-2	NO2, O3, PM2.5, PM10	3D-Var
SILAM	IFS, 1 hourly	CAMS-REG-AP	CAMS-Global IFS	CBM-IV / Sofiev (2000)	NO2, O3, CO, SO2, PM2.5, PM10	Intermittent 3D-Var /

Binkowski, F. and Shankar, U.: The Regional Particulate Matter Model .1. Model description and preliminary results, J. Geophys. Res., 100, 26191–26209, 1995.

De Meij, A., Pozzer, A., Pringle, K. J., Tost, H., and Lelieveld, J., EMAC model evaluation and analysis of atmospheric aerosol properties and distribution, Atmos. Res., 114–115, 38–69, 2012.

Friese E, Ebel A. Temperature dependent thermodynamic model of the system H(+)-NH_4(+)-Na(+)-SO_4^{2-}-NO_3^--Cl^--H_2O. J Phys Chem A. 2010 Nov 4;114(43):11595-631. doi: 10.1021/jp101041j. PMID: 21504090.

Frohn, L. M.: A study of long-term high-resolution air pollution modelling, Ministry of the Environment, National Environmental Research Institute, Roskilde, Denmark, 444 pp., 2004.

Gong, S. L., Barrie, L. A., Blanchet, J.-P., von Salzen, K., Lohmann, U., Lesins, G., et al. (2003). Canadian aerosol module: A size-segregated simulation of atmospheric aerosol processes for climate and air quality models 1. Module development. Journal of Geophysical Research, 108(D1), 4007. https://doi.org/10.1029/2001JD002002.

Sofiev, M.: A model for the evaluation of long-term airborne pollution transport at regional and continental scales, Atmos. Environ., 34, 2481–2493, 2000.

Strand, A., and Hov, Ø.: A two-dimensional global study of tropo- spheric ozone production, J. Geophys. Res., 99, 22877–22895, 1994.

2. The manuscript allocates receptors to categories including background, urban, traffic and industry. Does the current classification fully capture the diversity of the environments? A clear definition of what each category (e.g., "traffic," "industry") represents is needed, along with *justification for why these specific categories were chosen. In my mind, urban areas often exhibit both traffic-related pollution and residential zones, what's the difference between "urban" and "traffic"? Does "traffic" mean receptors adjacent to road, while "urban" refers to receptors away from road but in urban residential area?*

Thank you for pointing out this important issue on station types.
The Air Quality Directive (2008/50/EC) and the new Ambient Air Quality Directive (Directive 2024/2881/EU) 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 analyze pollution trends and enforce regulatory limits. We use mostly the urban types to identify the most important behaviours in air pollutant concentrations. The reason for this is that we believe that there are more import differences between station types than station environments.
The key definitions are:
A **traffic station** is located in areas where pollution levels are significantly influenced by emissions from road traffic. These stations are typically placed:

- Near major roads, highways, or intersections.
- Where vehicle emissions (such as NO_2, PM10, PM2.5) dominate the air quality levels.
- In locations ensuring that they reflect the exposure of the population to pollution from road transport.

An **urban station** represents the overall air quality in an urban area without being directly affected by a specific pollution source like traffic or industrial emissions. These stations are:

- Located in residential, commercial, or mixed areas.
- Reflecting the exposure of the general urban population.
- Measuring background pollution levels influenced by a mix of sources.

Industrial stations are located near significant industrial sources, such as factories or power plants. The stations:

- Monitor emissions from industrial activities and their impact on surrounding areas.
- Typical pollutants: SO_2, NO_2, heavy metals, VOCs.

A **rural station** is placed in areas away from direct local pollution sources, representing regional air quality. These stations:

- Measure background pollution levels from natural and transboundary sources.
- Are located in the countryside or suburban areas far from significant emissions (e.g., cities, industrial areas, or major roads).
- Help assess long-range transport of pollutants.

The Air Quality Directives provides detailed criteria for air quality monitoring station. Below are the definitions and references to the relevant sections given:

1. Traffic Stations
These stations measure pollution primarily from road traffic and are located where the highest concentrations of pollutants due to traffic emissions are expected.

They should be at least 25 meters from major intersections but no more than 10 meters from the road. They must be positioned to represent the population's exposure to pollution from traffic.

2. Urban Background Stations
These stations measure general air quality in urban areas without direct influence from traffic or industry. They must be more than 50 meters away from major roads and more than 4 km away from industrial sources. Their purpose is to assess the average exposure of the urban population to air pollution.

3. Rural and Suburban Background Stations
These stations are located in areas with minimal direct pollution sources, representing the regional or background air quality. Rural stations are placed at least 20 km from urban areas and 5 km from industrial sources. Suburban stations can be closer to cities but should not be influenced by local sources.

We summarized the above information regarding the station types and added this to the manuscript in Section 2.

3. The current FAIRMODE MQOs considers four air pollutants including NO2, O3, PM2.5 and PM10, why don't the authors include more air pollutants such as SO2, CO, and PM2.5 chemical species? Additionally, the paper considers 8-hour maximum O3 values, how about 1-h max O3 peaks?
In this study we selected NO_2, O_3, $PM_{2.5}$ and PM_{10} to investigate the usefulness of the indicators. It is important to note that building a MQI for one pollutant and time aggregation requires information on the associated measurement uncertainty. This is not straightforward to obtain. This is why we focused on the four main pollutants and for each only considered one short and one long time aggregation. Work is currently ongoing to extend these MQI to additional pollutants and time averages.

4. Given the complexity of air quality modelling, including an uncertainty analysis or a discussion of the confidence in the model's predictions would be valuable. This would provide more insight into the reliability of the proposed indicators and how they could be applied in practice.
The Reviewer has a valid point.
We are not sure to understand your point but here is an explanation of what we try to achieve with our approach. Estimating the modelling uncertainty is almost impossible, as it would require a large number of model simulations where each parameter is modified independently. Given this difficulty, we assume in our approach that the modelling uncertainty is proportional to the measurement uncertainty. The more uncertain the measurement, the more flexibility we allow to the model results. This coefficient of proportionality is obviously challenging to fix. It should lead to a threshold that is sufficiently stringent to ensure sufficient quality but not too stringent that no model fulfills it. The sensitivity analysis consists in selecting a large number of model simulations and test them against different threshold levels to identify the relevant level of stringency. Our work constitutes one test in this context but more tests will be performed in future.

5. After introducing the new set of indicators, it would be helpful to provide a full table summarizing the complete set of MQO indicators. Comparisons with other well-established model performance indicators from different regions (e.g., the US, China, and India) are also necessary. This would provide a more comprehensive evaluation and context for the proposed indicators.

As suggested by the reviewer we included the use of model performance indicators applied in other regions in the world and placed this is a new section Discussion. We added the following to the section Discussion:

“”As mentioned earlier, indicators and the associated quality criteria are crucial for model evaluation, guiding improvements, and ensuring that the models can effectively inform air quality management strategies.

In the United States of America (USA), modeling guidance and performing evaluation was firstly introduced by the US Environmental Protection Agency (EPA) in 1991. Followed by introducing 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), Normalized Mean Bias (NMB), Normalized 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 like $PM_{2.5}$, NO_2 and O_3.

For O_3 modeling a model is considered acceptable if:

- NMB is within ±15%
- NME is ≤ 25%

For $PM_{2.5}$ the performance goals are:

- NMB within ±30%
- NME ≤ 50%

Also, 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. (2022) 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. (2022) methodology is based on Emery et al., (2017), applying goals and criteria for NMB, NME, FB, FE, IOA and R. Also, in that study recommendations are given to provide a better overview of model performance. For example, for PM2.5 the NMB should be within 10 % and 20 % and R should lay between 0.6 and 0.7 for hourly and daily PM2.5 and between 0.70 and 0.90 for monthly PM.2.5 concentration values, Also, different temporal resolutions for PM2.5 calculated values are introduced. Furthermore, benchmarks for speciated PM components (elemental/organic carbon, nitrate, sulphate and ammonium) were recommended.

“”

Model performance depends on the quality of the input data (e.g. emission and meteorology) 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, such as model resolution (e.g. De Meij et al., 2007, Wang et al., 2015), chemistry (Thunis et al., 2021a, Clappier et al., 2021), meteorology (De Meij et al., 2009 and references therein), emission inventories (Thunis et al., 2021b and references therein). Huang et al., (2022) showed that improving the spatial resolution improves the model performance, but further increasing the resolution (e.g. < 5km) would not improve the model performance skill in calculating e.g. PM2.5 concentrations. Changing the above-mentioned processes will impact the model performance, which could be investigated in the future. ""

Note that the goals and criteria proposed in the US or in 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 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).

- Mean Bias: Measures the average difference between modeled and observed values. A positive MB indicates overprediction, while a negative MB indicates underprediction.
- Normalized Mean Bias: A normalized version of MB to compare across different datasets.
- Mean Absolute Error: Represents the absolute difference between model and observations, helping to understand overall deviations.
- Root Mean Square Error: Quantifies the average magnitude of model errors, giving more weight to large deviations.
- Fractional Bias: Used in regulatory applications to evaluate whether a model consistently over- or underpredicts concentrations.
- Normalized Mean Error: Similar to NMB but considers absolute differences, preventing positive and negative errors from canceling out.
- Pearson Correlation Coefficient: Measures the linear relationship between modeled and observed values (ranges from -1 to 1).
- Index of Agreement (IOA): A normalized metric that evaluates how well the model reproduces variations in observations.

“”

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.

A. Clappier, P. Thunis, M. Beekmann, J.P. Putaud, A. de Meij, Impact of SOx, NOx and NH3 emission reductions on PM2.5 concentrations across Europe: Hints for future measure development, Environment International, Volume 156, 2021, ISSN 0160-4120, https://doi.org/10.1016/j.envint.2021.106699.

De Meij, A., S. Wagner, N. Gobron, P. Thunis, C. Cuvelier, F. Dentener, M. Schaap, 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., 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.

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 Ma., 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.

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: PM2.5 and chemical species, Atmos. Chem. Phys., 21, 2725–2743, https://doi.org/10.5194/acp-21-2725-2021, 2021.

Thunis, P., Clappier, A., Beekmann, M., Putaud, J. P., Cuvelier, C., Madrazo, J., and de Meij, A.: Non-linear response of PM2.5 to changes in NOx 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, 2021.

Thunis, P., Crippa, M., Cuvelier, C., Guizzardi, D., De Meij, A., Oreggioni, G., Pisoni, E.: Sensitivity of air quality modelling to different emission inventories: A case study over Europe, Atmos. Env., X, Vol. 10, 100111, ISSN 2590-1621, https://doi.org/10.1016/j.aeaoa.2021.100111; https://www.sciencedirect.com/science/article/pii/S2590162121000113, 2021b.

""

Specific Comments:
6. Please ensure that the use of subscripts and superscripts for air pollutants and other variables is consistent throughout the manuscript. For example, NO2 should be NO_2; µg/m3 should be µg/m^3.
Corrected.

7. On page 4, line 93, the MQO is first mentioned, but its definition is provided later in line 99. The abbreviations should be defined at the first time it appears.
Corrected.

8. I recommend adding a more detailed explanation of the variables used in each formula. Many variables in the manuscript are not clearly defined, which could lead to confusion for readers. A thorough description of each term will enhance the clarity of the model formulations.
A detailed description of each variable addressed in this study is provided in Janssen et al, (2022). This is also mentioned in the manuscript. We believe the descriptions of the variables are sufficient, keeping in mind that the goal of this work is to evaluate the usefulness of the variables. Detailing each variable would make the manuscript become unnecessary lengthy.

9. All tables in this manuscript are missing table numbers, titles or captions. Please provide clear titles for all tables to give context to the data being presented.
Corrected. Thank you.

10. Section titles with a single variable name (e.g., "NO_2") do not provide enough information about the content of the section. I suggest adding brief summaries to section titles to help readers understand the focus of each section.
Done.

11. In some radar plots, the brackets around serial numbers are partially obscured, and some incomplete solid lines extend outside the borders of other figures. These issues detract from the overall appearance of the figures and should be corrected to improve the presentation.
Corrected.

12. On page 8, line 207, the phrase "for Traffic, Industry, All and Background stations for Germany" is unclear. What is meant by "All stations"? Is this the sum of traffic, industry, and background stations? If so, why does Figure 2 show lower NO_2 concentrations at all stations compared to traffic stations? This requires further clarification.
Average of all station types considered. And the reason why the NO_2 concentrations are lower for "All stations", is that also the background concentrations are considered. Note that the number of stations for each station type (urban, traffic, industry) also differs, which affects the NO_2 concentrations when all stations are considered.

13. The font size within Figure 4 varies, which impacts the readability and visual quality. I recommend enlarging the font size to improve consistency and clarity.
We have corrected the font size to enhance the readability of the figure where applicable.

14. Line260, "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." suggests a direct causal relationship between model resolution and the short life of NO_2. This could be misleading; I recommend rephrasing to avoid suggesting that insufficient model resolution directly impacts the short lifetime of NO_2. The two phenomena are not causally linked in this manner.
As suggested by the reviewer we rephrased the sentence. It now reads as follows:
"The reason for this is that the model resolution is not fine enough to capture the traffic emissions. The short lifetime of NO_2 (about one hour) requires high model resolution to capture well the non-linear production and loss of NO_2 concentrations."

15. Line 277, the word "that" is duplicated in the sentence.
Corrected.

16. The Conclusion section primarily summarizes the findings but does not delve into a deeper discussion or implications of the results. I suggest expanding this section to discuss the broader implications of the proposed indicators, including how they could influence model evaluation in other regions or in future air quality studies.

Initially, FAIRMODE introduced a single model performance indicator, the MQI. While this indicator provides a relevant pass/fail test, passing the MQI does not necessarily guarantee that the modeling results are fit for purpose. To address this, additional indicators have been progressively introduced, particularly to assess how models capture temporal and spatial aspects. At this stage of evaluating the usefulness and relevance of these indicators, we analyzed five countries and three air pollutants to better determine whether a given indicator is useful and relevant for a specific pollutant. The methodology presented in this study will be applied to a broader range of air pollutants and countries in the future. Also, our methodology could be applied in other regions in the world where some model performance indicators are already used, like the EPA in the USA and in China to enhance the robustness of the modelling air quality results.

17. There are several typographical errors throughout the manuscript (e.g., "u" should be "µ"). A careful proofreading is required to correct these and improve the manuscript's overall quality.
Corrected.