We thank the reviewer 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 3

The authors expose the strategy developed in the framework of FAIRMODE in order to qualify the performance of model outputs, starting with the specific case of the CAMS modelling framework (for models), and the Airbase network for measurements. Specifically, they present a new set of indicators to evaluate more thoroughly the performance of air quality models in the framework of the European CAMS ensemble simulation. This new set of criteria permit to evaluate not only the general performance of the model in standard statistical fashion, but develops new metrics to focus on specific features such as the weekly, diurnal and seasonal cycle, and spatial differences between different station types..

The matter of this article (improving and complementing the set of criteria and metrics used to benchmark model performance) is of interest and seems timely. However, I have strong concerns. The bibliography of the article is almost non-existent (five studies are cited, including 4 by the same authors as this papers), highlighting the fact that the proposed method is not compared to other efforts in other countries, with other approaches. In my opinion, it is impossible to publish a research article without placing the work in the context of the international state-of-the-art.

Also, the performance criteria are based essentially on measurement uncertainty, a possibly interesting approach, very different from what is done elsewhere, but the authors do not discuss their efforts in light of other existing strategies, reducing the scientific interest of the paper. The authors spend most of the time in the manuscript to showcase the application of these new criteria for validation of model outputs on specific European countries (Spain, France, Italy, Germany and Poland), but without really discussing the methodological basis for these criteria, and how their criteria differ (or improve upon) other methodologies. In this respect, it seems to me that the present paper is designed more like an internal technical report rather than a scientific paper presenting criteria intended to be used by others, and compared to the production of others.

Therefore, I recommend rejection of this article. Since the matter is of interest, I recommend a new submission of a totally revised and reoriented manuscript focused on discussing the design of the criteria and placing the methodology of the authors in a wider context.

We accept that there is room for improvement, and inserted in the manuscript a new chapter Discussion, which provides an overview of previous work that apply Model Performance Indicators and Criteria in the USA and China. This has also led to the extension of the bibliography.
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. “”

The basis for the selection of 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 MQI. While this indicator provides 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.
For example, 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 e.g. PM2.5, making these additional indicators highly valuable.