

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 1**

### *General comments*

*This paper presents a well-structured study and methodology about to evaluate air quality modelling performance, with insightful analysis. The authors have addressed a systematic analysis of the FAIRMODE work related to this subject and have contributed to produce a reference scientific paper for future air quality modelling applications. However, while the research is robust, there are some minor and major points that should be addressed and corrected.*

### *Scientific questions/specific comments*

- *Abstract: The list of indicators tested should be identified previously, in particular the new "increment" indicators that were included in this study. It is not clear what type of indicators were used.*

We added to the text the following to explain better the type of indicators: "Model Quality (bias) and Model Performance (temporal and spatial) Indicators.

- *Lines 57-59: a reference is needed here to support this idea*

We added the following references to this part of the text:

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., Kadygrov, 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.

And

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%3A+European+air+quality+reanalyses+data+documentation>

- *Line 63: which indicators thresholds?*

A threshold is associated to each indicator and corresponds to a level of quality that we assume sufficient for the use of modelling to support policy. Since all indicators are normalized by a quantity

proportional to the measurement uncertainty, this threshold is one for all indicators. We removed this mention to threshold in the text as this was not needed there.

- *Line 73: Why focusing on “the following statistical parameters” – this should be explained*

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. We added this to the text.

- *Line 82: What is a “complete time series”?*

In our work, 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 AAQD (90%) to increase the available number of measurement stations for validation. We however impose that available data are representative of the full year. We added this to the text.

- *Line 96: The paper was submitted after the new AQDirective enter into force, so it should be mentioned*

Indeed, we now refer to the new AADD in the text.

- *Line 102: “temporal or spatial correlation”: it shouldn’t be “and” instead of “or”?*

The reviewer is right. Indeed, the spatial and temporal indicators are based on temporal or spatial correlation. One indicator is not based on both at the same time.

Thank you.

- *Page 6: stations are only considered in terms of influence, and what about environment type (urban, suburban, rural)?*

Thank you for pointing out this important issue on station types.

In our work, we follow the definitions provided in the Air Quality Directive (2008/50/EC) and the new Ambient Air Quality Directive (Directive 2024/2881/EU) of the European Commission. These definitions are given for different types of air quality monitoring stations based on their location and the pollution sources they are exposed to.

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 important differences between station types than station environments.

- Page 276: *(1-R) lower than 1 do not mean that "models are good for these indicators", only if it close to zero... and also, "good for these indicators" is not the most appropriate scientific expression to evaluate model skills*

We rephrased the sentence which now reads as follows:

The normalized 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 fulfill the objective. Values closer to zero indicate even better performances.

- *More general comment: the analysis of results (section 3) is mainly focused on the behaviour of each model in each country/area/pollutant, which I don't think is the main goal of the paper/study. Only in the conclusions section is discussed the main question: how the different indicators are useful and should be used for each different pollutant. The ideas that are written in the conclusions should be already addressed and presented before, during the analysis of the results - that should be always focused on the indicators.*

We agree with the reviewer. However, while the country analysis does not address the main question of the usefulness of the indicators, we nevertheless need this analysis to assess how these indicators behave across Europe. We went through the manuscript and tried to stress these points where relevant and useful.

*Technical corrections:*

- *Using different terminology ("Air Chemistry Transport Models (ACTMs)" and "Air quality models" can be confusing.*

We have made the corrections in the text where appropriate. Thank you.

- *The number of atoms in the chemical formula of the pollutants (NO<sub>2</sub>, O<sub>3</sub>) should be subscript*  
Corrected.

- *Line 50: Replace model by models on "More details on the model"*  
Done.

- *Line 58: 0.1x0.1 is approximately 11km (and not 10km)*

We disagree. The distance in kilometers between two longitude points differ when moving away from the North Pole, i.e. the distance is getting larger. Depending on the position on the Earth, the distance in kilometers varies. For example, 0.1 degrees longitude around Tromsø (Norway) is around 3.8 km, while 0.1 degrees longitude around Athens is ~8.9 km. The distance of 0.1 degrees longitude at the equator is around 11.1 km.

We removed "approx. 10 km" to avoid confusion.

- *Line 62: replace "calculated" by "simulated"*  
Done.

- *Line 90: add the symbol of the mean measured concentration "(U(O)):"*  
Done.

- *Page 3, Line 65: avoid the 2 words together "simulated calculated"*  
Corrected. Thank you.

- *Line 127: "relevant<sup>1</sup>"?*  
The number 1 refers to the first footnote in the manuscript. To make the meaning of 1 clearer, we have placed it as a superscript.

- *Page 5: tables are not numbered*  
Done, thank you.

- *figures in pdf do not show good quality*  
We've tried to improve the quality of the figures.

- *Line 376: sentences should never start with "While"*  
We disagree with the reviewer. The Cambridge Grammar of the English Language (Huddleston & Pullum, 2002) confirms that "while" can be used at the beginning of a sentence.