Review of egusphere-2025-2739 entitled “How accurate are operational dust models in predicting Particulate Matter (PM) levels in the Eastern Mediterranean Region? Insights from PM Surface Concentrations” by Andreas Eleftheriou et al.

General

The manuscript entitled “How accurate are operational dust models in predicting Particulate Matter (PM) levels in the Eastern Mediterranean Region? Insights from PM Surface Concentrations” by Andreas Eleftheriou et al. provides an assessment of the performance of eleven operational dust forecasting models and a multi-model ensemble through comparisons against surface PM measurements at three sites in the Eastern Mediterranean Region (EMR), Ayia Marina (AM) in Cyprus, Be’er Sheva (BS) in Israel and Finokalia (FKL) in Crete. The evaluation is done using specific established statistical metrics, namely correlation coefficient, R, Mean Bias, MB, and Root Mean Square Error, RMSE. The obtained results reveal a substantial variability in the models’ accuracy that no single model consistently achieves accurate predictions across all three regions and in all conditions (entire study period and days with high dust loadings).

The manuscript adds to the scientific community’s knowledge about the performance of operational dust models. Nowadays, a significant effort is made on developing and implementing such models to monitor dust levels in the atmosphere, and also warning the public about hazardous dust episodes, at regional or global scales. Given that these models differ between them in their spatial resolution, meteorological drivers, emission schemes or data assimilation procedures, it is important to intercompare them and to draw conclusions on which model(s) outperform. In this meaning, the study is interesting, although the conclusion drawn is not clear as to which model does so, in overall. While this is a bit disappointing, the study proves and convinces the reader that this happens given the multi-parametric problem, in the sense that many and combined factors play role and determine the overall model performance. While some questions remain unanswered (as explained below), it is more or less understood that a model can perform well at some site, while not in another, or better/worse under different dust loading conditions.

Based on the above, and the fact that the analysis is correct and complete at a significant level, while the text is well organized and written, I recommend publication of the manuscript subject to some corrections and recommendations for revision suggested below.

Main Comments

1. Further reference to the existing literature dealing with Mediterranean dust storms/episodes should be made. In particular, additional references should be made to papers dealing with increasing/decreasing desert dust storms in the Mediterranean Basin.
2. The style of discussion of the results obtained is a bit tiring for the reader. For example, discussing the results station by station in section 4.1 may get the reader

tired and lost in the details given. What matters is the comparative analysis, which is discussed in the last paragraph (section 4.1.1). This kind of discussion can be expanded referring to Figures 3 and 4. Also, Table 3 can be restructured to show rankings of the 11 models performance regarding the 3 different stations and the statistical metrics.

3. Improve and make more descriptive the captions of some Figures (e.g. Figures 9, 10, 11) to help the reader to more easily and rapidly understand what is shown.
4. The style of discussing results in section 5 should be improved by: (i) reporting the values of statistical metrics wherever recalled/reported in the text, (ii) specifying if the statements made with reference to the models' performance concern the entire study period or other conditions, e.g. the 95th percentile pf observed PM concentrations.
5. Provide arguments to infer responses to questions as to why a model, e.g. the NASA-GEOS, outperforms at AM and BS but not in FKL, or why a model, e.g. NOA-WRF outperforms during intense dust events, but not the same in overall.

Minor comments

1. Line 64: Yet, the satellite detection algorithms have progressed with time, including geostationary satellites as well (e.g. Kolios and Hatzianastassiou, 2019) exempt from the limitations of the polar orbiting satellite-based algorithms.
2. Line 93: the reference "Varga et al., 2014)" misses in the list of references.
3. Section 2: Dust storms in the MB and their spatiotemporal characteristics are better captured by satellites (e.g. Gkikas et al., 2013, 2016).
4. Line 110: The range of measured/exported particles' size should be given (it is necessary information to be used in the comparison with the corresponding sizes of the 11 models) since differences with models can partly explain the PM overestimations/underestimations by the models.
5. Line 131: replace "develop …" by "developed …".
6. Lines 133: It would be useful to indicate these dust storm days on Fig. 2, probably using different symbols for each station. Thus, readers can have an idea about the intensity of these dust storms at every site. Also, are there common days in the 3 stations out of the reported 106, 88 and 101 dust storm days?
7. Table 1: What kind of radiation interactions are those reported in this Table? Aerosol-radiation interactions or others as well, e.g. aerosol-cloud? Please specify.
8. Table 2: at what height is the first level above surface for the NASA_GEOS and NOA_WRF models, why are the corresponding values missing in the Table?
9. Line 167: use a parenthesis after the sum symbol, i.e. put the difference "M_i-O_i" in a parenthesis.
10. Line 179: relative bias and relative RMSE would be equally interesting metrics to show (as bias and RMSE).

11. Figure 3: Change “MEDIAN” to “MMM” for consistency with the rest of paper. Do the same in Figures 5 and 7.

12. Line 310: Figures 3 and 4 are not discussed and mentioned in the text, reference to them should be made.

13. Line 330: does “… R=0.62 …” should read “… R=0.55 …”?

14. Discussion in section 4.1.2: It is also interesting to discuss if models perform better or worse for high-dust events compared to all cases, also providing possible explanations for the improvement/deterioration. Some models, e.g. NOA_WRF, perform better for high-dust events in some locations (in FKL in this case) while doing worse in other locations (BS and AM), why does this happen?

15. End of section 4.1: A general assessment should made referring to whether the findings for the identified dust days by Achilleos et al. (2020) methodology are similar to those drawn from the high-dust days analysis of the previous section or not. It is essential to see if the methodology applied to identify dust events affects the results referring to the comparative model’s performance analysis (keeping in mind that basically they should not do so).

16. Discussion of Figures 9, 10 and 11: Make a comment on existing differences between Figures 10 and 11 since the nature of the results shown on these figures is about similar (as they both refer to days with high dust loadings).

17. Lines 502-505: This is a (probably the most) typical statement reflecting the complexity of the problem addressed, concerning the distinction of which is/are the model/models that perform better than others, overall. In spite of the differences existing between the models at various levels, they are all performing more or less similarly, perplexing the situation/problem.

18. Lines 508-510: Remove the empty line and link the text (from “and” to “Table 2”).

19. Section 5: The results show that different factors, e.g. vertical/horizontal resolution, height of first model level etc. are used as criteria for evaluating the performance of the models. Yet, it seems that some models are superior to others with respect to a specific factor, while other model(s) are superior with respect to other factor(s). Thus, it seems that a kind of counterbalance exists, leading to a roughly similar performance of the models. Would it be possible to draw, based on the overall performance of the models, a conclusion about which factor/factors is/are the most important for the model’s performance?

20. Lines 591-595: What is reported here is somewhat worrying. Why/how the performance of a model should change depending on the selected approach/method for the evaluation?

References

Gkikas, A., Hatzianastassiou, N., Mihalopoulos, N., Katsoulis, V., Kazadzis, S., Pey, J., Querol, X., and Torres, O.: The regime of intense desert dust episodes in the Mediterranean based on contemporary satellite observations and ground

measurements, Atmos. Chem. Phys., 13, 12135–12154, https://doi.org/10.5194/acp-13-12135-2013, 2013

Gkikas, A., Basart, S., Hatzianastassiou, N., Marinou, E., Amiridis, V., Kazadzis, S., Pey, J., Querol, X., Jorba, O., Gassó, S., and Baldasano, J. M.: Mediterranean intense desert dust outbreaks and their vertical structure based on remote sensing data, Atmos. Chem. Phys., 16, 8609–8642, https://doi.org/10.5194/acp-16-8609-2016, 2016

Kolios, S.; Hatzianastassiou, N. Quantitative Aerosol Optical Depth Detection during Dust Outbreaks from Meteosat Imagery Using an Artificial Neural Network Model. Remote Sens. 2019, 11, 1022. https://doi.org/10.3390/rs11091022