Reply to Referee 3

Paper: **Operational and Probabilistic Evaluation of AQMEII-4 Regional Scale Ozone Dry Deposition. Time to Harmonise Our LULC Masks**, by Ioannis Kioutsioukis et al., 2025, ACP

We are grateful to the reviewer for the thorough analysis of the manuscript and the careful reading and suggestions. They all greatly improved its quality.

*Reply* to the specific comments:

This authors indicate that the paper is focused on an evaluation of model estimates of ozone deposition but the main activity is comparing ozone concentrations (not deposition) and considering the various controlling processes. However, the paper is part of the broader AQMEII effort that includes companion papers that cover related model components, such as Clifton et al. 2023 that already describe model evaluations with direct measurements of ozone deposition and other papers on other trace gases, etc. so this manuscript ties this to model ozone concentration estimates. It would be helpful if this were explained early in the manuscript along with a summary of the results of the other papers and how it relates to this manuscript.

*Thank you for your comment. As a matter of fact, this aspect is explained extensively in the Introduction. The number one goal of this analysis is the operational evaluation which precedes the probabilistic and diagnostic ones (Dennis et al. 2010). As outlined in Dennis et al. (2010), before diving into the latter two it is important to evaluate the basics of the model performance (i.e. concentrations at minimum). This is to set the basis for the comparison of other variables that directly depend on these basic ones. This paper serves as reference for the subsequent probabilistic and diagnostic analysis performed in this paper but also for all the other papers that are part of the SI. We also treat quite extensively deposition in both the probabilistic and diagnostic evaluations. Clifton et al. (2023) deals with the specificity of the deposition modules used as 0D models for very detailed case studies and datasets. In contrast to this analysis of deposition modules in Clifton et al. (2023), the current manuscript analyses the full regional scale models that include implementations of these deposition modules. The real companion paper of this one, as detailed in the manuscript (end of the Introduction), is not Clifton et al (2023) (though a very relevant contribution to the special issue) but Hogrefe et al. (2025) where the diagnostic analysis of dry deposition in regional models introduced here is taken to a deeper level of consideration and detail.*

The paper mentions differences in gas-phase mechanisms but it is not clear how they interact with and influence deposition processes. A more in-depth discussion on the importance of these chemical scheme differences would be useful.

*The reference is in section 5.2. The connection is straight forward. If the chemical mechanisms are different, one may expect a different spatio-temporal determination of air concentrations which in turn affects dry deposition fluxes or totals being the former a driving component of the process. These differences are documented in detail in the original paper publications and in the present one it would be a diversion on the topic if they were discussed in this context.*

A major finding of this manuscript is that LULC is important. The authors point out the model implications of this (that LULC data should be accurate and consistent for all models) but they do not discuss the implications regarding the importance of different LULC (e.g., urban green spaces) for air pollution control. Recent papers suggest that urban green spaces may not be an efficient abatement measure for air pollution (e.g, Venter et al. 2024, doi.org/10.1073/pnas.230620012). The authors should consider whether their results provide any insights on this.

*Though we consider the topic raised by the reviewer relevant we also consider it out of the scope of this paper. Our focus is the continental scale where a large variety of LULC are present and are currently accounted for in a very inhomogeneous way across models. The grid resolution used by regional scale AQ models does not allow the detailed level of analysis that would be required to quantitatively assess the effects of green portions of urban settlements on air pollution. We are interested in the continent scale deposition and the effects of the missing representation of all LULC that characterise it on deposition.*

The manuscript emphasizes the importance of having the correct LULC but doesn't consider whether any of the current LULC schemes are adequate for characterizing ozone deposition. For example, are the ozone uptake capabilities of all evergreen needleleaf trees the same? If there is significant variability within a given LULC type, do these LULC schemes need to be modified to represent these differences?

*As the reviewer has certainly noticed, the level of sophistication of LULC and the way it is used in regional scale air quality models is far from being able to handle within-class variability. As the paper shows we are at the stage now of discovering that a surface characterization is not the same for all models and therefore still far from considering the in-species variability. In the light of the findings our aim is to stimulate the community to harmonize the surface characterizations in their models as they have comparable topographies for example. Until this fundamental step, that pertains to making sure that all models represent the same 'objective' surface land use and cover, adding inter-species variabilities would just contribute to piling up uncertainties that at one point will* *have to be disentangled. This is the first time that an analysis has been performed with such a level of breakdown of model variables. We acknowledge the validity of the reviewer question but unfortunately, we are far from meeting the minimum conditions necessary to include this next step of sophistication.*

The authors focus on evergreen needleleaf forests and briefly present results for other LULC types shown in the supplement. It would be useful to have more discussion of these other LULC types to show their differences/similarities. Even though there are fewer representative sites, it could still show the importance of differences in LULC types such as the range of ozone uptake capabilities.

*Indeed. However, the main scope of the paper is operational and incidentally the probabilistic and diagnostic analysis. As detailed first in the paper at the end of the Introduction and in several other sections, a detailed diagnostic analysis considering a variety of LULC types is performed in Hogrefe et al (2025, this issue) where the good point of the reviewer is extensively addressed.*

I recognize that different ozone units (ppb, ug/m3) are typically used in Europe and North America but for this exercise it would be better to be consistent and just use one. At least explain the rational if you don't want to do this.

*Indeed, we could convert them, at the same time only the macro differences between the two continental air sheds modelled are compared, whilst for the details the two cases stand alone. Furthermore, the measurements are provided with these units. According to us, sticking to the original units is the best way to preserve the integrity of data that are not under our direct control. Finally, the conversion from ppb to ug/m3 for O3 and NO2 is approx. 2 and for NO 1.25 (at 25 deg C) which are manageable conversion factors in case anyone would be interested in a detailed comparison. It should also be considered, as explained in the paper that: '* ... since ozone values are reported in ppb over NA and ug/m3 over EU, the range of the colour scales over both continents has been set such that the same colours represent the same absolute errors (note the difference in the numerical values for the colour bars for these figures), to account for unit differences and allow for a visual comparison between continents.' *An explanatory sentence of this choice has been added to the text in section 2.*

Why were those specific years chosen and why are they different in NA and Europe?

*As mentioned in Section 2., in the technical note Galmarini et al (2022) part of the SI, the description and technical details about the setup of the cases are presented. Therein also this question finds an answer. The technical note was prepared with the specific intention of grouping all this kind of information so that no space would be taken away in the other publications of the SI to explain and repeat details that are common to all. In this way more space is left for detailing the results of every specific research piece. Further explanation in given in Section 2.2 of Makar et al (2025) this issue.*

The criteria for "optimal" ensembles are based on minimizing RMSE, which does not capture all aspects of model skill, especially the ability to reproduce the maximum values that are a concern for air quality managers. There should be some discussion of the implications of this.

*The scope of the ensemble analysis is to go beyond the mean treatment of a set of models (as we have done in the operational analysis) and to determine the level of redundancy in the latter and the optimal combination of all available model results. The same analysis could have been done of the peak values, however the maximum value analysis is something that is of interest for regulators at local scale rather than continental or sub-continental one, since it determines the population exposures to peak pollution events. In this context it would have not been very meaningful and would have disrupted the logical sequence of the paper.*

Some figures, such as Figure 6, are difficult to see. Others, especially those displaying multiple model results (e.g., Figure 11), are challenging to interpret due to the amount of information. I appreciate the attempt to get all the information in one figure but perhaps clearer differentiation could enhance readability.

*Thank you for the comment. All figures were re-created in high resolution, adding different line styles where appropriate and inserting titles and missing units.*

Why do forest canopy shading effects increase NOx? (see Line 270)

*As explained in the paper: Model NA3 includes two forest canopy two effects. The first of these reduces the coefficients of vertical diffusivity in the region below the forest canopy. Gases emitted below the canopy (for example from surface emissions sources of NOx) thus have reduced turbulent mixing and hence may reach higher concentrations below the canopy. The second effect is the reduction in photolysis due to shading below the canopy. This changes the NOx chemical regime from more rapid NO2 photolysis, cycling between NO and NO2 and NO termination reactions (i.e. daytime NOx chemistry) to relatively low photolysis level chemistry (closer to nighttime, where NO2 titration of O3 dominates). The main effect on NOx is likely the turbulence part of this effect. The former line 270 has been modified to read, "Model NA3 includes a forest canopy parameterization (Makar et al., 2017), which takes into account reduced vertical coefficients of thermal diffusivity and photolysis levels below the forest canopy – these in turn reduce turbulent mixing (resulting in higher NOx concentrations from surface sources, and also shift the chemical regime from ozone production to ozone destruction by NOx titration below the forest canopy)."*