Reply to Referee 1

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:

Line 23: The abstract is missing information on why evaluation of ozone deposition is important for models/air quality.

Thank you for your suggestion the text has been modified accordingly.

Lines 48 – 52: The introduction would benefit from a brief description of the AQMEII4 activity and importance of deposition evaluation for this study.

Thank you for your suggestion the text has been modified accordingly.

Line 71: LULC acronym is introduced here without being defined.

Corrected thanks

Line 80: It is not clear why the NA and EU models are chosen for the years 2016 and 2010 for the evaluation. The text should make clearer why these years are considered for the analyses.

Table 1: Missing specifics on the model resolutions. NA and EU are described as regional scale models covering general domains but the specifics on model domain and resolution could be including here and within the text as well. Adding the model versions here might be helpful as well since this distinction is mentioned later in the text (i.e., Line 187).

Lines 83-88: It is not clear in the text which emissions are using for anthropogenic, fire, etc. missing relevant citations here.

This paper is part of an ACP special issues that contains several studies using the AQMEII4 model results and data. In order to avoid repeating for every other paper the description of the motivations of the activity and the specifications of the multi model exercise, all such information has been condensed in the Technical note: AQMEII4 Activity 1: evaluation of wet and dry deposition schemes as an integral part of regional-scale air quality models, by Galmarini et al. published with reference: ACP, 21, 15663–15697, 2021, https://doi.org/10.5194/acp-21-15663-2021. This editorial choice has been

made to avoid repetitions that would have reduced the publishing space for actual results. Therefore, all the answers to the above comments can be found in Galmarini et al. (2021) with a great level of detail. The note is cited in this paper whenever a reference to the case set up or a model's specifications is present. We understand that the early date of publication of the technical note (2021) may look like an independent publication totally disconnected from the current one, but as a matter of fact it is an integral part of the special issue. For further clarity this sentences has been added in the paper: "Details on the choice of years, model resolution and domains, model versions and choice of emissions input data may be found in Galmarini et al (2021)." and " Details on the model deposition parameterizations for ozone are found in Clifton et al (2024), and a discussion on the details of deposition for acidifying species can be found in Makar et al. (2025)"

Line 135: As this section is long, it may be useful to split this section into further subsections that discuss the results from the NA models, EU models, and comparisons.

Thank you, the section has been broken down into sub thematic sub-sections.

Lines 206-210: Reference?

Unfortunately it is not clear what reference is required here since the text relates to the finding of this study.

Line 250: It might be useful to discuss the seasonal and diurnal cycles underneath a separate subsection for clarity.

Done.

Line 320 – 324: The conclusions could be more clearly stated here. This section should be broken up into multiple sentences.

Done.

Lines 383 – 386: The phrasing here is unclear and should be restructured accordingly.

Done.

Line 688 – 692: Please reword for clarity.

Done.

Technical Corrections:

Figure 1: Images are blurry. The color bar is missing units to denote differences between the numerical amounts shown. Regions (i.e., R1, R2, R3) should be defined in the figure caption.

Units were added in the title of each subplot and the figure was exported in 300dpi. Moreover, the following sentence was introduced in the figure caption: "The rectangular areas represent the four selected sub-regions (R1, R2, R3, R4)".

Figure 2: The figure should have a title or color bar label indicating that RMSE is what is being shown. The image resolution is poor.

A title was inserted and the figure was exported in 300dpi.

Figure 3: Same as above, labeling the figure as MB or adding a color bar label.

A title was inserted and the figure was exported in 300dpi.

Figure 4/5: Same as above. Images have poor resolution.

A title was inserted and the figures were exported in 300dpi.

Figure 6: Resolution is low quality. Titles for (a) and (b) would help make this figure more readable.

A title was inserted and the figure was exported in 300dpi.

Figure 7/Figure 8: Image key should be placed outside of the figure for readability.

Titles were inserted, legend was placed outside and the figures were exported in 300dpi.

Figure 14/15: In the caption, the corresponding figure labels for wind speed, PBL height, solar radiation, and deposition velocity are missing.

The missing legends were inserted and the figures were exported in 300dpi.

Figure 17: Figure labels could be larger and /or boldened

The Figure labels here and in Figure S10 were updated to use bold font.

Reply to Referee 2

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:

One general comment about the figures, especially those for North America. I found it difficult to identify results for individual models. It would be helpful if a variety of line types (e.g., solid, long dash, short dash) and symbol types could be used for these figures in addition to different colors.

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.

Specific Comments

In the description of the surface ozone measurement data that were used in this study in Section 2 (lines 100-110), no information is provided about any data filtering that was applied before use, including how roadside monitors were treated. For example, Solazzo et al. (2012b) noted that in the AQMEII-1 operational evaluation they only used ozone measurement data from rural receptors below an altitude of 1000 m with at least 75% annual data availability. The present manuscript also notes better performance statistics in general for the EU case vs. the NA case, but this difference might be influenced by differences in characteristics of the measurement data sets that were used for the evaluation. For example, ozone statistics are typically quite different for urban stations vs. rural stations. Were the urban-rural splits for the EU ozone data set and the NA ozone data set similar?

A greater percentage of the land area in the European domain is urbanized or close to urbanized, while the North American domain contains large regions of uninhabited land. This may influence the results - for example, the gas phase chemical mechanisms for the models typically are based on high concentration smog chamber observations, and hence may provide better ozone formation and destruction and hence better evaluation than in remote areas. EMEP, AIRBASE and AIRS have similar definitions for the 'Rural' stations considered in this manuscript.

2. The text introducing Figures 2 to 10 at the beginning of Section 3.1 (lines 136-139) seems inconsistent and out of order. For example, Figures 2-5 show station-level RMSE and MB plots but not Figure 6, Figure 10 does not show box plots, and there is no reference to domain-level monthly and diurnal time series.

All figure numbering has been checked and corrected. The origin of the problem has been a last-minute request from the editorial office of Copernicus and a rushed implementation of the former from our side. Thank you for spotting this non secondary problem.

3. I found the discussion in lines 182-197 of the similarities and differences between the three WRF-Chem versions that were run for North America a bit hard to follow. First, there is a reference to Figure 5, which shows results for the European simulations. Then the three different versions of WRF-Chem that were used for North American simulations are described, but this is followed by references to "the former" and "the latter", implying that only two model versions are being discussed. Then the discussion turns to "relatively minor differences" and "larger differences" without stating differences in which quantity: model configuration choice or RMSE or MB or something else?

A less convoluted wording of this paragraph is now presented with the right figure references which should clarify the statements.

4. The discussion of Figure 4 (lines 198-212) states that model EU3 has the best results while EU1 and EU4 have less good results north of the Mediterranean area. To my eye, though, model EU2 had worse RMSE scores than the other three models over Germany, Poland, and Hungary, and it also has the highest median RMSE value in Fig. 6,b but EU2 is left out of the discussion. Am I misinterpreting these figures?

Your interpretation is correct, your analysis has been included for the sake of making it more precise and detailed. Thanks.

Very late in the paper another significant difference between NA6/NA8 and NA7 is mentioned (lines 741-743)

That is true but we consider those explanations best placed at that level of the paper.

5. There are multiple discussions in the manuscript concerning the differences between two of the models used for the North American simulations: GEM-MACH(Base) [or NA3] and GEM-MACH(Ops) [or NA5] (see l. 177-181, l. 222-232, l. 269-278, l. 294-297, l. 411-413, l. 612-619, and l. 745-748). These discussions are somewhat hard to follow because the differences are mentioned incrementally wth wide gaps between mentions: (a) use/non-use of a canopy shading and turbulence scheme (l. 179) and different treatments of area-source emissions injection (l. 181); (b) use/non-use of a vehicle-induced turbulence scheme (l. 271); (c) full feedback vs. no feedback of meteorology on chemistry (l. 412); and (d) different treatments of seasonality of LAI (l. 747). These

differences are then summed up on line 616, where it is stated that GEM-MACH(Base) uses "very different physical parameterizations than" GEM-MACH(Ops). This seems like an overstatement; isn't it more of a case that GEM-MACH(Base) uses two additional physical parameterizations compared to GEM-MACH(Ops)? And despite these differences, O3, NO, and NO2 predictions made by these two model versions have very similar scores (e.g., Figs. 2, 3, 6, 7, 9, S1, S2, S3, S5). It thus appears somewhat unbalanced that there is considerable discussion of GEM-MACH(Base) and its forest canopy and vehicle-induced turbulence (VIT) parameterizations, even though GEM-MACH(Ops), which does not employ either the canopy or VIT parameterizations, had comparable scores and was the most frequently included member of high-performing ensembles for NA (Table 2).

The full list of differences is larger than the reviewer suggested – there are five differences in process representation between the two models – we've clarified this in the revised manuscript, with the entry at former line 177-181 becoming:

"The relatively low MB for models NA3 and NA5 reflect the use of a similar deposition velocity algorithm, while differences between these two models reflect the use of process representations in NA3 which are absent in NA5 (for canopy vertical turbulence different approaches for canopy vertical mixing and photolysis (Makar et al., 2017), feedbacks between chemistry and meteorology (Makar et al., 2015a,b), vehicle-induced turbulence (Makar et al., 2021) and satellite derived leaf area index (Zhang et al., 2020), while NA5 makes use of a simplified means of adding surface emissions in the model which assumes that fresh emissions are evenly mixed into the first two model layers).

The intent of the other sections was to help explain possible differences despite the same deposition code being used, not to focus on the NA3 parameterizations as such. We've tried to clarify this as well as better summarize these points as follows

Lines 222-232 has been shortened and made more to the point (that the physical locations of positive ozone biases in many of the models compared here correspond spatially to forest canopy locations, and that a methodology developed to represent this process has had some success in improving O3 performance in both the GEM-MACH and CMAQ models):

"Some of these biases may be attributable to the need for physical process representation for forest canopy shading and turbulence (see Makar et al., 2017, which intercompares multiple models), and has been found more recently to improve the performance of the CMAQ model (Campbell et al., 2021, Wang et al., 2025). Many of the regions with the highest ozone biases in models EU1, EU2 and EU4 correspond to areas with high forest canopy and leaf area index values, as does the eastern seaboard of the USA and Canada, and the negative biases in EU1 and EU4 for NO and NO_2 are consistent with the absence of the more realistic reduction in thermal diffusivity coefficients and

photolysis rates expected under forest canopies (Makar et al., 2017); the performance of these models may be improved through the inclusion of forest canopy processes."

Line 269 – 278 has been simplified to:

"As noted above, Models NA3 and NA5 include process represention which can enhance the vertical transport of freshly emitted NOx out of the lowest model layer; at least some of superior performance may be related to this faster dispersion."

Lines 294 to 297 in the original manuscript have been simplified to:

"As noted above, process representation of forest canopy shading and turbulence is one such such possible means of model performance improvement[1]."

Line 411-413: The line has been changed to :

"The difference in soil versus cuticle terms dominating in NA3 and NA5 likely reflects differences in meteorology between these two model implementations; as noted above, NA3 includes feedbacks between meteorology and chemistry, in turn resulting in differences in the meteorological terms controlling these two deposition pathways."

Line 612-619 has been changed to:

"Five process representation differences between NA3 and NA5 have been summarized above – one of these is different driving meteorology, which may influence differences between these two models in Figure 11."

Lines 745-748 have been modified to read:

"We note that the differences noted above for NA3 versus NA5 include different LAI information (with different sources and seasonal dependance)."

6. I found it striking how the observed monthly O3 profile differs in NA in the autumn between regions R1 and R2-R4 (Figure 9). This is not true in EU (Figure 10), where the observed monthly O3 profiles are similar across the four regions (and also with NA R1). The discussion of Figure 9 (lines 337-349) notes this behavior indirectly, but refers to the model overpredictions rather than the sharp reduction in observed concentration.

We are actually not quite sure what "autumn" differences the reviewer is referring to. The September - November period looks similar between R1 and R2-R4, with decreasing trends in all regions. It is difficult to tell what the causes of these differences might be, but we can speculate. We note that the observed NA time series (Figure 9) for regions R2, R3, R4 have O_3 peaks in April, R1 peaks in June, in contrast to the EU time series (Figure 10) which peak in June to July. The key difference is the April peak in the R2, R3, R4 observed values – these regions are all downwind of North America's western cordillera mountains, and the April peak may represent the impacts of downmixing of upper Tropospheric air along the eastern side of the mountains, an effect which is known to maximize in the

springtime (see for example, Pendlebury et al., https://doi.org/10.1016/j.atmosenv.2017.10.052, 2018).

We have added the following lines to the discussion on Figure 9: "We also note that the time series of observed O_3 for North America shows April peaks for regions R2, R3 and R4, while R1 peaks in June. One possible cause for the observed early spring peak in the latter regions is the transport of upper Tropospheric O3 downwind of the western cordillera, a process which is known to be at its maximum in the springtime (Pendlebury et al., 2018)."

Note too that Figure S3 is not referred to in the manuscript unlike other figures in the Supplement, but there are similarly significant regional differences in monthly NO profiles for the NA1 vs. NA2-NA4 subregions but not the EU1-EU4 subregions. And for the sentence that begins "The same behavior observed" (lines 347-348), should a parenthetical "(not shown)" be appended to this sentence since I don't think there are supporting figures provided.

Corrected, thanks

7. I think this manuscript only presents quantitative analyses of ozone concentrations and ozone dry deposition flux. Can lines 494-495 be reworded to remove the reference in line 494 to "deposition velocity performance for NA3, NA5 here"?

Thanks, but the text mentioned in your comment refers to what is discussed in Makar et al. (2024) and Clifton et al. (2023) not this paper.

8. In Section 4 the number of possible ensembles in line 472 for NA (254) and in line 498 for EU (18) don't seem correct. Based on the sum of rows of Pascal's triangle, shouldn't these numbers be 255 and 15? Also, in line 499 since we are excluding 4C1 (i.e., 4C2 + 4C3 + 4C4) it should say "Four out of the 11 combinations of ...". And for Table 3 the color scheme appears to be inconsistent with Table 2 -- shouldn't three columns (one 2nd order, one 3rd order, and the 4th order) be colored orange?

You are right the right numbers have been inserted

9. Would it be possible to recheck Eqns. 2b and 3 in Section 5.1? Just quickly looking at Eqns. 1 to 3 as a whole, my sense was that f/2 should be e/2 in Eqn. 2b and [a] should be [c] in Eqn. 3, but I stand to be corrected. And to help the reader, line 565 could be expanded to state something like "where the shared fraction of variation explained by X1 and X2 is (similarly for e and f)".

Thank you for the comment. The right figure S7 was now added in the supplementary which makes the interpretation of the figures easier.

10. Section 5 has two parts, a subsection on ozone deposition flux variability and a subsection on ozone concentration variability. However, the text frequently refers just to

"ozone flux" and "ozone variability". Using more exact terminology would be helpful to the reader in Sections 5 and 6. A similar comment applies to the use of "deposition" instead of "dry deposition" throughout the manuscript.

More care has been put in the use of flux, concentration and dry deposition

11. I have two minor concerns about the discussion of Figure 12 on pages 20-21. First, the comment that "the picture changes completely from NA" seems too strong to me, especially since the discussion that follows is much more nuanced. My sense is that the WRF-Chem results in Figure 12 (EU1, EU2) are in fact broadly similar to those in Figure 11 (NA6-NA8). The one big qualitative difference is the very small contribution from stomatal effective flux in the three winter months for EU1 and EU2 (which is different from NA6). My second minor concern is about terminology. The term "meteorological driver" generally suggests a numerical weather prediction model like WRF whose inputs are fed to a chemical transport model. Wouldn't "driving meteorology" or just "meteorology" be better here? I do wonder too whether the implementation of the seasonal dependence of stomatal conductance for these European simulations isn't also a contributing factor (cf. line 213), especially when compared to the NA6-NA8 results in Figure 11?

We agree with your assessment, these aspects have been corrected thank you.

12. Following the discussion of Figure 17 in lines 789-800, there is no discussion about the corresponding LULC statistics for Europe that are presented in Figure S10. Moreover, Figure S10 is incomplete -- either the top or bottom panel is missing, and it would also be helpful to have panel labels similar to Figure 17. In addition, the Abstract mentions an "introductory diagnostic evaluation" but Section 5 does not mention this aspect.

Thank you for pointing out that the top portion of Figure S10 (LULC distributions across all non-water grid cells in the common EU domain) was missing, this has been corrected in the revised submission.

13. One important finding noted on page 27 of the Conclusions (line 848) is that the predominant LULC types at ozone receptor locations are {\vf LULC types for which deposition is relatively low}". This is a second factor to explain the similarities in VP of ozone concentration variability shown in Figures 15b and 16, but this finding does not appear to be mentioned in Section 5.2 in the discussion of Figures 17 and S10.

As a matter of fact these aspects are tackled in section 5.2 has shown in the text here: '' This result also confirms the hypothesis made at (3) in the previous section; the operational ozone monitoring sites are not suitable for the analysis of deposition results for specific LULC classes. A similar conclusion can be drawn for the EU case (Figure 12b)

which is presented back-to-back with the evergreen needle-leaf forest case. To corroborate the last statement, Figure 17 shows a comparison of the fraction of the entire NA common domain (excluding grid cells dominated by water, i.e. water fraction > 0.5) covered by each LU type to the LU distribution of all grid cells corresponding to O_3 receptor locations (EU results are shown as Figure S10 in the SM). As can be noticed, existing O_3 receptor locations are characterised mainly by Planted/Cultivated, Shrub land and urban LULC with a 10% coverage of deciduous broadleaf forest (Figure 17b). At these locations all models appear to have the same distribution of the main LULC type apart from Shrubland (NA3, 4 and 5 20% more abundant) and Planted/Cultivated (same models 10 % less abundant). However, the distribution of LULC from the overall NA common model domain (Figure 17a) demonstrates that the current receptor site LULC poorly represent the relative amount of land use occurring throughout the domain, with, for example, much higher Evergreen Needleleaf and Grassland fractions, and much lower urban land use LULC in the all-domain data of Figure 17a compared to the observing station values of Figure 17b.''

14. One important finding noted by the authors in Section 3.1 (lines 278-294) that in my opinion represents an important conclusion for the entire AQMEII4 project is not mentioned at all in the Conclusion section. This is that the two models whose ozone concentration predictions for North America were found by the operational evaluation to have the highest skill were previously found by Activity 2 of AQMEII4 to have dry deposition modules with larger errors than other models, thus suggesting that their grid-scale performance was almost certainly due to error compensation between now-known errors in their dry deposition module and compensating errors in one or more process representations and that the other models applied for North America likely also had these same unknown errors, which resulted in their less good operational performance for ozone concentration due to their smaller errors for representation of dry deposition.

While the two North American domain models GEM-MACH-Basecase and GEM-MACH-Ops did have a common error associated with the dry deposition module employed, it would be more accurate to say that this likely contributed to their grid scale performance rather than that the performance was solely due to the effect of this error. This is under investigation by the GEM-MACH modelling group and they hope to have a publication submitted to ACPD on the correction to the model results and its influence on model performance later this summer. The cause of that difference has since been tracked down to misinterpretations of LAI dependence from different literature sources in the deposition code implementation in those model versions (see the footnote in the original

and in the revised manuscript, section 3.1.3). However, the reviewer is making an unwarranted assumption that this error extends to other models in the ensemble -this is simply not the case. At the same time, we agree that the potential for compensating errors (in a more general sense) is always possible within an air-quality model, and other examples appear in the literature (c.f. Makar et al., 2014: https://gmd.copernicus.org/articles/7/1001/2014/).

15. The manuscript title mentions both operational and probabilistic evaluations, the Abstract mentions operational, probabilistic, and diagnostic evaluations, but the Conclusions section only mentions the operational evaluation directly. Perhaps findings for the other two evaluation types could be added in the Conclusions section and "deposition-focused model evaluation" could be referred to in this section as a type of diagnostic evaluation.

You are right, the title has been corrected as well as the text.

16. The References section needs some attention:

All suggestions about the references have been corrected. Thank you for taking the time to check those out.

- Hogrefe et al. (2023) is not cited in the manuscript but there is a reference;

- Kioutsioukis et al. (2014) is cited three times in the manuscript but there is no matching reference; however, Kioutsioukis and Galmarini (2014) is not cited at all but there is a reference;

- Solazzo et al. (2015) is cited two times in the manuscript but there is no matching reference; however, Solazzo et al. (2013) is not cited at all but there are two references for Solazzo et al. (2013);

- The Makar et al. (2024) reference should be updated to Makar et al. (2025) (https://acp.copernicus.org/articles/25/3049/2025/);

- Would the Campbell et al. (2022) journal publication be a better reference than the Campbell et al. (2021) conference presentation?

[Campbell, P. C., Tang, Y., Lee, P., Baker, B., Tong, D., Saylor, R., Stein, A., Huang, J., Huang, H.-C., Strobach, E., McQueen, J., Pan, L., Stajner, I., Sims, J., Tirado-Delgado, J., Jung, Y., Yang, F., Spero, T. L., and Gilliam, R. C.: Development and evaluation of an advanced National Air Quality Forecasting Capability using the NOAA Global Forecast

System version 16, Geosci. Model Dev., 15, 3281–3313, https://doi.org/10.5194/gmd-15-3281-2022, 2022.]

17. The "Contributions" section (lines 880-886) also needs some attention. First, in what ways did KM and JP contribute? Second, should "WRF-Chem(IASS)" be "WRF-Chem(RIFS)" and should YHC be YHR? Perhaps you could also refer here to "WRF/CMAQ-M3Dry" and "WRF/CMAQ-STAGE" to be consistent with Table 1.

Technical and Editorial Corrections/Suggestions

Thanks for the detail analysis, corrections and suggestions. All corrections have been implemented all suggestions have been considered. Some specific clarification is included hereafter.

p. 1, l. 27 \ \ Perhaps "The collective evaluation begins with an operational evaluation, namely a direct comparison of model-simulated predictions with monitoring data aiming at assessing model performance (Dennis et al., 2010)."

p. 1, l. 37 \ \ "... as a variable to be ..."

p. 2, l. 50, 52 \ \ Perhaps "dry deposition process modelling" and "standalone dry deposition modules" -- Galmarini et al. (2021) mentions wet deposition as a focus of AQMEII4 but that doesn't seem relevant for this manuscript.

p. 2, l. 63 \ \ Perhaps "The operational evaluation also provides important context ..."

p. 2, l. 72 \ \ Perhaps "... of modelled ozone dry deposition fluxes and velocities can be found ..."

p. 3, l. 108 \ \ Perhaps "For the European case the monitoring network databases employed included:"

p. 4, l. 117 \ \ Perhaps "... and the yearly average measured ozone at these sites"

p. 4, l. 123 \ \ Re "greater activity density" - I am not sure what is meant here by activity density

p. 4, l. 127 \ \ "... less detail, and ..."

p. 5, l. 147 \ \ Change "from the Figure 2-5" to "from Figure 2"

p. 5, l. 158 \ \ Delete "(see also Figure 3" -- superfluous?

p. 5, l. 168-173 \ \ Very long sentence?

p. 7, l. 224 \ \ Wang et al. (2025)?

p. 7, l. 233 \ \ Normalized mean bias?

p. 7, l. 235 \ \ Instead of "notice", would "note" be more appropriate? Same comment for lines 582, 646, 726, and 801.

p. 8, l. 238 \ \ Perhaps "... have ozone bias values closest to zero, followed ..."

p. 8, l. 240 \ \ From inspection of Fig. S1 I see three models with medium NRMSE values and only one with a high value.

p. 8, l. 250 \ \ Perhaps "... observed and modelled seasonal and diurnal cycles for North America for ozone, NO and NO2"

p. 9, l. 277 \ \ Perhaps "transported upwards away from the model surface"

p. 10, l. 314-315 \ \ The fact that models with smallest NO and NO2 biases do quite well for NO and NO2 shouldn't be surprising.

p. 11, l. 334 \ \ "... confidence in mobile and stack emissions, which ..."

p. 11, l. 338 \ \ I think reference here should be to Figures 9 and 10 rather than Figures 5 and 6.

p. 11, l. 340 \ \ "regions"

p. 11, l. 351-352 \ \ It is probably obvious from manuscript context but why not insert the word "dry" here, as in "ozone dry deposition fluxes"?

p. 11, l. 355 \ \ Perhaps "is not only due to these resistances but also"

p. 12, l. 368 \ \ "in which"

p. 12, l. 374-375 \ \ How many EU grid cells for "Deciduous Broadleaf Forest", and which continent for "Mixed Forest" and "Urban" values and what about the other continent? *Results not shown explanation provided*

p. 12, l. 380 \ \ Are 6130 cells or 6108 cells really "very few"? Perhaps "relatively few".

p. 13, l. 404-406 \ \ Double negative: keep "not" and change "neither ... nor" to "either ... or"

p. 13, l. 417 \ \ It is not clear to me what the "rule of model differences" is.

p. 14, l. 426-427 \ \ "but sometimes there is contributing seasonality in non-stomatal flux" -- awkward wording

p. 18, l. 561 \ \ "in this case ozone deposition flux"

p. 19, l. 590 \ \ Perhaps "... show that different models have very different ..."

p. 19, l. 596 \ \ Perhaps "... ozone flux variability is more equally distributed across ..."

p. 20, l. 619 \ \ Is "vehicles on highways" strictly true? From a quick perusal of Makar et al. (2021) it appears that the parameterization is based on VKT on roadways in general, not just highways.

Specifically, the parameterization is used for "on-road mobile source emissions"; the text has been updated to include this phrase.

p. 20, l. 629 \ \ "WRF-Chem"

p. 21, l. 638 \ \ Change "S8-S10" to "S7-S9"

p. 21, l. 645 \ \ Change "9" to "13"

p. 21, l. 649-650 \ \ Change "8" to "12"

p. 21, l. 650, 788 \ \ Perhaps change "back-to-back to" to "side-by-side with"

p. 21, l. 653 \ \ Perhaps "... stomatal flux, though disagreeing on the exact ..."

p. 24, l. 731 \ \ "solar radiation"

p. 24, l. 736 \ \ "wind speed"

p. 24, l. 737 \ \ "... though contributing on average 30% of the resolved varibility"?

p. 24, l. 738-739 \ \ Why this particular order of models: NA2, NA6, NA7, NA4, NA1?

p. 25, l. 764-765 \ \ "emissions variability" would be better

p. 25, l. 762, 767 \ \ "Interesting is the ..." -- awkward wording

p. 25, l. 778-779 \ \ Change "12" to "16" and "11b" to "15b"

p. 25, l. 782-783 \ \ "NA1, NA2, NA3, and NA5" would be more consistent with rest of manuscript. Should NA8 be included in this list?

p. 25, l. 784-785 \ \ "... contributor to ozone concentration variability at receptor locations"

p. 25, l. 787 \ \ Change "12b" to "15b"

p. 26, l. 802-804 \ \ To support this statement, could the following addition be made: "..., in conditions of uniform LU characteristics and dominance of urban and Planted/Cultivated LULC types, as shown in Figures 15b and 16 the models tend to produce comparable results in terms of contributors to ozone variability"

p. 26, l. 817-818 \ \ "... performance of dry deposition schemes ...", "Ozone dry deposition, in particular, ..."

p. 27, l. 821 \ \ One EU model has NRMSE for NO2 of 35% (Fig. S1)

p. 27, l. 834 \ \ "over North America"

p. 32, l. 988 \ \ For Hogrefe et al. (2025) reference, change "2005" to "2025".

p. 34, l. 1056 \ \ Separate these two references.

p. 38, l. 1131 \ \ Should "orange" be "yellow"?

p. 50, l. 1192 \ \ "columns"

p. 52, l. 1210 \ \ "Same as Fig. 11 but at the locations of ..."?

p. 52, l. 1196-1197 \ \ Is this sentence necessary?

Yes it is according to us, thank you

p. 55, l. 1226 \ \ Should be "Same as Fig. 14 but at ..."

Figs. 1-4 captions \ \ Add units of MB and RMSE to captions.

Units are there

Figs. 7-8 captions \ \ Should note that time units are local time.

Fig. 9 vs. Fig. 10 \ \ Labels of former are smaller and harder to read

Figs. 7-8 captions \ \ Should note that time units are local time.

Fig. S1 caption \\\ "Figure S1: Soccer plot diagrams of O3, NO2, NO2 and NO." would better reflect the figure layout.

Fig. S3 \ \ There seem to be two NO panels for the NA case.

Figs. S3 and S4 \ \ Could the line thicknesses in the legends be increased so that the line color is easier to see?

Different line styles were used for similar colors and the figures were exported in 300dpi.

Figs. S5 and S6 \ \ What are units of O3FLX? Perhaps "Monthly ozone dry deposition flux". The resolution of these figures is too low -- it is very difficult to see the line colors of the legend.

The figures were exported in 300dpi and O3FLX units were inserted in the caption:

"... Ozone dry deposition flux (O3FLX) in g/ha, ..."

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)."

Reply to Referee 4

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:

While the manuscript presents a detailed and robust analysis highlighting the inconsistencies in the LULC masks and their implications for ozone dry deposition modelling, it has become extremely long and dense. Several sections, particularly operational evaluation, are notably verbose, with extensive model-by-model commentary that could be streamlined. I recommend that authors condense or relocate some of these detailed discussions to the supplements, especially where the text reiterates statistical patterns already evident in the figures. Additionally, the manuscript would benefit from clearer synthesis or summary paragraphs at the end of each major section to reinforce the key takeaways and guide the reader through the analysis.

We have condensed slightly the section following the reviewer suggestions; however, we consider the details in it important for the scope of the paper as they are there to document the model performances that are needed not only of the current study but also all the others present in the special issue. Following a suggestion of Reviewer 1 the section 'operational evaluation' has been broken up in subsections. Thank you for your comment.

Line 83-88: What are the emission inventories used for lightning NOx, forest fires, biogenic emissions, and other natural sources? Please provide a brief discussion on the internal model processing of these emissions.

As explained in the introduction all the information relating to the case studies and the common input to all models have been summarized in the technical note Galmarini et al. (2021). This editorial choice offers a one stop-shop for all necessary information and avoids repeating it in the various contributions to the special issues, which would take away space for actual research items. Briefly, the lightning NOx as well as forest fire emissions were harmonized across all models with more details provided in Galmarini et al. (2021) while the representation of biogenic and other natural emissions was decided by each modelling group. The following sentences have been added fo further clarity: 'Details on the choice of years, model resolution and domains, model versions and choice of emissions input data may be found in Galmarini et al (2021). Details on the model deposition parameterizations for ozone are found in Clifton et al (2024), and a

discussion on the details of deposition for acidifying species can be found in Makar et al. (2025).'

Line 89-90: Please provide a reasoning behind choosing the spatial and temporal domains for this study. What are the spatial and temporal resolutions used in the model runs?

This information is contained in Galmarini et al. (2021).

Line 136-139: Ozone values in the figures are reported in ppb over North America, while in µg/m3 over Europe. The use of these two different units for ozone concentration complicates the direct comparison. While the authors attempted to align the color scales, I recommend standardizing the units across both regions for the convenience of readers and easier comparisons.

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.*

Line 161: Figure 4 presents results for Europe, which is incorrectly referred to as showing results for North America. I recommend that authors carefully review the entire manuscript to identify and correct such figure and section reference inconsistencies. While minor, these errors can significantly impact the clarity and interpretation of the results and may confuse readers.

Thank you, we noticed this too and corrected it accordingly.

Line 190-194: The authors hypothesize that the relatively minor difference between WRF-Chem (UPM) and WRF-Chem (UCAR) is primarily due to the difference in gas-phase chemistry mechanisms. However, this claim is made without presenting supporting analysis or citing a reference that explicitly evaluated the gas-phase chemical mechanisms used in these two configurations. I recommend that the authors provide a reference to a past work supporting this hypothesis or rephrase this discussion as a hypothesis.

A good point – we should have included a reference in the revised manuscript (Knote et al., 2015). The impact of the two mechanisms on model performance was evaluated in a previous AQMEII ensemble cycle (AQMEII2) by Knote et al., Atm. Env., 115, 553-568, 2015. Figure 3 of this paper shows that the use of MOZART4 versus CBMZ resulted in opposite signs and magnitudes for O3 biases in North America. We've broken the original sentence into two parts, and have inserted the following sentence between them: "Knote et al. (2015) conducted a comparison of the two gas-phase mechanisms (CBMZ and MOZART4) within the same modelling framework, and showed that two mechanisms to have biases opposing in both magnitude and sign over North America)."

Line 308-311: What are the factors driving the underestimation of wintertime NOx?

While we can't provide a certain answer, we can speculate – we have included the following sentence in the revised manuscript:

"Potential factors which might drive an underestimate of wintertime NOx include underestimates in the emissions of NOx from combustion sources such as wintertime home heating from fossil or wood fuels (van der Gon et al., 2015), underestimates of atmospheric stability (i.e. if the simulated atmosphere is more unstable than the actual atmosphere, NOx emissions may build up to higher concentrations in the model than is observed), and the potential for HONO cycling in the presence of snow on surface leading to longer lifetimes of NOx (Michaud et al., 2015)."

Line 390-393: Are there any implications of combining LCAN and soil for models that distinguish these two terms?

The reviewer has raised a good question. The LCAN term is relatively small - it is sufficiently small that some deposition algorithms leave it out altogether (Clifton et al., 2024) - hence its inclusion with the ground resistance is unlikely to influence O3 deposition significantly, and the ground resistance term is unlikely to be significantly influenced by being combined with the LCAN term.

Line 529-534: The authors identify factors that are expected to be relevant in the determination of ozone concentration variability at the surface but fail to discuss the methodology used in identifying these factors. I recommend that authors briefly discuss the methods used to identify these factors, either in the main text or the supplement.

As a matter of fact, the methodology is fully explained also in mathematical terms in the paper, so it is hard to understand what specifically the reviewer is referring too. The same set of four variables has been selected simply to represent the core mechanisms of O3 fate in the atmosphere (convection, advection, dry deposition, photochemistry) in different domains, months and LULC classes.