

Operational, Diagnostic and Probabilistic Evaluation of AQMEII-4 Regional Scale Ozone Dry deposition. Time to Harmonise Our LULC Masks

Ioannis Kioutsioukis¹, Christian Hogrefe², Paul A. Makar³, Ummugulsum Alyuz⁴, Jesse O. Bash², Roberto Bellasio⁵, Roberto Bianconi⁵, Tim Butler¹⁰, Olivia E. Clifton^{6,7}, Philippe Cheung³, Alma Hodzic⁸, Richard Kranenburg⁹, Aura Lupascu^{10,11}, Kester Momoh⁴, Juan Luis Perez- Camaño¹², Jonathan Pleim², Young-Hee Ryu¹³, Roberto San Jose¹², Donna Schwede^{2,14}, Ranjeet Sokhi⁴, Stefano Galmarini^{15,*}

[1] Department of Physics, University of Patras, Patras, Greece

[2] ORD, U.S. EPA, Research Triangle Park, NC, USA

[3] Environment and Climate Change Canada, Toronto, Canada

[4] Centre for Climate Change Research (C3R), University of Hertfordshire, UK

[5] Enviroware srl, Concorezzo, MB, Italy

[6] NASA Goddard Institute for Space Studies, New York, NY, USA

[7] Center for Climate Systems Research, Columbia University, New York, NY, USA

[8] NSF National Center for Atmospheric Research, Boulder, CO, USA

[9] TNO, Utrecht, the Netherlands

[10] Research Institute for Sustainability – Helmholtz Centre Potsdam, Germany

[11] ECMWF, Bonn, Germany

[12] Tech. U. of Madrid (UPM), Madrid, Spain

[13] Yonsei University, Seoul, South Korea

[14] (Retired)

[15] European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract

We present the collective evaluation of the regional scale models that took part in the fourth edition of the Air Quality Model Evaluation International Initiative (AQMEII). The activity consists of the evaluation and intercomparison of regional scale air quality models run over North American (NA) and European (EU) domains for 2016 (NA) and 2010 (EU). The focus of the paper is ozone dry deposition. Dry deposition is among the most important processes of removal of chemical compounds from the atmosphere and an important contributor to the overall chemical budget of the latter. Furthermore ozone dry deposition is very important as it can be severely detrimental to vegetation physiology. 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). Following the AQMEII protocol and Dennis et al. (2010), we also perform a probabilistic evaluation in the form of ensemble analyses and an introductory diagnostic evaluation. The latter, analyses the role of dry deposition in comparison with dynamic and radiative processes and land-use/land-cover types (LULC), in determining surface ozone variability. Important differences are found across dry deposition results when the same LULC is considered. Furthermore, we found that models use very different LULC masks, thus introducing an additional level of diversity in the model results. The study stresses that, as for other kinds of prior and problem-defining information (emissions, topography or land-water masks), the choice of a LULC mask should not be at modeller's discretion. Furthermore, LULC should be considered as a variable to be evaluated in any future model intercomparison, unless set as common input information. The differences in LULC selection can have a substantial impact on model results, making the task of evaluating dry deposition modules across different regional-scale models very difficult.

*Corresponding author: Dr Stefano Galmarini, European Commission, Joint Research Centre (JRC), Ispra, Italy, email: Stefano.galmarini@ec.europa.eu

1. Introduction

This paper presents the results of the operational and probabilistic evaluation of the regional scale models taking part in the Air Quality Model Evaluation International Initiative phase 4 (AQMEII-4) activity. As presented in Galmarini et al. (2021), the AQMEII-4 focus is dry deposition process modelling within regional scale models (AQMEII-4-Activity 1) as well as standalone dry deposition modules (AQMEII-4-Activity 2) as detailed in Clifton et al. (2023).

As traditionally done in past editions of the AQMEII activity (Solazzo et al. 2012a, Im et al., 2014, Solazzo et al. 2017a), and in agreement with the protocol described by Dennis et al. (2010), prior to any detailed analysis of specific process modelling (diagnostic evaluation), a thorough analysis of the overall performance of the model must be conducted via operational and probabilistic evaluation. The scope of such an approach is to verify the positioning of the models participating in AQMEII with respect to observations or any other model simulating the case study or against a multi-model ensemble (Galmarini et al. 2013). Such an analysis has the scope of assisting the interpretation of any other detailed (diagnostic) result in this paper or other contribution to the special issue and understanding how the different processes contribute to the model spread. Examples of this approach can be found in Solazzo et al. (2012a and b), Vautard et al. (2012), Im et al. (2015, 2018), Giordano et al. (2015), Brunner et al. (2015), and Kioutsioukis et al. (2016). The operational evaluation also provides important context for the interpretation of diagnostic results – for example, the contrast in diagnostic comparisons between models with higher and lower evaluation performance helps to identify specific processes which may contribute to the differences (an example of this approach appears in Makar et al. (2025), this issue, for sulphur and nitrogen dry deposition, and Vivanco et al. (2018)).

Since the operational and probabilistic analysis is instrumental to the interpretation of ozone dry deposition-related results (the focus of the fourth edition of AQMEII), we shall concentrate on the variables that are directly or indirectly connected to description of dry deposition processes within the models, namely: atmospheric concentrations, land-use/land-cover (LULC) masks and meteorology. A detailed diagnostic analysis of modelled ozone dry deposition can be found in Hogrefe et al. (2025, this issue).

2. Models, domains, and years of consideration

The setup of the AQMEII-4 Activity 1 is detailed in Galmarini et al. (2021). In essence, the activity consists of running regional scale models on the North American (NA) and European (EU) domains for the years (2010, 2016) and (2009, 2010) respectively. The motivations behind the selection of these for years are given in Galmarini et al. (2021). The models that took part in AQMEII-4 are listed in Table 1, where details on the institutions in charge and the cases simulated are also provided. These models and in particular their dry deposition schemes are described more in detail in Galmarini et al. (2021, this issue), Makar et al. (2025, this issue) and Hogrefe et al. (2023 and 2025, this issue). Note that simulations took place with harmonized input emissions fields (Galmarini et al., 2021, this issue); all models started with the same anthropogenic, lightning NO_x, and forest fire emissions inventory for North America and Europe, respectively (Galmarini et al., 2021), while biogenic emissions and other natural sources of emissions such those of sea-salt particles were carried out as part of internal model processing and should be considered “part of the model” in the analysis that follows.

The analysis described here will only focus on two year-long simulations: 2016 for the NA case and 2010 for the EU case in the interest of synthesis. The following aspects will be considered in detail in this paper:

- Analysis of space and/or time averaged ozone concentrations
- Analysis of seasonal, diurnal, and spatial variations of ozone (and to a lesser extent nitric oxide, and nitrogen dioxide concentrations, in order to assist in the ozone analysis).
- Ensemble analysis of modelled ozone concentrations
- The role of variability in effective fluxes for specific pathways in determining the variability of ozone dry deposition flux over different LULC types
- The role of variability in wind speed, mixed layer height, dry deposition, and radiation in determining the variability of ozone concentrations at the surface.

Model values will be evaluated against ozone and precursor concentrations collected by regional operational networks during the year in consideration. More specifically, for North America the monitoring network databases employed included: the U.S. Environmental Protection Agency’s Air Quality System (AQS; https://aq5.epa.gov/aqsweb/airdata/download_files.html), the Canadian National Air Pollution Surveillance (NAPS) program (<https://www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/national-air-pollution-program.html>), and the Canadian National Atmospheric Chemistry database (<https://www.canada.ca/en/environment-climate-change/services/air-pollution/monitoring-networks-data/national-air-pollution-program.html>).

climate-change/services/air-pollution/monitoring-networks-data/national-atmospheric-chemistry-database.html). For the European case the monitoring network databases employed include: the European Monitoring and Evaluation Programme (EMEP; <https://www.emep.int/>), and the European Air Quality Database (AIRBASE; <https://eeadmz1-cws-wp-air02-dev.azurewebsites.net/download-data/>). The databases provide measurements in ppb for the NA case and $\mu\text{g}/\text{m}^3$ for the EU case. We opted for sticking to the original units avoid a conversion of one into the other to preserve the integrity of datasets and avoid the instruction of uncertainties that would penalise the quality of one or the other.

Given the continental dimension of the two regional domains simulated under AQMEII-4, the latter have been divided into sub-regional domains for analysis. These group portions of the network that share common features such as atmospheric circulation and possible sources of ozone precursors, and also provide continuity with past AQMEII model evaluation phases (Solazzo et al. 2012a and b).

Figure 1 shows the sub-regions selected within the two modelling domains, the corresponding sampling sites and the yearly average measured ozone (a and b). As noted by Solazzo et al (2012a), from the distributions of the pollutants, it is easy to identify the reason for those specific division in subdomains. In North America, a longitudinal divide is present between western (R1), central (R2) and eastern parts of the continent while the latter also requires a latitudinal division in two smaller subdomains (R3 and R4) due to the different kind of precursors' distributions and consequent ozone formation potentials. In Europe, the spatial distribution of emitters is different from North America and shows greater spatial density. There exist areas that require specific attention being almost decoupled from the rest of the continental air shed. These are typically the Iberian Peninsula and southern Mediterranean basin (R4), the Po Valley (R3) and Eastern Europe (R2). These NA and EU analysis sub regions were first defined in Solazzo et al (2012a), though with less detail and have been used in subsequent AQMEII analyses (e.g., Hogrefe et al., 2018) with different subdivisions but with the same goal of identifying regions with more homogeneous chemical potentials. For the sake of synthesis and in the absence of direct measurement of ozone dry deposition, this paper will concentrate exclusively on the model performance with respect to ozone concentrations with a few references to nitrogen oxides to give a more comprehensive sense of the quality of the performance of the individual models and the ensemble.

3. Operational evaluation

3.1 Ozone and nitrogen oxides surface air concentrations

3.1.1 NA case

The model performances at continental level and for the whole year are presented in Figure 2-6. For the two continents, the Root Mean Square Error (RMSE) and Mean Bias (MB) are computed from hourly ozone values for the entire year and are shown for each model in Figure 2-3 North America, and Figure 4, Figure 5 for Europe. Figure 6 shows the spatially averaged results presented in Figures 2 through 5 as box plot diagrams. In general, RMSE for the NA case (and in particular, for two models, namely NA7 (WRF-Chem (UPM)) and NA8 (WRF-Chem (NCAR))) appears to be larger than the EU case. Note that, since ozone values are reported in ppb over NA and $\mu\text{g}/\text{m}^3$ 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. Most differences from the observations are found in the eastern and south-eastern parts of the NA domain. As from Figure 2-5, three groups of behaviours can be distinguished for the NA case. Relative to the rest of the models, NA1, NA2, NA3 and NA5 (respectively WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Base), GEM-MACH (Ops)) show low RMSE values and comparable behaviours. NA4 (GEM-MACH (Zhang)) and NA6 (WRF-Chem (RIFS)) show slightly higher errors in the mid to east coast part of the domain whereas NA7 (WRF-Chem (UPM)) and NA8 (WRF-Chem (NCAR)) show markedly higher errors in the mid to eastern part of the domain and along the west coast. Looking at the biases (Figure 3), the analysis presented above is confirmed with some nuances though. In fact, we can see that the grouping can be more refined. A first group is made of the two EPA models NA1 and NA2 (WRF/CMAQ (M3Dry) and WRF/CMAQ (STAGE)) with a widespread overestimation across the continent. NA3 and NA5 (GEM-MACH (Base) and GEM-MACH (Ops)) produce the smallest biases of the group (see also Figure 3) and with a clearer West-East regional separation compared to NA1 and NA2. Finally, NA4, NA6, NA7 and NA8 (GEM-MACH (Zhang), WRF-Chem (RIFS), WRF-Chem (UPM), WRF-Chem (NCAR)) have larger biases, with NA8 having the largest mean bias (MB) of all (Figure 4). This analysis helps to distinguish the impacts of different dry deposition modules from the impacts of differences in other aspects of the model on simulated ozone. For example, WRF/CMAQ (M3Dry) and WRF/CMAQ (STAGE) differ only in their dry deposition modules, and the differences between these two simulations are generally smaller than their differences relative to the GEM-MACH and

WRF-Chem simulations. On the other hand, the dry deposition scheme has an important effect when we look at NA4 (GEM-MACH (Zhang)) vs. NA3 (GEM-MACH (Base)). These two models share the same regional scale system but use a different dry deposition scheme. The effect of the dry deposition schemes on the ozone concentration is quite remarkable. Recent work emphasizes a substantial effect of the magnitude of dry deposition velocity on ozone concentration (e.g. Baublitz et al. 2020; Wong et al., 2019; Clifton et al., 2020b). The results are consistent with those in Clifton et al. (2023) where the individual dry deposition module performances were evaluated (see discussion below). Therein larger differences were shown existing between the Zhang and Base schemes used in GEM-MACH than between the M3Dry and STAGE schemes used in CMAQ. Comparing NA3 (GEM-MACH (Base)) to NA5 (GEM-MACH (Ops)) reveals the impacts of model configuration and science option choices other than dry deposition, since both simulations use the Wesely scheme but differ in a number of other modelling aspects, as described in more detail in Makar et al. (2025). 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 effects of model configuration choices are also evident in the results of the three remaining models (WRF-Chem (RIFS), WRF-Chem (UPM), and WRF-Chem (NCAR)) that share the same dry deposition model and overall model code but utilize different configuration options. These simulations show a consistent overestimation that cannot be attributed clearly to one factor (see also Figure 3). The three implementations are also with respect to three different WRF-Chem version numbers (3.9.1, 4.0.3 and 4.1.2 respectively); versions 3.9.1 and 4.0.3 use the Grell and Devenyi (2002) cumulus parameterization, version 4.1.2 uses the Grell and Freitas (2014) parameterization. Furthermore, both WRF-Chem (RIFS) and WRF-Chem (UCAR) employ the same gas-phase mechanism (Emmons et al., 2010), while that of WRF-Chem (UPM) differs from the other two models. The relatively minor differences between WRF-Chem (UPM) and WRF-Chem (UCAR) shown in Figure 6a, may thus reflect differences in the gas-phase chemistry, with the former's mechanism resulting in slightly lower positive bias levels. 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). The larger differences (same figure) with the RIFS implementation reflecting differing cloud amounts and hence differing photolysis rates within the two implementations. The large overestimation of ozone by the WRF-Chem (UCAR) configuration may thus be linked to the underestimated precipitation in this model reported elsewhere (e.g. Makar et al. 2025), which also implies smaller cloud amounts and stronger solar radiation.

3.1.2 EU case

In Figures 4 and 5, RMSE and MB in Europe are presented, respectively. The errors have more a hot-spot character that is mainly evident in the southern part of the domain and therein at well-recognized critical regions like the Po Valley in the north of Italy, Greece and the Iberian Peninsula. This result is confirmed in the MB plots that also show EU3 (LOTOS/EUROS) as the best-performing model of the four though in many cases underestimating ozone concentration levels. EU2 shows worse RMSE scores than the other three models in particular over Germany, Poland, and Hungary, and scores the highest median RMSE value (Figure 6b). As for the rest of the domain, smaller RMSE values can be noticed throughout the region for all models. EU1 (WRF/Chem (RIFS)) and EU4 (WRF/CMAQ (STAGE)) show comparatively larger errors, especially in the southern and northern parts of the domain respectively. This behaviour of EU1, EU2 and EU4 may be associated with the prediction of NO₂ and NO concentration (see later discussion).

In this case, a model implementation/user effect can be an element of consideration since the EU4 is the same model that is used by EPA in the NA case (NA2), but in this instance run by the University of Hertfordshire. In the implementation of EU4, the primary differences lie in the meteorological model and the MEGAN biogenic emissions input. These variations in meteorological drivers and biogenic emissions can introduce differences, potentially contributing to the observed model biases when compared to other implementations of the same model. However, it should also be noted that the CMAQ simulations in North American (models NA1, NA2, Figure 3) also show positive biases, particularly along the US eastern seaboard. 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 NO₂ 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. .

From the analysis of NO, NO₂ and O₃ Normalised Root Mean Square Error vs Normalised Mean Bias in the soccer plots of Figure S1 in the Supplementary Material (SM, from now on) for the two continents, we note that the two precursors to ozone show an error smaller than 15% for all models except two. For the NA case, the ozone soccer plots confirm the grouping of the results qualitatively derived from the regional analysis of Figure 2. Figure 6 shows that GEM-MACH models NA3 and NA5 have ozone bias values closest to zero, followed by CMAQ (NA1 and NA2), while CMAQ has the lowest RMSE values, closely followed by the GEM-MACH NA3 and NA5 implementations. Four models show small error (<15%), two with medium (>15% and <20%) and two with high (>20%). The ozone goal plots for the EU (Figure S1) show a statistical tendency to produce smaller errors than the NA case and in particular more coherence between the errors for ozone and its precursors.

The Taylor diagram depicted in Figure S2 in the SM also evaluates the correlation between simulated and observed ozone values. The results show a higher correlation of model predictions with observations in the EU case while the other statistical parameters in the diagram confirm what has been presented in the other plots. The multi model ensemble (MME) is also presented for the two cases, showing in both instances an improved performance with respect to the individual model simulations.

3.1.3 Diurnal and seasonal variability

Figure 7 shows a comparison of observed and modelled seasonal and diurnal cycles for North America for ozone, NO and NO₂. These cycles were constructed by averaging the underlying raw hourly data available for the entire year over a given month-of-year or hour-of-day, respectively. At the monthly level, the figure clearly shows that for ozone in NA, almost all models over-estimate the concentration during summer. The multi-model mean fails to reproduce the ozone maximum in April by overshooting by approximately 3 ppb and presenting a maximum in June. This result is driven by 4 out of 8 models (NA4 (GEM-MACH (Zhang)), NA6 (WRF-Chem (RIFS)),

NA7 (WRF-Chem (UPM)) and NA8(WRF-Chem (NCAR))). Although slightly overestimating the concentration, two models (NA3 (GEM-MACH (Base)), and NA5 (GEM-MACH (Ops))) manage to reproduce very accurately the seasonal evolution. NA1 and NA2 (WRF/CMAQ (M3Dry) and WRF/CMAQ (STAGE)) capture the trend and seasonality and just slightly overestimate the ozone peak value.

The tendency for overestimating ozone concentration and underestimating NO is also clear from Figure 7 (for NA) and Figure 8(for EU). Figure 7's diurnal variation panels (right hand column) in particular show that the models NA3 and NA5 have the closest values to observations for O₃, NO and NO₂, though all models underestimate the NO_x totals. This is especially evident for NO and NO₂ in the mid-day hours (10 to 18 local time), when the simulated NO and NO₂ values are the closest in the ensemble to the observations. The monthly variation panels (Figure 7 left column) show that the relative impact of the NO_x underestimates is smaller in the summer than in the winter, and models NA3 and NA5 have the closest NO values to observations and slightly overestimate NO₂ in the summer. 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 NO_x concentrations from surface sources, and also shift the chemical regime from ozone production to ozone destruction by NO_x titration below the forest canopy). Model NA3 also includes the effects of vehicle-induced turbulence on NO_x emissions from vehicles (Makar et al., 2021), an effect which results in more efficient dispersion of these emissions out of the surface layer. Model NA5 assumes the area emissions of NO_x are evenly and instantaneously distributed over the first two vertical levels of the model rather than incorporating these emissions as a flux boundary condition on the diffusion equation. As noted above, Models NA3 and NA5 include process representation which can enhance the vertical transport of freshly emitted NO_x out of the lowest model layer; at least some of superior performance may be related to this faster dispersion. The ozone dry deposition velocity used in NA3 and NA5 versus that of NA4 is also a driver for the differences between these models, as noted in Clifton et al. (2023), which noted that NA3 and NA5 shared a scheme which significantly overestimated ozone dry deposition velocities relative to observations in the summer while providing reasonable estimates during the winter, while the Zhang scheme, used in NA4, showed little seasonal variation (tending to be flat over time, with overestimates during winter, underestimates during summer). It is of note that the models that reproduce the seasonal evolution of ozone most accurately during summer when the rest of the models struggle, have the dry deposition schemes with the largest positive biases in summertime ozone dry deposition

velocity and the greatest seasonal amplitude (Clifton et al. 2023). This implies (1) that the factors affecting the ozone concentrations have a strong seasonal dependence (models NA4 versus NA3 and NA5), (2) and that while one means of helping achieve that seasonal dependence is through an overestimation of the ozone dry deposition velocity relative to observations (models NA3 and NA5), (3) *other seasonally dependent process improvements than dry deposition velocity are required to better simulate ozone* (given that the other models considered here which incorporate more accurate ozone dry deposition schemes, relative to the observations in Clifton et al. (2023) also have high positive biases in parts of NA and EU (Figure 3 and Figure 5). As noted above, process representation of forest canopy shading and turbulence is one such possible means of model performance improvement¹. The other consideration worth examining is the interdependence between model cloud cover and surface photolysis rates, given the variation between NA WRF-Chem models NA6, NA7, and NA8, where the largest differences in ozone positive bias correspond to the use of differing cloud parameterizations.

For NO and NO₂, the models show seasonal cycles which differ between the models (Figure 7, Figure 8, left-hand columns) versus the observations and between the NA and EU observations. Observed NA ozone peaks in April (month 4, Figure 7 upper left panel), while observed EU ozone peak in July (month 7, Figure 8 upper left panel). As noted above, models NA1, NA2, NA3, and NA5 all capture the NA O₃ seasonality (CMAQ and Base and Ops GEM-MACH configurations) while the WRF-Chem models predict a late summer peak, similar to observations in EU. All models tend to overestimate compared to observed ozone concentrations (exceptions: NA3 and NA5 in April and May, Figure 7, EU2 and EU3 from November to April). All models underestimate wintertime NO_x (though NA models NA1, NA2, NA3, NA5, and NA7 have close NO₂ performance to observations from July through October, Figure 7), and EU3 NO values closely match observations, while EU2 NO₂ is biased high relative to observations. All NA models have significant (factor of two or more) negative biases in NO, and the largest seasonal NO₂ negative biases in winter. As a consequence, all NA models strongly underestimate the amplitude of the observed seasonal cycle.

¹ We note that subsequent investigation at ECCO of the GEM-MACH dry deposition algorithm described in Makar et al. (2018), following the results published in Clifton et al (2023) identified two key errors added to the code in the version subsequent to the code version used in Makar et al (2017). Specifically, the cuticle resistance formula (Makar et al, 2018 equation S.8, Clifton et al (2023) equation (42) made use of Zhang et al (2002) dry cuticle resistance coefficients (rc_{cuti} , rlu) which should not have been scaled by inverse leaf area index, and made use of Zhang et al (2002) coefficients for the lower canopy resistance (Makar et al, 2018 equation (S.2), Clifton et al 2023 equation (44) which did not include the required scaling of the coefficients by $(LAI^{0.25})/(u^*)^2$. Subsequent to these corrections, a much closer fit to the observations in Clifton et al. (2023) was achieved. (K. Toyota, A. Robichaud, personal communication, 2024).

Potential factors which might drive an underestimate of wintertime NO_x include underestimates in the emissions of NO_x from combustion sources such as wintertime home heating from fossil or wood fuels (van der Gon, 2015), underestimates of atmospheric stability (i.e. if the simulated atmosphere is more unstable than the actual atmosphere, NO_x 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 NO_x (Michaud et al., 2015).” Figure 8 also shows, not unexpectedly, that the models with the smallest NO and NO₂ biases (EU2 (WRF-Chem (UPM)) and EU3 (LOTOS/EUROS)) do quite well for O₃, NO and NO₂, and the EU NO and NO₂ biases for these models are in general much smaller than the NA model biases. At the diurnal level (Figure 7, Figure 8 right panels) the results are consistent with what is found at the seasonal level in terms of over- or underestimations. At the diurnal level, EU2 outperforms the others showing a good capacity of catching the average time evolution of the three pollutants.

The monthly averaged ozone, NO and NO₂ concentrations breakdown at the sub regional level are presented in Figures S3 and S4 for NA and EU, respectively. From Figure S3 one can conclude that the major contribution to the domain-wide estimation presented earlier is essentially coming from regions R2, R3 and R4 (i.e. the eastern part of the domain) whereas all model results in R1 are rather similar and in agreement with the measurements throughout the year with some models overestimating cold seasons but by a lesser extent than in the other regions. The summertime ozone overestimation over the Eastern U.S. for NA1 and NA2 (WRF/CMAQ (M3Dry) and WRF/CMAQ (STAGE)) is consistent with the findings of Appel et al. (2021). It is also worth noting that all of the NA models (Figure 9) overestimate O₃ in the period from July through September in regions R2, R3, R4; an observed effect largely absent in the EU models (Figure 10). We also note that the time series of observed O₃ 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). From figure S4, referring to the EU case, we see that EU1 and EU4 underestimated NO and NO₂, whereas EU2 largely overestimates for all European sub-regions. Such model performances can explain the ozone biases as they affect ozone titration at night. This effect is apparently exacerbated in the Po Valley area, which is known for high NO_x emission levels. The observational sites in the Scandinavian Peninsula are mainly from the EMEP network which is

representative of the remote background whereas the AirBase network rural background sites are more prone to local sources of pollution.

These regional differences will be instrumental to the analysis of dry deposition processes. The same behaviour observed in sub regions is found at both the seasonal and hourly level. From Figure 10 we can see the situation in Europe, which lacks the large positive biases in the NA simulations.

3.1.4 Summary of the analysis

One overall conclusion from the comparisons with observations for NO, NO₂ and O₃ is that:

- the models which most closely match NO and NO₂ (EU2, EU3) also have the best performance for O₃,
- that those models with negative biases for NO and NO₂ also have positive biases for O₃, and that the magnitude of the NO_x negative biases is inversely proportional to the magnitude of the O₃ positive biases, for all models.
- The relative magnitude of the “freshly emitted” component of NO_x (i.e. NO) tends to be underestimated, with the exception of model EU3 (LOTOS/EUROS).

These results all point towards excessive vertical mixing of fresh NO emissions up from the lowest model layer as a root cause of the model biases in the other models. The reasons for this conclusion are:

- (1) the relative fraction of NO_x that is NO will be highest in air dominated by fresh emissions;
- (2) the relationship between positive ozone biases and negative NO biases indicates that the ozone biases are due to insufficient NO titration;
- (3) the effect is exacerbated in winter in all NA models and some EU models - a time when the atmosphere tends to be more stable, and photolysis rates in the northern hemisphere are low, both conditions which favour NO_x titration.

A secondary cause may be missing NO emissions in the wintertime, though this seems less likely due to the relatively high confidence in mobile emissions and stack emissions, which dominate the NO_x emissions totals, and the relatively good performance of EU3 relative to the other EU models when making use of the same emissions inventory.

381

382 **3.2 Ozone dry deposition fluxes**

383 We start our examination of O₃ dry deposition fluxes with the direct comparison of the
384 effective and total fluxes calculated by the models. Effective flux is a convenient way of examining
385 the contribution of the resistances of various pathways towards bulk dry deposition, taking into
386 account that variability is not only due to these resistances but also surface ozone concentrations
387 (Galmarini et al., 2021). The definition of effective fluxes is analogous to the definition of effective
388 conductances (Paulot et al., 2018; Clifton et al., 2020b). Specifically, by definition, the sum of the
389 effective fluxes equals the total ozone dry deposition flux, and this equality is used in the
390 subsequent analysis. Within AQMEII-4, the relevant effective conductances were defined *a priori*
391 and every participating modelling group was requested to determine the combination of all
392 relevant resistances accounted for in their systems, necessary to produce the effective
393 conductances requested. The definitions of the effective conductances, the dry deposition
394 modelling approaches and the detailed formulation of effective fluxes for each model are
395 presented in Galmarini et al. (2021, this issue). Because effective conductances and ozone
396 concentrations can co-vary on daily timescales, it was important to archive high-frequency
397 effective fluxes; for this same reason, conclusions about drivers of variations in effective fluxes
398 may be distinct from those regarding effective conductances. The analysis of effective and total
399 fluxes is performed only for the grid cells in which all models share the same LULC (for details on
400 the common LULC classifications see Galmarini et al. (2021, this issue)). By restricting the analysis
401 to locations sharing the same characteristics of land use across models, we reduce the impact of
402 LULC variability on the resulting analysis, thus allowing us to compare only the response of models
403 to the different dry deposition schemes employed for a given LULC. We present model results at
404 grid cells that are covered by at least 85% of respectively, Evergreen Needleleaf Forest (NA: 1544
405 cells, EU: 2531 cells), and Planted-Cultivated (NA: 6130 cells, EU: 6108 cells). In addition, we also
406 define an ‘Ozone Receptor’ case that corresponds to the grid cells where ozone is monitored at
407 the surface in the two continents (NA 1551 cells, EU 1656) independently from the underlying
408 LULC type, which can therefore be different from model to model. In the SM the Deciduous
409 Broadleaf Forest (581 cells), Mixed Forest (705 cells) are also presented for NA case only for the
410 sake of synthesis.

411 An important finding is obtained by simply imposing the data the selection criterion
412 described above. As can be noted, for the same continent the models share relatively few grid

cells with the same dominant LULC. This is a clear indication of the fact that individual LULC masks, employed in the models, were obtained from substantially different sources (Table 1). Such results raise a significant issue: is it acceptable that the characterisation of the land surface differs so much? In principle LULC masks adopted by the AQMEII 4 models should be very comparable, especially when sources of this information with a high degree of spatial resolution are now available. More discussion may be found in Section 5 and in our companion paper (Hogrefe et al., 2025, this issue).

Figures S5 and S6 show seasonal cycles of the total ozone dry deposition flux and its decomposition into the three different effective fluxes. The pathways represented by these effective fluxes are (1) lower canopy and soil conductances combined in one factor (LCAN+SOIL) since some models did not distinguish these two terms, (2) cuticular conductance (CUT) and (3) stomatal conductance (ST).

The following features can be appreciated across the model results:

- The magnitude peak of the ozone flux varies considerably from model to model in some cases (NA8) being almost twice that of others (NA4) for the monthly average.
- Typically, the flux is highest during summer and lowest during winter. In some cases, some fluxes show nearly constant values throughout the summer season (NA2, NA3, NA5 and NA7). In others, there is a stronger midsummer peak (NA1, NA4, NA6) in July or August. NA8 shows a double peak shape. Given the dry deposition scheme is the same in NA8 as NA7 and NA6, this suggests this double peak is either meteorologically driven, or ozone driven.
- In the EU case more homogeneity appears between EU1 and EU2 behaviours while EU4 shows a slightly different performance at this macro level analysis at least.

The breakdown of the contributions of the specific pathways to the total ozone flux does not appear to identify any common behaviour either across models, or within the same LULC type nor across time. It is particularly notable that the relative contributions of the different pathways vary between models, (e.g., compare the relative magnitude of stomatal flux in NA1 and NA2, Figure S5(a)). Some models employing the same dry deposition algorithm nevertheless have different contributions associated with the different pathways (see NA3 versus NA5, which have the same dry deposition algorithm, yet the soil term dominates in NA3 and the cuticle term dominates in NA5). 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..

We note that an exception to the explanation presented above is for the “planted-cultivated” LULC, where ST and LCAN+SOIL tend to dominate the flux. There is also a clear summer maximum in ST across models (Figure S5e), but the exact seasonality of ST differs significantly between models. LCAN+SOIL tends to have a bi-modal seasonality for this LULC type – with minima during winter and during times of maximum ST. CUT tends to be low – with NA1 and NA2 suggesting slightly higher values – with weak but noticeable seasonality with a broad growing season peak. To a certain extent, this pattern in seasonal variation in the different pathways and their contribution to the total flux also shows up for deciduous forests (Figure S5c), but less so for CMAQ than for the other models. In general, stomatal flux tends to drive seasonality in the ozone flux, as Clifton et al. (2023) found for ozone dry deposition velocity at the individual flux sites, but sometimes there a seasonal contribution in non-stomatal flux. The models also all differ in the relative contributions of LCAN+SOIL, CUT, and ST, as also found by Clifton et al. (2023). For example, cuticular flux is very low in some models (e.g., WRF Chem) but a dominant contributor (about 1/3 except over crops) in NA1 and NA2. Perhaps the primary conclusion is that model behaviour can be grouped around the model type. In fact, clear similarities can be found among NA3 and NA5 (GEM-MACH (Base) and GEM-MACH (Ops) for several land-use types), as well as NA6, NA7 and NA8 (WRF-Chem (RIFS), (UPM) and (NCAR) respectively). In the EU case, EU1 and EU2 (both WRF-CHEM) have comparable yearly characteristics, while EU4 (WRF/CMAQ (STAGE) used by the University of Hertfordshire) shares a similar breakdown with NA2 (WRF/CMAQ (STAGE) run by the USA-EPA).

Although relevant for operational evaluation, the analysis in Figures S5 and S6 does not easily reveal the significance of dry deposition processes and pathways in determining ozone variability across models. Toward this end, hierarchical and variation partitions are considered in Section 5.

4. Probabilistic evaluation

The ensemble analysis described in this section aims to identify the models that contribute to an improved ensemble result and the best combination of models that improves the ensemble

skill. Such analysis is part of the probabilistic evaluation described in Dennis et al. (2010) and constitutes one of the four pillars of evaluation defined therein and adopted in the overall AQMEII activity. In past phases of AQMEII, ensemble analysis was also presented as an integral part of the model evaluation (Solazzo et al. 2012a, Solazzo et al. 2012b, Galmarini et al. 2013, Kioutsioukis and Galmarini, 2014, Im et al. 2015, Solazzo et al. 2013, Kioutsioukis, et al. 2016, Solazzo et al. 2017b, Galmarini et al., 2018). The ensemble mean of the model results has already been presented in the operational analysis. However, identifying which and how many models contribute to improved ensemble results is another question to be addressed in this context. The analysis uses ozone mean concentration measured at the monitoring sites as reference and techniques based on model combination to determine the optimal results as described in earlier studies (Solazzo et al. 2012a, Solazzo et al. 2012b, Solazzo et al. 2013, Galmarini et al. 2013, Kioutsioukis and Galmarini 2014, Kioutsioukis, et al. 2016, Galmarini et al., 2018).

The skill of an ensemble increases if we combine accurate and diverse models (Kioutsioukis and Galmarini, 2014). As shown by Solazzo et al. (2012a) the skill normally reaches a maximum for an ensemble composed of less than half of the available models and then deteriorates when more models are added until reaching an asymptotic value. Given m available models, several combinations of model results in groups of $n \leq m$ can be produced. In this analysis, we aim at identifying the minimum number of models that produce the optimal result and which are the models that produce the highest ensemble skill. We therefore consider all ensembles obtained by combinations of members in each group constructed from the m models (i.e., a total of $\sum_{n=1}^m \binom{m}{n}$ ensembles where $\binom{m}{n}$ represents the combination of n models out of a total of m available). For each combination, we calculate the RMSE with respect to the measured values and identify the ensemble with the least error. Note that these ensembles cover the full range of possible combinations from first-order (one model ensemble) to m^{th} order ($m = 8$ models for NA case and $m = 4$ models for EU). To avoid the exclusion of yet meaningful results and at the same time to study how the variety of models analysed combines toward those, we also present the results of ensembles with RMSE within 10% of the optimal one. Lastly, we determine the frequency with which each model is selected as part of an optimal ensemble.

In Table 2 the results from NA are presented. The analysis of the 255 ensembles obtained by combining the models in groups of 1, 2, 3 through 8, gives a RMSE ranging from 3.77 to 11.89 ppb. The results from Solazzo et al. (2012a) are confirmed in this study, therefore in the Table 2 we will present only results up to order 4 (i.e. four members in the ensembles) in the NA case,

since for higher orders the results only deteriorate. The ensembles with the least error are obtained from the average of two and three models results (i.e. a 2nd and 3rd order ensemble, blue columns). The models that contribute to these two optimal ensembles are WRF/CMAQ (STAGE) and GEM-MACH (Ops) for order 2 and WRF/CMAQ (STAGE), GEM-MACH (Base) and GEM-MACH (ops) for order 3. The second-best ensembles (yellow columns) are also of order 2 and 3 and are composed of GEM-MACH (Base) and GEM-MACH (Ops) results, and WRF/CMAQ (M3Dry), GEM-MACH (Base) and GEM-MACH (Ops), respectively. In particular, it is worth noting that (a) order one features two of the models most present in the ensembles and their individual result is still within 10% of the best higher order ensembles (b) WRF-CMAQ and GEM-MACH are the most frequent contributors, (c) WRF-Chem versions (RIFS, UPM, NCAR) are never contributing to any ensemble set. We note that both WRF/CMAQ and GEM-MACH (Base and Ops) are used for operational air-quality forecasting in the USA and Canada, respectively, and hence (1) they are frequently evaluated against monitoring data under the principle that new model versions must improve the forecast before replacing old model versions, (2) the ongoing evaluation process will tend to select model configurations with the best performance with respect to ozone concentrations, (3) this ongoing evaluation process is for the model as a whole, while individual processes tend to be evaluated based on other data, and are incorporated into the base code, (4) this process can result in the adoption of processes with compensating errors (c.f. Makar et al. 2014, and note the contrast between dry deposition velocity performance for NA3, NA5 here versus the dry deposition velocity performance in Clifton et al., 2023). As new data such as the dry deposition observations of Clifton et al. (2023) become available, compensating errors come to light, allowing for corrections and updates to the model codes to be carried out.

The EU ensemble (Table 3) has 4 models, which generates 15 ensembles with RMSEs ranging from 7.51 to 14.59 $\mu\text{g}/\text{m}^3$. Four out of the 15 combinations of 2nd, 3rd and 4th order have errors within 10% (yellow column) of the optimal combination generated from LOTOS/EUROS and WRF-Chem (RIFS) for the second order (blue column). No 1st order ensemble has a RMSE smaller than the 2nd order best ensemble, meaning that no individual model run on the EU case performs better than the combination of the two shown in the 2nd order grouping. LOTOS/EUROS is present in all the ensembles created but yet alone is not doing better than when its results are averaged with those of WRF/CMAQ (STAGE). The latter, operated by the University of Hertfordshire for this case study, is present 80% of the time as a contributor to the second- and third-best ensembles. We note that LOTOS/EUROS, like the GEM-MACH and WRF/CMAQ models in NA, provides

operational forecasts of O₃, NO₂, and PM₁₀, and hence will likely benefit from ongoing evaluation and selection of process representation that gives the most accurate model results. Since the results of all orders are shown in Table 3 we can see that the conclusion of Solazzo et al. (2012a) is confirmed to the extent that a combination of half of the available members tends to outperform any single model or larger ensemble of results. It should be clear that the number of models is only an indication to the extent to which the combination of specific models allows one to produce the best results with a reduced number of ensemble members.

5. Variance analysis of ozone fluxes and the role of conductances, turbulence, radiation and wind speed to ozone variability on common LULC types

At this stage of the analysis it is important to determine the overall role of dry deposition and other relevant factors in determining the variability of ozone concentrations at the surface. Having established which grid cells are representing the same LULC characterisation (Section 3.2), we proceed with the analysis of dry deposition data by identifying a set of parameters that are expected to be relevant in the characterisation of the ozone flux, namely:

- Lower canopy and soil effective flux (LCAN+SOIL) combined as one factor,
- Cuticular effective flux (CUT)
- Stomatal effective flux (ST)
-

We also identify the factors that are expected to be relevant in the determination of ozone concentration variability at the surface, namely:

- Boundary layer height,
- Solar radiation,
- Wind speed,
- Dry deposition velocities.

Chemical transformation is a dominant factor in creating ozone variability together with the abundance of ozone precursors. However, it is challenging to represent the influence of these factors through a specific variable, although solar radiation can be viewed as a proxy of

photochemical activity. We note that air temperature can also have a significant influence on photochemical formation of ozone, but air temperature will also influence the dry deposition pathways; the two influences would be difficult to differentiate. Although the analysis will be performed over all the months of the analysed years, the main focus will be around the summer months, when the ozone production and mixing ratios are normally at maximum levels, and when models are performing the worst, at least over NA.

5.1 Relative relevance of pathway fluxes in ozone flux variability

Variation partitioning of a single response variable (Y, e.g. total O₃ flux, or O₃ concentration) is based on the adjusted R² in a regression framework (Peres-Neto et al., 2006; Lai et al., 2022). For example, the variation partitioning of O₃ flux between three sets of predictors (X1: LCAN+SOIL, X2: CUT, X3: ST) can be achieved through the estimation of the fractions (represented here by the dummy variables: a, b, c, d, e, f, and g) based on one (X_i), two (X_i, X_j) or three (X_i, X_j, X_k) variables (Figure S7):

(1) fractions based on one variable:

$$\begin{aligned} [a + d + f + g] &= R_{Y|X1}^2 \\ [b + d + e + g] &= R_{Y|X2}^2 \\ [c + e + f + g] &= R_{Y|X3}^2 \end{aligned} \tag{1a}$$

(2) fractions based on two variables:

$$\begin{aligned} [a + b + d + e + f + g] &= R_{Y|(X1,X2)}^2 \\ [a + c + d + e + f + g] &= R_{Y|(X1,X3)}^2 \\ [b + c + d + e + f + g] &= R_{Y|(X2,X3)}^2 \end{aligned} \tag{1b}$$

(3) fraction based on all three predictor variables:

$$[a + b + c + d + e + f + g] = R_{Y|(X1,X2,X3)}^2 \quad (1c)$$

Y in equations 1 (a,b,c) is the predictor variable in this case ozone deposition flux. From the above expressions, we can estimate the *sole* and *shared* contributions of each predictor. For example, the *sole* and *shared* fraction of variation explained by X_1 are respectively:

$$sole = [a] = [a + b + c + d + e + f + g] - [b + c + d + e + f + g] \quad (2a)$$

$$shared = [d/2 + f/2 + g/3] \quad (2b)$$

where (similarly for the other fractions):

$$[d] = [a + b + c + d + e + f + g] - [c + e + f + g] - [a] - [b]$$

The analysis proceeds by carrying out multiple regressions for the equations 1(a) through 1(c); the values of the left-hand side terms that minimize the differences between left and right-hand sides of the equations are then compared - these provide the relative contribution of the component terms towards the net correlation coefficient between the ozone flux and the three predictors.

For the sake of synthesis in the main paper, we shall present results of the variance decomposition analysis for the two most relevant LULC cases (evergreen needle leaf forest and ozone receptors). The analysis for all other LULC types selected and listed in Section 3.2, is presented in the Supplement.

Figure 11 presents the contribution to the ozone dry deposition flux variability of the three effective fluxes (total or 'sole' plus 'shared': first and third column of plots in each figure) and their decomposition into 'sole' and 'shared' fractions (second and fourth column panels) for all months of 2016 and for the eight models participating in the NA case study for shared cells covered by at least 85% evergreen needle leaf forests.

Considering the first and third columns of Figure 11 (where the sum of Eqs. 2a and 2b is presented) we note that for all models the fractional contributions to ozone flux variance add up to 1 as expected. For the summer period, we can see that the models can be divided into three main groups. The first group is where stomatal effective fluxes dominate in defining the ozone flux variability (WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE)), a second group where the dominant

pathway to ozone flux variability is through the cuticular effective flux (GEM-MACH (Base), GEM-MACH (Ops) and GEM-MACH (Zhang)) and a third group where the main factor is the combined soil and lower canopy effective flux combined (WRF-Chem (RIFS), WRF-Chem (UPM), WRF-Chem (NCAR)). This constitutes a significant result that is also in line with those obtained by Clifton et al. (2023), but extends their finding. For example, Clifton et al. (2023) show that different models have very different relative partitioning across effective conductances at individual sites. Our result here suggests that spatial variability in the ozone flux across the same LULC type is mainly determined by different flux pathways. Given the fact that the grid cells selected were dominated by the same land-use type, differences between the three groups can be attributed to substantial differences in the dry deposition modules, concentration gradients, and meteorology. In the winter and autumn months, the contribution to ozone flux variability is equally distributed across the three pathways for all models for this LULC type. We also note that the seasonal cycle of the “sole” terms varies as a function of model. The stomatal conductance term dominates the CMAQ implementations (NA1, NA2) in the summertime, while for the GEM-MACH implementations (NA3, NA4, NA5), summertime seasonality is mostly driven by the soil + lower canopy term, while for WRF-Chem implementations (NA6, NA7, NA8), stomatal and soil+lower canopy terms both have a weak maximum in the summer.

In Figure 11 the results of the decomposition obtained according to equation (2) are independently presented (columns 2 and 4). For the sake of presenting the results in a clearer way, the contributions to the variation obtained from equation 2b, are plotted after changing their sign to better distinguish them from the others; but the total sum of the negative and positive values should be 1. This more detailed analysis allows us to verify the previous one with additional details. For example, the predominance of stomatal flux in WRF-CMAQ at the warm season is due to the sole contribution of stomatal flux whereas at the other seasons the shared contributions dominate. For GEM-MACH, the importance of the cuticular flux seen earlier arises from its shared contributions except GEM-MACH (Zhang) where its sole fraction appears equally high throughout the year. 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. . We note that the WRF-Chem models are also being used in feedback mode and have less variation than the GEM-MACH case, potentially indicating a smaller impact of differing model parameterizations on the feedback portions of the WRF-Chem code.

Last, for WRF-Chem, the shared contribution of soil and canopy flux is important all year, but its sole contribution becomes equally high in the warm season.

Figure 12a shows the same analysis for the EU continent where the picture differs from NA, indicating very different meteorological condition between the two regions. This is not unexpected, in that EU meteorology is strongly influenced by the ocean circulation of the Gulf Stream, while the NA meteorology is over a broad region that has a much broader range of conditions in a “continental” climate. In two of the three models (WRF-Chem), the importance of soil-lower canopy and stomatal effective fluxes in the warm season (mid spring through October) is due to their shared fractions while the sole contribution of the cuticular effective flux in winter drives the variation of the total O₃ flux. The seasonality of the EU stomatal component is shared with that of NA6, while the EU soil components have a greater degree of seasonality compared to the NA WRF-Chem models. The other model -- EU4 (WRF/CMAQ (STAGE)) -- shows a more even distribution of the stomatal contribution across the year, and a more equal distribution across the three pathways during the year. EU3 is not presented since no data were delivered for effective conductances.

From the figures S8-S10 one can deduce that the rest of the land covers (Deciduous Broadleaf Forest, Mixed Forest, Planted-Cultivated) still exhibit a dominance of stomatal effective flux during the summer. These LULCs all have a significant deciduous component, and the summertime dominance is in part due to the wintertime absence of foliage in the more northerly parts of the model domains. Depending on the model, cuticular and soil are at times the second contributor to variability of ozone flux.

The category ‘Ozone Receptor’ groups the results at grid cells containing an ozone sampling location regardless of the land cover adopted by individual models (Figure 12b for EU and 13 for NA). It is interesting to note that the Ozone Receptors case shows a remarkable consistency across models in terms of the contribution of the different effective fluxes and their variability in time, a behaviour not seen when performing this analysis for grid cells dominated by specific LULC types. This can be appreciated from Figure 12 where the evergreen needle-leaf forest case (12a) is presented side-by-side with the ozone receptor case (12b) for the EU domain. There is some disagreement for the EU about the stomatal flux contribution during winter (zero or low) and on the exact partitioning during warm months, but generally all the models show substantial contributions from the stomatal flux, though disagreeing on the exact non-stomatal partitioning. The consistency for the Ozone Receptors case is also visible across the continents (Figure 13 for

the NA case) where the contribution has a remarkable resemblance across models for seasonality and the partitioning of the ozone flux variance across the effective fluxes, compared to individual land use type values. For the NA case, models suggest moderate to strong contributions for LCAN+SOIL during winter, yet small to moderate contributions during summer; the contribution of cuticular effective flux tends to be constant and moderate throughout the year, with three models (WRF-CHEM) suggesting smaller contributions in winter; with stomatal effective flux making up the difference, roughly a third of the total, but sometimes as low as 10% or as high as 50%.

This result calls for some important considerations:

1- The remarkable consistency and similarity found among the model results at the ozone receptor locations could be due to the lack of dominance of any specific LULC type at this subset of grid cells considered. This would be in agreement with the fact that the locations have presumably been chosen for air quality monitoring activities and by-design are intended to be neutral to any prevailing process such as the removal of pollutants from the atmosphere by dry deposition processes, thus extending the spatial representativity of the monitoring locations.

2- The variance decomposition into contributions of both *sole* fluxes and *shared* fluxes (columns 2 and 4 of Figure 13 and column 2 in Figure 12b) does not show the same agreement found for the total fluxes (columns 1 and 3 of Figure 13 and column 1 of Figure 12b). This indicates that every dry deposition model maintains a peculiarity in its behaviour for individual land use types. This specificity is lost in the results when the ozone monitoring station are considered. This suggests that while the monitoring station locations show the models perform in a similar fashion for mixtures of LULC types, the model performance for individual land use types (represented by a much smaller number of stations) may differ significantly. Given that model performance is judged using observation station values, this may indicate that dry deposition algorithms have been inadvertently tuned towards providing similar results in the regions where mixtures of LULC values are present - but require single LULC type stations for the evaluation of individual LULC performance. We note that this tuning is not intentional, but a product of the purpose for which monitoring stations have been set up (e.g. human health impacts, and hence closer to human habitations than remote locations which may have a single LULC) and the availability of infrastructure (roads, electrical power) for station operations. This result underscores the importance of land-use specific dry deposition sites such as those used in point model dry deposition velocity analysis in Clifton et al. (2023, this issue) when evaluating dry deposition

algorithms, and suggests that subsets of monitoring network stations located in single LULC types should be identified (or constructed if none are available) in order to further improve model performance within those LULC types. The result is that the dry deposition algorithms are achieving similar results for dry deposition flux relative to observations – but sometimes via very different pathways, especially across different LULCs. This is in line with suggestions from recent work examining a single model (Silva and Heald, 2018), a review paper on modelling ozone dry deposition (Clifton et al., 2020a), and the results of the single-point modelling AQMEII Activity 2 paper (Clifton et al., 2023). These findings and the above analysis illustrate

- a strong need to generate observational datasets which focus on specific dry deposition components for model evaluation (e.g., as suggested by Clifton et al. 2020a),
- the need for dry deposition velocity observation to evaluate dry deposition algorithm performance
- the need for monitoring network locations that represent specific LULCs, to improve model performance in regions where one LULC dominates.

The current evaluation practice with mixed LULC monitoring stations used for dry deposition algorithm evaluation prevents progress in algorithm improvement in specific LULCs, and allows for LULC-specific compensating errors to be missed in dry deposition algorithm development.

3- If (1) and (2) can be confirmed one should consider comparing dry deposition results obtained at operational monitoring sites with care – the net results of the comparison may be that the regional models and possibly their dry deposition fluxes agree – on average, for regions with multiple land-use types - but the agreement is the result of regional model evaluation procedures as opposed to mechanistic dry deposition velocity algorithm evaluation that is LULC-specific. Furthermore, this may give an appearance of agreement among regional models that may be illusory, since in grid cells with shared dominant LULC types more disagreement has been demonstrated in the above analysis. An important implication of this finding is the need to evaluate regional models using both single-land-use and multiple-land-use type stations in the future, and for representation in single-land-use type locations to be a consideration in monitoring network design.

5.2 Non-linear contributions of other factors to the ozone concentration variance

The analysis of the non-linear contributions to the ozone variance has been conducted by introducing other factors considered to be relevant in influencing ozone variability at the surface level, namely: boundary layer height, solar radiation, wind speed, and dry deposition velocity. In a way, this analysis allows us to determine the role of dry deposition in relation to other factors influencing the variation of ozone concentrations at the Evergreen Needleleaf Forest cells and therefore estimate its relevance as a driver of ozone variance in a regional scale model. Figure 14 presents the analysis for the NA case while Figure 15 shows results for the EU case.

From Figure 14 we firstly note that the selected components have a very relevant role in the determination of the surface ozone variance as, overall, they account on average for 60 to 80% of ozone variance. The remaining portion can be attributed to variations in emissions and chemical reactions that cannot easily be represented by a specific variable, or to other factors not considered in this analysis. Throughout the 8 models participating in the NA case study, we can note the dominance of solar radiation followed by PBL height and dry deposition velocity whereas wind speed seems to be relevant throughout the year only for three of the eight (WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Zhang)). We note that correlation does not necessarily imply causation - the wind speed dependence effects noted here may reflect model dependence on the friction velocity, which can be expressed as a function of the wind speed, logarithmic profile, and surface roughness. The contribution of wind speed across models is very scattered in time though contributing on average for 30% of the resolved variability. In some models it appears to be among the dominant factors in winter more than in summer (WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Zhang), WRF-Chem (RIFS), WRF-Chem (UPM)). While WRF-Chem (UPM) uses the CBMZ mechanism (see Makar et al., 2025, this issue), the dry deposition implementation for CBMZ accounts only for 4 seasons, while the other two WRF-Chem models (RIFS and NCAR) employ the MOZART chemical mechanism, for which the dry deposition algorithm has tabulated entries on a monthly basis which are used in dry deposition. That is, the WRF-Chem dry deposition implementations which are linked to different gas-phase mechanisms have differing degrees of seasonal resolution.

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

It appears that in North America a seasonality in the contribution of the various components is more evident. The differences between GEM-MACH (Base) and GEM-MACH (Ops) can be attributed at least partially to the meteorology change associated with feedbacks, but also may

partially result in the differing seasonality in LAI inputs. The no-feedback model (GEM-MACH (Ops)) has less ozone variability associated with wind speed, and more with solar radiation, compared to the feedback model GEM-MACH (Base); feedbacks exacerbate meteorological variability. GEM-MACH (Base) versus GEM-MACH (Zhang) shows how much the dry deposition scheme can affect the variability, via the feedbacks: GEM-MACH (Base) and GEM-MACH (Zhang) are otherwise identical models. This quantifies the impact of feedbacks on meteorology and hence dry deposition velocity variance. WRF-Chem is also a feedback model as well, and the impact of the feedbacks is showing up as differences in the relative importance of meteorology versus ozone dry deposition velocity itself between the different implementations.

In EU we see from Figure 15a that the contributions have a greater degree of scatter than for NA. WRF-Chem (UPM) and WRF/Chem (RIFS) share an important contribution of dry deposition velocity in February and of PBL in April, November and December. Interesting is the fact that across the year the components account for a smaller portion of the total variance (<50%) than in the NA case. This could be due to drastically different conditions and the dominance emissions variability (and consequently chemistry) on the ozone variability. Each of the models are using different driving meteorology, but the variation in observed conditions across EU may be less than across NA, as noted above. The March case of WRF-Chem (RIFS) is particularly interesting where the PBL height, solar radiation, wind speed and dry deposition velocity contribute to less than 5% of the ozone variance. Another difference between the NA and EU case studies is the contribution of dry deposition compared to the other processes in determining ozone variability. In NA, dry deposition velocity contributes 10 to 25% to ozone variability during summer and 10 to 50% during winter. In the EU, however, the summer contribution is much lower and in February two models out of four show a 70% contribution.

All these results clearly point toward a relevance of dry deposition in determining ozone variability and concentrations at the surface and yet they also show that important differences are present in the process description in individual models that can greatly influence the outcome.

When the same analysis is performed at the O₃ Receptor cells, we can clearly demonstrate hypothesis (1) and possibly (2) presented in the previous section. Figures 16 for the NA case and 15b for the EU case show the results for the O₃ receptor cells. The eight models in the NA case clearly show that at those grid cells the contribution of dry deposition velocity to ozone variability is generally much smaller compared to the results for grid cells with specific common LULC types, for example with respect to Evergreen Needle-leaf Forests. Despite this general trend, NA1, NA2,

NA3 and NA5 (WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Base), GEM-MACH (Ops) respectively) still show that during winter, dry deposition can be a significant contributor to ozone concentration variability at receptor locations. 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 dry deposition results for specific LULC classes. A similar conclusion can be drawn for the EU case (Figure 15b) 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 O₃ receptor locations (EU results are shown as Figure S11 in the SM). As can be noted, existing O₃ 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.

In this respect, it is important also to note that in spite of the formal differences among dry deposition modules (Galmarini et al., 2021), in conditions of uniform LU characteristics and dominance of urban and Planted/Cultivated LULC types, the models tend to produce comparable results in terms of contributors to ozone variability. This result further underlines the importance of a correct and uniform characterization of the both the input LULC data and the extent to which monitoring station data reflect LULC across the domain, both of which are driving factors in determining the differences among dry deposition modules.

6. Conclusions

An operational evaluation has been conducted on the models that took part to the AQMEII-4 activity (Galmarini et al., 2021). A total of 12 models were analysed, 8 of which were run over the North American continental air quality simulation of the year 2016 and the rest were run over

Europe for the year 2010. The scope of the evaluation is to determine the level of agreement of the models against available measurements and how they compare with one another. This is normally referred to as operational evaluation and according to Dennis et al, (2010) is the first necessary step prior to any more detailed evaluation or inter-comparison of model results. The focus of the fourth phase of AQMEII is the analysis of the performance of dry deposition schemes in regional scale models, therefore the operational evaluation has been performed having that goal in mind. Ozone dry deposition, in particular, is the focus of this analysis. Ozone average annual concentration errors ranged between 10 and 30% in NA and between 10 and 15% in EU except for one model (35% error). Errors for NO and NO₂ were on the order of 5-10% and 10-15% respectively in NA and 15% for both pollutants in EU. The sub regional analysis confirmed these findings, considering the expected sub regional variability related to different emission patterns. The models can be distinctively grouped by performance with WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Base) and GEM-MACH(Ops) showing a better overall capacity of predicting ozone concentrations in NA followed by GEM-MACH (Zhang) and WRF-Chem (RIFS), while WRF-Chem (RIFS) and WRF-Chem (NCAR) show larger errors throughout the year and the domain. In the EU case LOTOS/EUROS outperforms the two WRF-Chem versions (RIFS and UPM) and WRF/CMAQ (STAGE). This result is also very evident from the probabilistic analysis where all combinations of possible ensembles were calculated and reflect the results of the operational evaluation.

As far as the dry deposition is concerned, a diagnostic evaluation was performed aiming at analysing the variance contribution of the different pathways to the variance of the overall ozone dry deposition fluxes. All cells covered with at least 85% of the same land-use types were considered in this analysis. Across grid cells containing mostly needleleaf forests over NA, the main example used in our study, the analysis shows the mixed response of the various dry deposition schemes adopted in the regional scale models; one group of models shows a prevailing contribution of the stomatal effective flux in determining spatial ozone flux variability, one shows the three pathways contribute rather equally, and the last group of models for which the lower canopy and soil effective flux is the prevailing contributor. Thus, models are simulating very different drivers of ozone flux variability in space, even for the same land use type. The contribution to ozone variability of wind speed, dry deposition velocity, solar radiation and boundary layer height was also investigated.

When the above-mentioned analysis was also performed for all grid cells where ozone monitors were present regardless of the LULC type, a remarkable result was found. Regardless of the EU or NA case considered, all the differences among models found for specific LULC types largely disappeared, showing a more uniform behaviour across models. This aspect was demonstrated to be attributable to a minor contribution of dry deposition at those sites in determining the ozone variability when compared with other factors. Other factors contributing to this behaviour are the presence of predominant LULC types for which dry deposition is relatively low and the uniform distribution of those types and other LULC types across the models at the observation station locations.

This result allows us to present important conclusions. The first conclusion is that the evaluation of dry deposition processes should not be conducted only at operational ozone monitoring sites. The latter's characteristics are selected on other considerations aside from dry deposition. They appear unsuitable for dry deposition algorithm evaluation. An analysis of dry deposition modelling at these sites may produce an illusory agreement among models that could be completely misleading and misrepresentative. Therefore, specific sites with a predominance of LULCs which induce high dry deposition should be selected among existing monitoring stations, or added to existing monitoring networks, for dry deposition-focused model evaluation.

The second conclusion is a recurring theme throughout AQMEII-4 regional modelling studies to date (e.g., Hogrefe et al, 2025, Makar et al., 2025), namely the necessity for a harmonisation of LULC data across regional scale air quality models, as a large diversity in the characterization of the surface is still present among all models, and this diversity has a significant impact on model performance. Considering the existence of detailed information in space and time on LULC (e.g., Copernicus Land Monitoring services, USGS, LandSat, etc.), we find the lack of agreement between models on the input land use data anachronistic and of great concern. Any interpretation of the behaviour of dry deposition schemes will be impaired by the lack of agreement of LULC masks and will inevitably include an inherent uncertainty difficult to quantify. The present situation is comparable to the one where models use different topographies or terrain elevations to the extent of including (excluding) specific reliefs or mountain ranges in (from) the domain. If there is an ambition to improve the performance of regional scale models in terms of dry deposition processes (effectively a sink in the concentration budget), the selection of up-to-date and common LULC data is a fundamental and necessary prerequisite. Considering the advances in the characterisation of land surface at very high spatial and temporal resolutions

(metre scale), such effort cannot be further delayed and should be taken on prior to any new model evaluation or intercomparison of dry deposition processes.

Disclaimer: The views expressed in this article are those of the authors and do not necessarily represent the views or policies of the U.S. Environmental Protection Agency.

Competing Interests: One of the co-authors (SG) is a member of the editorial board of Atmospheric Chemistry and Physics Author

Contributions: SG, IK, CH and PAM: study design. SG: manuscript writing. IK: analyses and plots. PAM, CH, OEC, SG: AQMEII-4 steering committee coordination. PAM and PC: GEM-MACH simulations. CH, JOB and JP: CMAQ-M3Dry, CMAQ-STAGE simulations. RB and RB: ENSEMBLE system for submission of model output, monitoring data selection and organization. AL and TB: WRF-Chem (RIFS) simulations. AH and YHC: WRF-Chem (UCAR) simulations, comments on manuscript. OEC and DS: comments on manuscript. RK: LOTOS-EUROS simulations. JLPC and RSJ: WRF-Chem (UPM) simulations, reanalysis of WRF-Chem output. UA, KM and RS: WRF-CMAQ (UH) simulations.

References

- Appel, K. W., Bash, J. O., Fahey, K. M., Foley, K. M., Gilliam, R. C., Hogrefe, C., Hutzell, W. T., Kang, D., Mathur, R., Murphy, B. N., Napelenok, S. L., Nolte, C. G., Pleim, J. E., Pouliot, G. A., Pye, H. O. T., Ran, L., Roselle, S. J., Sarwar, G., Schwede, D. B., Sidi, F. I., Spero, T. L., and Wong, D. C.: The Community Multiscale Air Quality (CMAQ) model versions 5.3 and 5.3.1: system updates and evaluation, *Geosci. Model Dev.*, 14, 2867–2897, <https://doi.org/10.5194/gmd-14-2867-2021>, 2021
- Baublitz, C. B., Fiore, A. M., Clifton, O. E., Mao, J., Li, J., Correa, G., Westervelt, D. M., Horowitz, L. W., Paulot, F., and Williams, A. P.: Sensitivity of Tropospheric Ozone Over the Southeast USA to Dry deposition, *Geophys. Res. Lett.*, 47, e2020GL087158, <https://doi.org/10.1029/2020GL087158>, 2020.
- Brunner D., N. Savage, O. Jorba, B. Eder, L. Giordano, A. Badia, A. Balzarini, R. Baró, R. Bianconi, C. Chemel, G. Curci, R. Forkel, P. Jiménez-Guerrero, M. Hirtl, A. Hodzic, L. Honzak, U. Im, C. Knote, P. Makar, A. Manders-Groot, E. van Meijgaard, L. Neal, J. L. Pérez, G. Pirovano, R. San Jose, W. Schröder, R. S. Sokhi, D. Syrakov, A. Torian, P. Tuccella, J. Werhahn, R. Wolke, K. Yahya, R. Zabkar, Y. Zhang, C. Hogrefe, S. Galmarini, Comparative analysis of meteorological performance of coupled chemistry-meteorology models in the context of AQMEII phase 2, *Atmospheric Environment*, Volume 115, 2015, Pages 470-498, ISSN 1352-2310 <https://doi.org/10.1016/j.atmosenv.2014.12.032>.
- 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.
- Clifton, O. E., Fiore, A. M., Massman, W. J., Baublitz, C. B., Coyle, M., Emberson, L., Fares, S., Farmer, D. K., Gentine, P., Gerosa, G., Guenther, A. B., Helmig, D., Lombardozzi, D. L., Munger, J. W., Patton, E. G., Pusede, S. E., Schwede, D. B., Silva, S. J., Sörgel, M., Steiner, A. L., and Tai, A. P. K.: Dry deposition of ozone over land: processes, measurement, and

948 modeling, Rev. Geophys., 58, e2019RG000670, <https://doi.org/10.1029/2019RG000670>,
949 2020a.

950 Clifton, O. E., Paulot, F., Fiore, A. M., Horowitz, L. W., Correa, G., Baublitz, C. B., Fares, S., Goded,
951 I., Goldstein, A. H., Gruening, C., Hogg, A. J., Loubet, B., Mammarella, I., Munger, J. W., Neil,
952 L., Stella, P., Uddling, J., Vesala T., and Weng, E.: Influence of dynamic ozone dry deposition
953 on ozone pollution, J. Geophys. Res.-Atmos., 125, e2020JD032398,
954 <https://doi.org/10.1029/2020JD032398>, 2020b.

955 Clifton, O. E., Schwede, D., Hogrefe, C., Bash, J. O., Bland, S., Cheung, P., Coyle, M., Emberson, L.,
956 Flemming, J., Fredj, E., Galmarini, S., Ganzeveld, L., Gazetas, O., Goded, I., Holmes, C. D.,
957 Horváth, L., Huijnen, V., Li, Q., Makar, P. A., Mammarella, I., Manca, G., Munger, J. W., Pérez-
958 Camanyo, J. L., Pleim, J., Ran, L., San Jose, R., Silva, S. J., Staebler, R., Sun, S., Tai, A. P. K., Tas,
959 E., Vesala, T., Weidinger, T., Wu, Z., and Zhang, L.: A single-point modeling approach for the
960 intercomparison and evaluation of ozone dry deposition across chemical transport models
961 (Activity 2 of AQMEII4), Atmos. Chem. Phys., 23, 9911–9961, [https://doi.org/10.5194/acp-](https://doi.org/10.5194/acp-23-9911-2023)
962 [23-9911-2023](https://doi.org/10.5194/acp-23-9911-2023), 2023.

963 Dennis, R., Fox, T., Fuentes, M. et al. A framework for evaluating regional-scale numerical
964 photochemical modeling systems. Environ Fluid Mech 10, 471–489 (2010).
965 <https://doi.org/10.1007/s10652-009-9163-2>

966 Emmons, L. K., Walters, S., Hess, P. G., Lamarque, J.-F., Pfister, G. G., Fillmore, D., Granier, C., Guenther, A.,
967 Kinnison, D., Laepple, T., Orlando, J., Tie, X., Tyndall, G., Wiedinmyer, C., Baughcum, S. L., and Kloster,
968 S.: Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4
969 (MOZART-4), Geosci. CMAQ (Hertfordshire)ev., 3, 43–67, 2010. [https://doi.org/10.5194/gmd-3-43-](https://doi.org/10.5194/gmd-3-43-2010)
970 [2010](https://doi.org/10.5194/gmd-3-43-2010)

971 Galmarini, S., Makar, P., Clifton, O. E., Hogrefe, C., Bash, J. O., Bellasio, R., Bianconi, R., Bieser, J.,
972 Butler, T., Ducker, J., Flemming, J., Hodzic, A., Holmes, C. D., Kioutsioukis, I., Kranenburg, R.,
973 Lupascu, A., Perez-Camanyo, J. L., Pleim, J., Ryu, Y.-H., San Jose, R., Schwede, D., Silva, S., and
974 Wolke, R.: Technical note: AQMEII4 Activity 1: evaluation of wet and dry deposition schemes
975 as an integral part of regional-scale air quality models, Atmos. Chem. Phys., 21, 15663–
976 15697, <https://doi.org/10.5194/acp-21-15663-2021>, 2021.

977 Galmarini S., Kioutsioukis I., Solazzo E., Alyuz U., Balzarini A., Bellasio R., Benedictow A.M.K.,
978 Bianconi R., Bieser J., Brandt J., Christensen J.H., Colette A., Curci G., Davila Y., Dong X.,

979 Flemming J., Francis X., Fraser A., Fu J., Henze D.K, Hogrefe C., Im U., Garcia Vivanco M.,
 980 Jiménez-Guerrero P., Eiof Jonson J.-E., Kitwiroon N., Manders A., Mathur R., Palacios-Peña
 981 L., Pirovano G., Pozzoli L., Prank M., Schultz M., Sokhi R., Sudo K., Tuccella P., Takemura T.,
 982 Sekiya T., and Unal A., Two-scale multi-model ensemble: is a hybrid ensemble of opportunity
 983 telling us more? *Atmos. Chem. Phys.*, 18, 8727–8744, [https://doi.org/10.5194/acp-18-8727-](https://doi.org/10.5194/acp-18-8727-2018)
 984 [2018](https://doi.org/10.5194/acp-18-8727-2018), 2018.

985 Galmarini, S., Kioutsioukis, I., and Solazzo, E.: *E pluribus unum*: ensemble air quality predictions,
 986 *Atmos. Chem. Phys.*, 13, 7153–7182, <https://doi.org/10.5194/acp-13-7153-2013>, 2013.

987 Giordano, L., Brunner, D., Flemming, J., Hogrefe, C., Im, U., Bianconi, R., , A. Badia, A. Balzarini, R.
 988 Baró, C. Chemel, G. Curci, R. Forkel, P. Jiménez-Guerrero, M. Hirtl, A. Hodzic, L. Honzak, O.
 989 Jorba, C. Knote, J.J.P. Kuenen, P.A. Makar, A. Manders-Groot, L. Neal, J.L. Pérez, G. Pirovano,
 990 G. Pouliot, R. San José, N. Savage, W. Schröder, R.S. Sokhi, D. Syrakov, A. Torian, P. Tuccella,
 991 J. Werhahn, R. Wolke, K. Yahya, R. Žabkar, Y. Zhang,... Galmarini, S. (2015). Assessment of
 992 the MACC reanalysis and its influence as chemical boundary conditions for regional air
 993 quality modeling in AQMEII-2. *Atmospheric Environment*, 115, 371-388.
 994 [doi:10.1016/j.atmosenv.2015.02.034](https://doi.org/10.1016/j.atmosenv.2015.02.034)

995 Grell, G. A. and Freitas, S. R.: A scale and aerosol aware stochastic convective parameterization for
 996 weather and air quality modeling, *Atmos. Chem. Phys.*, 14, 5233–5250,
 997 <https://doi.org/10.5194/acp-14-5233-2014>, 2014.

998 Grell, G. A., and D. Dévényi, A generalized approach to parameterizing convection combining
 999 ensemble and data assimilation techniques, *Geophys. Res. Lett.*, 29(14),
 1000 [doi:10.1029/2002GL015311](https://doi.org/10.1029/2002GL015311), 2002.

1001 Hogrefe, C., Liu, P., Pouliot, G., Mathur, R., Roselle, S., Flemming, J., Lin, M., and Park, R. J.: Impacts
 1002 of different characterizations of large-scale background on simulated regional-scale ozone
 1003 over the continental United States, *Atmos. Chem. Phys.*, 18, 3839–3864,
 1004 <https://doi.org/10.5194/acp-18-3839-2018>, 2018.

1005 Hogrefe, C., Bash, J. O., Pleim, J. E., Schwede, D. B., Gilliam, R. C., Foley, K. M., Appel, K. W., and
 1006 Mathur, R.: An analysis of CMAQ gas-phase dry deposition over North America through grid-
 1007 scale and LULC-specific diagnostics in the context of AQMEII4, *Atmos. Chem. Phys.*, 23,
 1008 8119–8147, <https://doi.org/10.5194/acp-23-8119-2023>, 2023.

1009 Hogrefe C., Galmarini S., Makar P. A. , Kioutsioukis I., Clifton O. E., Alyuz U., Bash J. O., Bellasio R.
 1010 , Bianconi R. , Butler T., Cheung P., Hodzic A., Kranenburg R., Lupascu A., Momoh K., Perez-
 1011 Camanyo J.-L., Pleim J.E., Ryu Y.-H. , San Jose R. , Schaap M., Schwede D. B., and Sokhi R., A
 1012 Diagnostic Intercomparison of Modeled Ozone Dry deposition Over North America and
 1013 Europe Using AQMEII4 Regional-Scale Simulations, EGUsphere [preprint],
 1014 <https://doi.org/10.5194/egusphere-2025-225>, 2025.

1015 Im U., Bianconi R., Solazzo E., Kioutsioukis I., Badia A., Balzarini A., Baró R., Bellasio R., Brunner D.,
 1016 Chemel C., Curci G., Flemming J., Forkel R., Giordano L., Jiménez-Guerrero P., Hirtl M., Hodzic
 1017 A., Honzak L., Jorba O., Knote C., Kuenen J.J.P., A. Makar P., Manders-Groot A., Neal L., Pérez
 1018 J.L., Pirovano G., Pouliot G, San Jose R., Savage N., Schroder W., Sokhi R.S., Syrakov D., Torian
 1019 A., Tuccella P., Werhahn J., Wolke R., Yahya K., Zabkar R., Zhang Y., Zhang J., Hogrefe C.,
 1020 Galmarini S., Evaluation of operational on-line-coupled regional air quality models over
 1021 Europe and North America in the context of AQMEII phase 2. Part I: Ozone, Atmospheric
 1022 Environment, Volume 115, 2015, Pages 404-420, ISSN 1352-2310,
 1023 doi: [10.1016/j.atmosenv.2014.09.042](https://doi.org/10.1016/j.atmosenv.2014.09.042).

1024 Im, U., Christensen, J. H., Geels, C., Hansen, K. M., Brandt, J., Solazzo, E., Alyuz, U., Balzarini, A.,
 1025 Baro, R., Bellasio, R., Bianconi, R., Bieser, J., Colette, A., Curci, G., Farrow, A., Flemming, J.,
 1026 Fraser, A., Jimenez-Guerrero, P., Kitwiroon, N., Liu, P., Nopmongcol, U., Palacios-Peña, L.,
 1027 Pirovano, G., Pozzoli, L., Prank, M., Rose, R., Sokhi, R., Tuccella, P., Unal, A., Vivanco, M. G.,
 1028 Yarwood, G., Hogrefe, C., and Galmarini, S.: Influence of anthropogenic emissions and
 1029 boundary conditions on multi-model simulations of major air pollutants over Europe and
 1030 North America in the framework of AQMEII3, Atmos. Chem. Phys., 18, 8929–8952,
 1031 <https://doi.org/10.5194/acp-18-8929-2018>, 2018.

1032 Kioutsioukis, I. and Galmarini, S.: *De praeceptis ferendis*: good practice in multi-model ensembles,
 1033 Atmos. Chem. Phys., 14, 11791–11815, <https://doi.org/10.5194/acp-14-11791-2014>, 2014.

1034 Kioutsioukis, I., Im, U., Solazzo, E., Bianconi, R., Badia, A., Balzarini, A., Baró, R., Bellasio, R.,
 1035 Brunner, D., Chemel, C., Curci, G., van der Gon, H. D., Flemming, J., Forkel, R., Giordano, L.,
 1036 Jiménez-Guerrero, P., Hirtl, M., Jorba, O., Manders-Groot, A., Neal, L., Pérez, J. L., Pirovano,
 1037 G., San Jose, R., Savage, N., Schroder, W., Sokhi, R. S., Syrakov, D., Tuccella, P., Werhahn, J.,
 1038 Wolke, R., Hogrefe, C., and Galmarini, S.: Insights into the deterministic skill of air quality

ensembles from the analysis of AQMEII data, *Atmos. Chem. Phys.*, 16, 15629–15652, <https://doi.org/10.5194/acp-16-15629-2016>, 2016.

Knote, C., Tuccella, P., Curci, G., Emmons, L., Orlando, J.J., Madronich, S., Baro, R., Jimenez-Guerrero, P., Luecken, D., Hogrefe, C., Forkel, R., Werhan, J., Hirtl, M., Perez, J.L., San Jose, R., Giordano, L., Bunner, D., Yahya, K., Zhang, Y., Influence of the choice of gas-phase mechanism on predictions of key gaseous pollutants during the AQMEII phase-2 intercomparison, *Atm. Env.*, 115, 553-568, <https://doi.org/10.1016/j.atmosenv.2014.11.066>, 2015.

Lai J., Y. Zou, J. Zhang & P. Peres-Neto. 2022. Generalizing hierarchical and variation partitioning in multiple regression and canonical analysis using the rdacca.hp R package. *Methods in Ecology and Evolution*, 13: 782–788.

Makar, P. A., Cheung, P., Hogrefe, C., Akingunola, A., Alyuz-Ozdemir, U., Bash, J. O., Bell, M. D., Bellasio, R., Bianconi, R., Butler, T., Cathcart, H., Clifton, O. E., Hodzic, A., Koutsoukis, I., Kranenburg, R., Lupascu, A., Lynch, J. A., Momoh, K., Perez-Camanyo, J. L., Pleim, J., Ryu, Y.-H., San Jose, R., Schwede, D., Scheuschner, T., Shephard, M., Sokhi, R., and Galmarini, S.: Critical Load Exceedances for North America and Europe using an Ensemble of Models and an Investigation of Causes for Environmental Impact Estimate Variability: An AQMEII4 Study, *EGUsphere* [preprint], <https://doi.org/10.5194/egusphere-2024-2226>, 2024.

Makar, P.A., Stroud, C., Akingunola, A., Zhang, J., Ren, S., Cheung, P., and Zheng, Q., Vehicle-induced turbulence and atmospheric pollution, *Atm. Chem. Phys.*, 21, 12291–12316, <https://doi.org/10.5194/acp-21-12291-2021>, 2021.

Makar, P.A., Staebler, R., Akingunola, A., Zhang, J., McLinden, C., Kharol, S.K., Pabla, B., Cheung, P., Zheng, Q., The effects of forest canopy shading and turbulence on boundary layer ozone. *Nat Commun* 8, 15243, 2017. <https://doi.org/10.1038/ncomms15243>

Makar, P.A., Gong, W., Hogrefe, C., Zhang, Y., Curci, G., Žabkar, R., Milbrandt, J., Im, U., Balzarini, A., Baró, R., Bianconi, R., Cheung, P., Forkel, R., Gravel, S., Hirtl, M., Honzak, L., Hou, A., Jiménez-Guerrero, P., Langer, M., Moran, M.D., Pabla, B., Pérez, J.L., Pirovano, G., San José, R., Tuccella, P., Werhahn, J., Zhang, J., Galmarini, S., Feedbacks between air pollution and weather, part 2: Effects on chemistry, *Atm. Env.*, 115, 499-526, 2015b. <https://doi.org/10.1016/j.atmosenv.2014.10.021>

1069 Makar, P.A., Gong, W., Milbrandt, J., Hogrefe, C., Zhang, Y., Curci, G., Žabkar, R., Im, U. Balzarini,
 1070 A., Baro, R., Bianconi, R., Cheung, P., Forkel, R., Gravel, S., Hirtl, M., Honzak, L., Hou, A.,
 1071 Jiménez-Guerrero, P., Langer, M., Moran, M.D., Pabla, B., Pérez, J.L., Pirovano, G., San José,
 1072 R., Tuccella, P., Werhahn, J., Zhang, J., and Galmarini, S., Feedbacks between air pollution
 1073 and weather, Part 1: Effects on weather, *Atm. Env.*, 115, 442-469, 2015a.
 1074 <https://doi.org/10.1016/j.atmosenv.2014.12.003>

1075 Makar, P.A., Nissen, R., Teakles, A., Zhang, J., Zheng, Q., Moran, M.D., Yau, H., diCenzo, C.,
 1076 Turbulent transport, emissions, and the role of compensating errors in chemical transport
 1077 models, *Geosci. Model Dev.*, 7, 1001-1024, 2014. [https://doi.org/10.5194/gmd-7-1001-](https://doi.org/10.5194/gmd-7-1001-2014)
 1078 [2014](https://doi.org/10.5194/gmd-7-1001-2014).

1079 Makar, P. A., Akingunola, A., Aherne, J., Cole, A. S., Aklilu, Y.-A., Zhang, J., Wong, I., Hayden, K., Li,
 1080 S.-M., Kirk, J., Scott, K., Moran, M. D., Robichaud, A., Cathcart, H., Baratzedah, P., Pabla, B.,
 1081 Cheung, P., Zheng, Q., and Jeffries, D. S.: Estimates of exceedances of critical loads for
 1082 acidifying dry deposition in Alberta and Saskatchewan, *Atmos. Chem. Phys.*, 18, 9897–9927,
 1083 <https://doi.org/10.5194/acp-18-9897-2018>, 2018.

1084 Michaud, V., Doussin, J.-F., Colomb, A., Afif, C., Borbonb, A., Camredon, M., Aumont, B., Legrand,
 1085 M., Beekmann, M., Strong HONO formation in a suburban site during snowy days, *Atm. Env.*,
 1086 116, 155-158, <https://doi.org/10.1016/j.atmosenv.2015.06.040>, 2015.

1087 Paulot, F., Malyshev, S., Nguyen, T., Crounse, J. D., Shevliakova, E., and Horowitz, L. W.:
 1088 Representing sub-grid scale variations in nitrogen dry deposition associated with land use in
 1089 a global Earth system model: implications for present and future nitrogen dry deposition
 1090 fluxes over North America, *Atmos. Chem. Phys.*, 18, 17963–17978,
 1091 <https://doi.org/10.5194/acp-18-17963-2018>, 2018.

1092 Pendlebury, D., Gravel, S., Moran, M.D., Lupu, A., Impact of chemical lateral boundary conditions
 1093 in a regional air quality forecast model on surface ozone predictions during stratospheric
 1094 intrusions, *Atm. Env.*, 174, 148-170, <https://doi.org/10.1016/j.atmosenv.2017.10.052>, 2018.

1095 Peres-Neto, P. R., Legendre, P., Dray, S., & Borcard, D. (2006). Variation partitioning of species data
 1096 matrices: Estimation and comparison of fractions. *Ecology*, 87, 2614— 2625.

1097 Silva, S. J. and Heald, C. L.: Investigating dry deposition of ozone to vegetation, *J. Geophys. Res.-*
 1098 *Atmos.*, 123, 559–573, <https://doi.org/10.1002/2017JD027278>, 2018.

1099 Solazzo E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R.,D., Appel, K.W., Bessagnet, B.,
 1100 Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X.V., Grell, G.,
 1101 Grossi, P., Hansen, A.B., Miranda, A.I., Nopmongcol, U., Prank, M., Sartelet, K.N., Schaap, M.,
 1102 Silve, J.D.r, Sokhi, R.S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S.T.,
 1103 Galmarini, S., Operational model evaluation for particulate matter in Europe and North
 1104 America in the context of AQMEII, Atmospheric Environment, Volume 53, 2012a, Pages 75-
 1105 92, ISSN 1352-2310, doi: [10.1016/j.atmosenv.2012.02.045](https://doi.org/10.1016/j.atmosenv.2012.02.045).

1106 Solazzo E., Bianconi R., Vautard R., Appel K.W., Moran M.D., Bessagnet B, Brandt J., Christensen
 1107 J.H., Chemel C., Coll I., van der Gon H.D., Ferreira J., Forkel R., Francis X.V., Grell G., Grossi
 1108 P., Hansen .B., Jeričević A., Kraljević L., Miranda .I., Nopmongcol U., Pirovano G., Prank M.,
 1109 Riccio A., Sartelet K.N., Schaap M., Silver J.D., Sokhi R.S., Vira J., Werhahn J., Wolke R.,
 1110 Yarwood G., Rao S.T., Galmarini S., Model evaluation and ensemble modelling of surface-
 1111 level ozone in Europe and North America in the context of AQMEII, Atmospheric
 1112 Environment, Volume 53, 2012b, Pages 60-74, ISSN 1352-2310,
 1113 doi: [10.1016/j.atmosenv.2012.01.003](https://doi.org/10.1016/j.atmosenv.2012.01.003).

1114 Solazzo E., C. Hogrefe, A. Colette, M. Garcia-Vivanco, and S. Galmarini [Advanced error diagnostics](#)
 1115 [of the CMAQ and Chimere modelling systems within the AQMEII3 model evaluation](#)
 1116 [framework](#) Atmos. Chem. Phys., 17, 10435–10465, [https://doi.org/10.5194/acp-17-10435-](https://doi.org/10.5194/acp-17-10435-2017)
 1117 [2017](https://doi.org/10.5194/acp-17-10435-2017), 2017a

1118 Solazzo, E., Bianconi, R., Hogrefe, C., Curci, G., Tuccella, P., Alyuz, U., Balzarini, A., Baró, R., Bellasio,
 1119 R., Bieser, J., Brandt, J., Christensen, J. H., Colette, A., Francis, X., Fraser, A., Vivanco, M. G.,
 1120 Jiménez-Guerrero, P., Im, U., Manders, A., Nopmongcol, U., Kitwiroon, N., Pirovano, G.,
 1121 Pozzoli, L., Prank, M., Sokhi, R. S., Unal, A., Yarwood, G., and Galmarini, S.: Evaluation and
 1122 error apportionment of an ensemble of atmospheric chemistry transport modeling systems:
 1123 multivariable temporal and spatial breakdown, Atmos. Chem. Phys., 17, 3001–3054,
 1124 <https://doi.org/10.5194/acp-17-3001-2017>, 2017b.

1125 Solazzo, E., Bianconi, R., Pirovano, G., Moran, M. D., Vautard, R., Hogrefe, C., Appel, K. W.,
 1126 Matthias, V., Grossi, P., Bessagnet, B., Brandt, J., Chemel, C., Christensen, J. H., Forkel, R.,
 1127 Francis, X. V., Hansen, A. B., McKeen, S., Nopmongcol, U., Prank, M., Sartelet, K. N., Segers,
 1128 A., Silver, J. D., Yarwood, G., Werhahn, J., Zhang, J., Rao, S. T., and Galmarini, S.: Evaluating
 1129 the capability of regional-scale air quality models to capture the vertical distribution of

1130 pollutants, *Geosci. Model Dev.*, 6, 791–818, <https://doi.org/10.5194/gmd-6-791-2013>,
 1131 2013.

1132 Solazzo, E., Riccio, A., Kioutsioukis, I., and Galmarini, S.: Pauci ex tanto numero: reduce redundancy
 1133 in multi-model ensembles, *Atmos. Chem. Phys.*, 13, 8315–8333,
 1134 <https://doi.org/10.5194/acp-13-8315-2013>, 2013.

1135 Vivanco M. G., Mark R. Theobald, Héctor García-Gómez, Juan Luis Garrido, Marje Prank, Wenche
 1136 Aas, Mario Adani, Ummugulsum Alyuz, Camilla Andersson, Roberto Bellasio, Bertrand
 1137 Bessagnet, Roberto Bianconi, Johannes Bieser, Jørgen Brandt, Gino Briganti, Andrea
 1138 Cappelletti, Gabriele Curci, Jesper H. Christensen, Augustin Colette, Florian Couvidat,
 1139 Cornelis Cuvelier, Massimo D'Isidoro, Johannes Flemming, Andrea Fraser, Camilla Geels, Kaj
 1140 M. Hansen, Christian Hogrefe, Ulas Im, Oriol Jorba, Nutthida Kitwiroon, Astrid Manders,
 1141 Mihaela Mircea, Noelia Otero, Maria-Teresa Pay, Luca Pozzoli, Efsio Solazzo, Svetlana Tsyro,
 1142 Alper Unal, Peter Wind, and Stefano Galmarini Modeled dry deposition of nitrogen and
 1143 sulfur in Europe estimated by 14 air quality model systems: evaluation, effects of changes in
 1144 emissions and implications for habitat protection *Atmos. Chem. Phys.*, 18, 10199–10218,
 1145 <https://doi.org/10.5194/acp-18-10199-2018>, 2018.

1146 van der Gon, J.A.C., Bergstrom, R., Fountoukis, C., Johansson, C., Pandis, S.N., Simpson, D., and
 1147 Visschedijk, A.J.H., Particulate emissions from residential wood combustion in Europe –
 1148 revised estimates and an evaluation, *Atm. Chem. Phys.*, 15, 6503–6519,
 1149 <https://doi.org/10.5194/acp-15-6503-2015>, 2015.

1150 Vautard, R., Moran, M., Solazzo, E., Gilliam, R., Volker, M., Bianconi, R., Chemel, C., Ferreira, J.,
 1151 Geyer, B., Hansen, A.B., Jericevic, A., Prank, M., Segers, A., Silver, J.D., Werhahn, J., Wolke,
 1152 R., Rao, S.T., Galmarini, S., 2012. Evaluation of the meteorological forcing used for the Air
 1153 Quality Model Evaluation International Initiative (AQMEII) air quality simulations. *Atmos.*
 1154 *Environ.* 53, 38e50.

1155 Wang, C.-T., Campbell, P. C., Makar, P., Ma, S., Ivanova, I., Baek, B. H., Hung, W.-T., Moon, Z., Tang,
 1156 Y., Baker, B., Saylor, R., and Tong, D.: Quantifying Forest Canopy Shading and Turbulence
 1157 Effects on Boundary Layer Ozone over the United States, *EGUsphere* [preprint],
 1158 <https://doi.org/10.5194/egusphere-2025-485>, 2025.

- 1159 Wong, A. Y. H., Geddes, J. A., Tai, A. P. K., and Silva, S. J.: Importance of dry deposition
1160 parameterization choice in global simulations of surface ozone, *Atmos. Chem. Phys.*, 19,
1161 14365–14385, <https://doi.org/10.5194/acp-19-14365-2019>, 2019.
- 1162 Zhang, L., Moran, M. D., Makar, P. A., Brook, J.R., and Gong, S.: Modelling gaseous dry deposition
1163 in AURAMS: a unified regional air-quality modelling system, *Atmos. Environ.*, 36(3), 537–
1164 560, 2002. [https://doi.org/10.1016/S1352-2310\(01\)00447-2](https://doi.org/10.1016/S1352-2310(01)00447-2), 2002.
- 1165 Zhang, J., Moran, M. D., Makar, P. A., and Kharol, S.: Examination of MODIS Leaf Area Index (LAI)
1166 Product for Air Quality Modelling, 19th Annual CMAS Conference, 26–30 October 2020,
1167 https://www.cmascenter.org/conference/2020/slides/ZhangJ_MODIS_LAI_CMAS_2020.pdf
1168 f (last access: 6 June 2025), 2020.

1185 Table 1: Institutions in charge and the models used in AQMEII-4 case studies.

Abbreviation	Modeling System (dep. scheme)	Domain	Modeling Group	Dry deposition Scheme	LU for Dry deposition Scheme
NA1 (10700)	WRF/CMAQ (M3Dry)	NA	U.S. EPA	M3Dry	MODIS
NA2 (10701)	WRF/CMAQ (STAGE)	NA	U.S. EPA	STAGE	AQMEII-4 (mapped from MODIS)
NA3 (10703)	GEM-MACH (Base)	NA	Environment and Climate Change Canada	Wesely	Robichaud (Robichaud et al. 2020)
NA4 (10704)	GEM-MACH (Zhang)	NA	Environment and Climate Change Canada	Zhang	Zhang et al. (2003)
NA5 (10705)	GEM-MACH (Ops)	NA	Environment and Climate Change Canada	Wesely	Robichaud (Robichaud et al. 2020)
NA6 (10702)	WRF-Chem (RIFS)	NA	Research Center for Sustainability (RIFS)	Wesely	USGS24
NA7 (10708)	WRF-Chem (UPM)	NA	Technical University of Madrid (UPM)	Wesely	USGS24
NA8 (10709)	WRF-Chem (NCAR)	NA	National Center for Atmospheric Research / Yonsei University	Wesely	USGS24
EU1 (10702)	WRF-Chem (RIFS)	EU	Research Center for Sustainability (RIFS)	Wesely	CORINE 33
EU2 (10708)	WRF-Chem (UPM)	EU	Technical University of Madrid (UPM)	Wesely	USGS24
EU3 (10707)	LOTOS/EUROS	EU	TNO	DEPAC	Mapped from Coordination of Information on the Environment (CORINE) land cover as described in Manders-Groot et al. (2023)
EU4 (10710)	WRF/CMAQ (STAGE)	EU	University of Hertfordshire	STAGE	MODIS + extended urban

1186

1187

1188

1189

1190

1191

1192

1193 **Table 2:** NA case. For all available combinations of models $\binom{m}{n}$ analysed, the table presents those that produce the minimum errors (blue
 1194 columns) as well as all other combinations that fall within 10% of that minimum error (yellow and orange columns). The minimum RMSE of 3.77
 1195 ppb is achieved by the second order $\binom{8}{2}$, all combinations of 2 models out of 8) combination of WRF/CMAQ (STAGE) and GEM-MACH (Ops) as
 1196 well as the third order $\binom{8}{3}$, all combinations of 3 models out of 8) combination of WRF/CMAQ (STAGE), GEM-MACH (Base), and GEM-MACH
 1197 (Ops). The combinations with the second lowest RMSE are shown in the dark gray. The frequency column shows the number of times each model
 1198 was part of an ensemble weighted by the number of ensembles considered.

MODEL	Model code	Frequency (%)	Order of Model Combination											
			1		2				3			4		
WRF/CMAQ (M3Dry)	NA1 (10700)	36.4				X					X	X	X	
WRF/CMAQ (STAGE)	NA2 (10701)	54.5			X		X		X	X		X	X	
GEM-MACH (Base)	NA3 (10703)	63.6		X			X	X	X		X	X	X	
GEM-MACH (Zhang)	NA4 (10704)													
GEM-MACH (Ops)	NA5 (10705)	81.8	X		X	X		X	X	X	X	X	X	
WRF-Chem (RIFS)	NA6 (10702)	9.1												X
WRF-Chem (UPM)	NA7 (10708)													
WRF-Chem (NCAR)	NA8 (10709)													
RMSE (ppb)			3.90	3.95	3.77	3.86	4.02	3.83	3.77	4.04	3.83	3.93	4.10	

1199

1200 **Table 3:** Same as Table 2a for the EU case

MODEL	Model code	Freq. (%)	Order of Model Combination				
			2		3		4
WRF-Chem (RIFS)	EU1 (10702)	60	X		X		X
WRF-Chem (UPM)	EU2 (10708)	40				X	X
LOTOS/EUROS	EU3 (10707)	100	X	X	X	X	X
WRF/CMAQ (STAGE)	EU4 (10710)	80		X	X	X	X
RMSE (ugm ⁻³)			7.51	8.15	7.92	8.11	8.17

1201

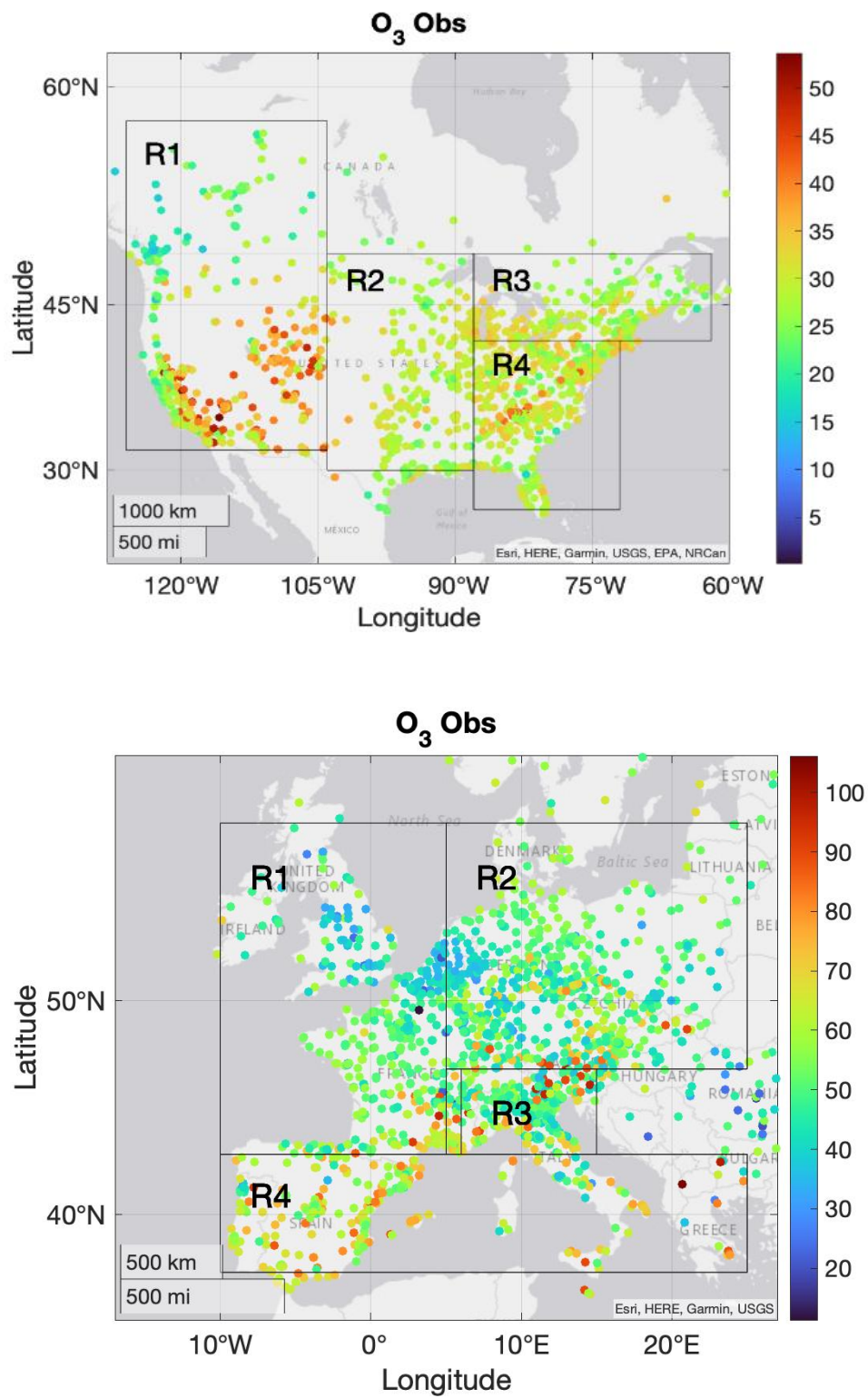


Figure 1: Annual average of ozone at all available monitoring stations in North America for 2016 (top panela) [ppb] and Europe for 2010 (bottom panelb) [µg m⁻³]. The rectangular areas represent the four selected sub-regions (R1, R2, R3, R4).

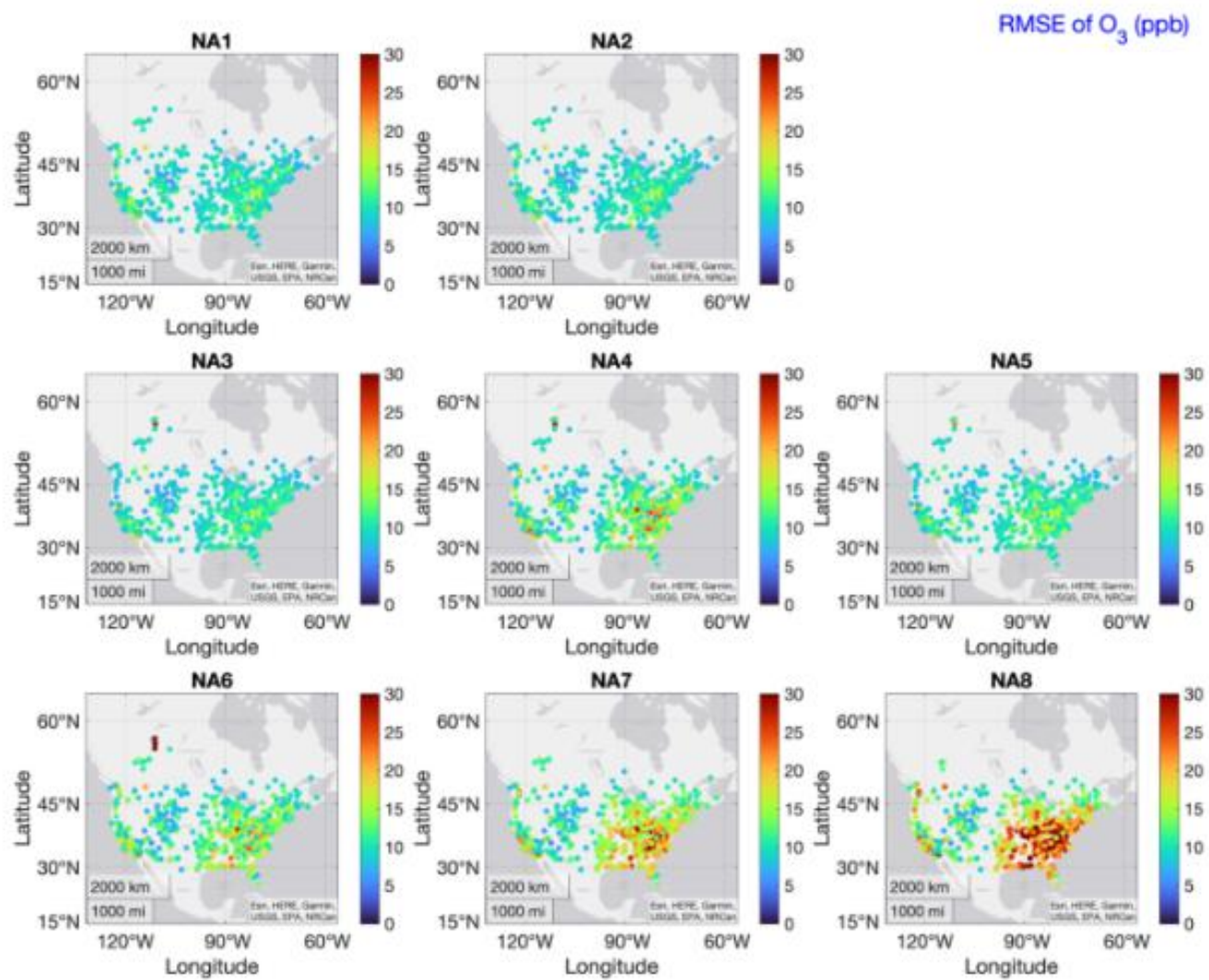
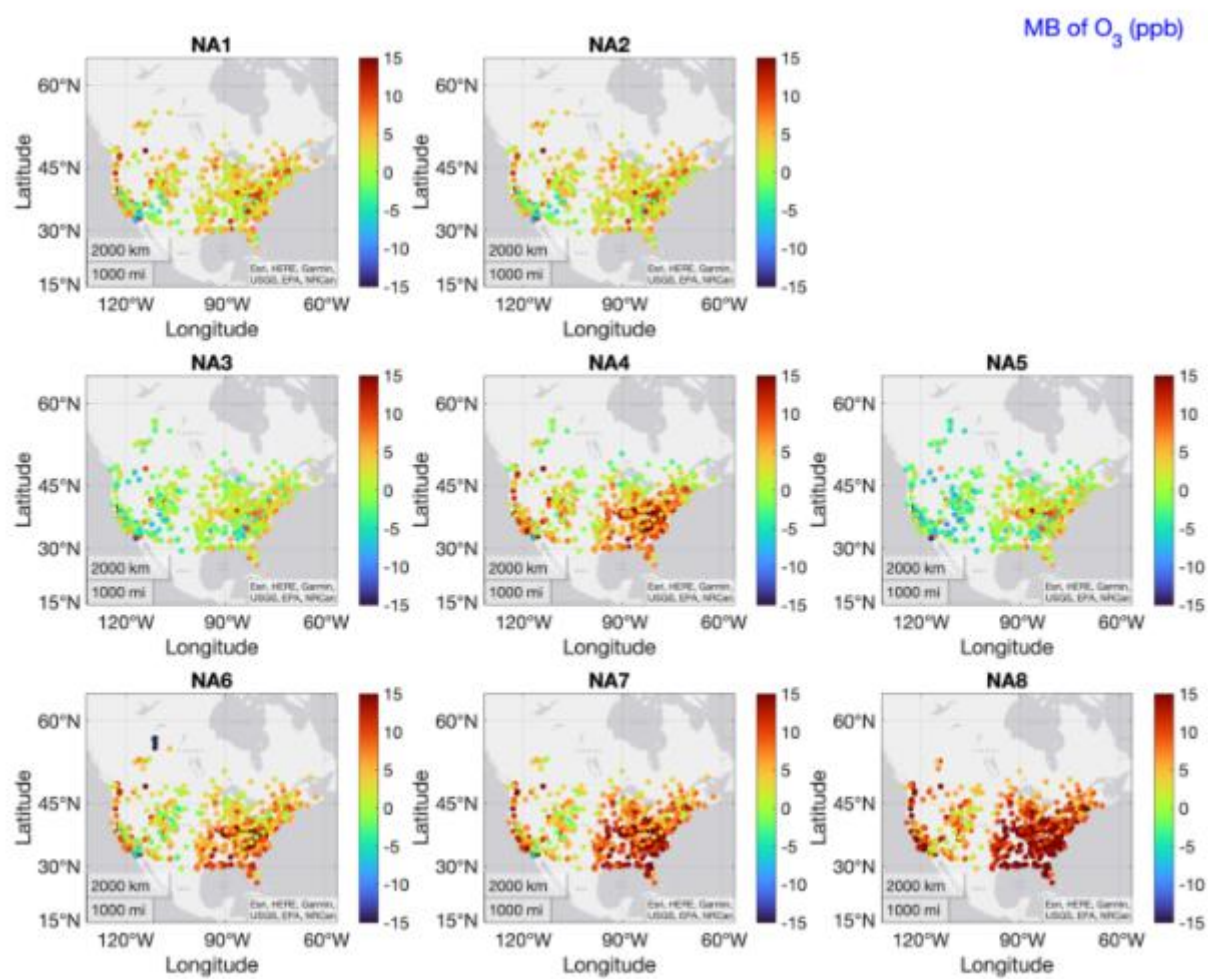


Figure 2: Individual model ozone RMSE calculated over the whole year (2016) over NA. From NA1 through NA8: WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Base), GEM-MACH (Zhang), GEM-MACH (Ops), WRF-Chem (RIFS), WRF-Chem (UPM), and WRF-Chem (NCAR). Units are in ppb.



1219

1220

1221

1222

1223

Figure 3: Individual model ozone MB calculated over the whole year (2016) over NA. From NA1 through NA8: WRF/CMAQ (M3Dry), WRF/CMAQ (STAGE), GEM-MACH (Base), GEM-MACH (Zhang), GEM-MACH (Ops), WRF-Chem (RIFS), WRF-Chem (UPM), and WRF-Chem (NCAR). Units are in ppb.

1226

1227

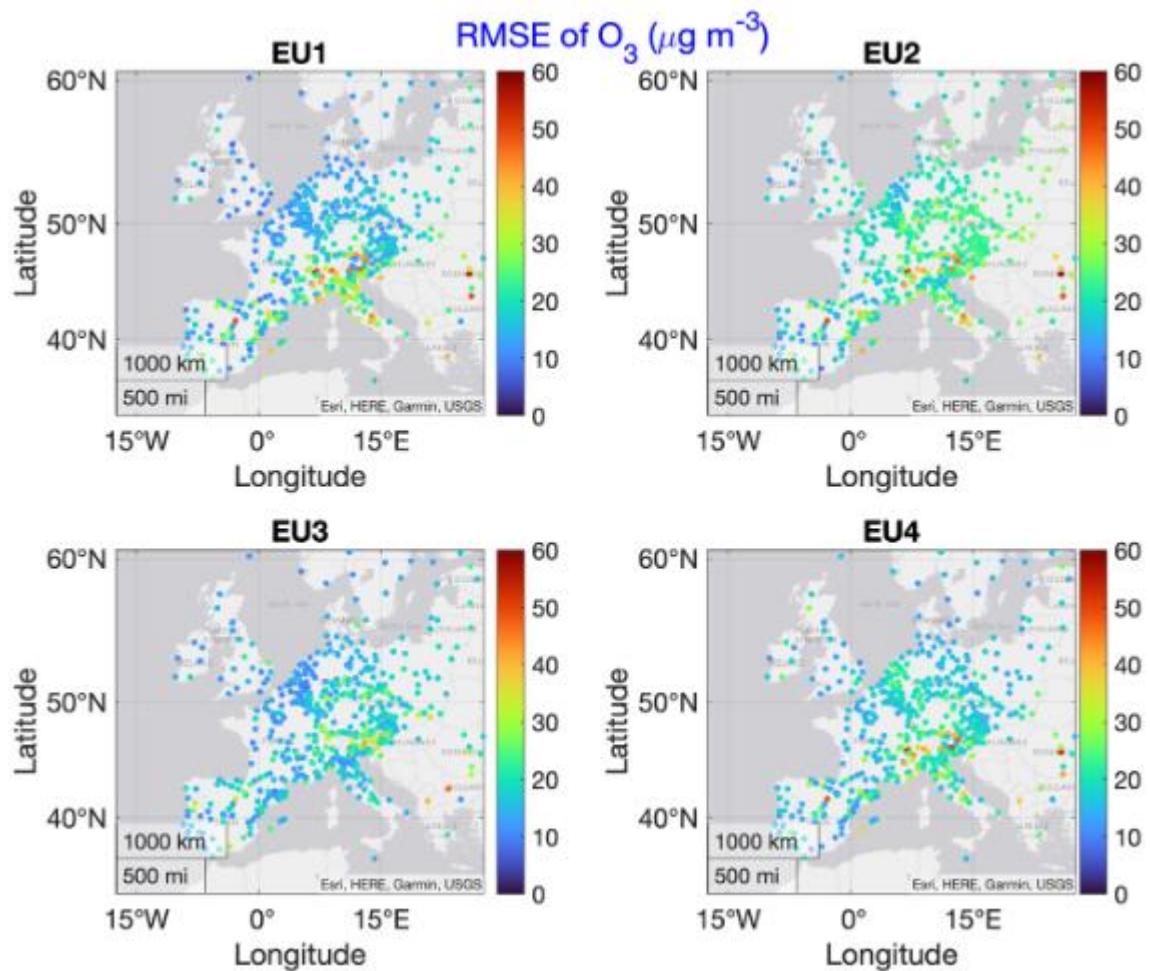
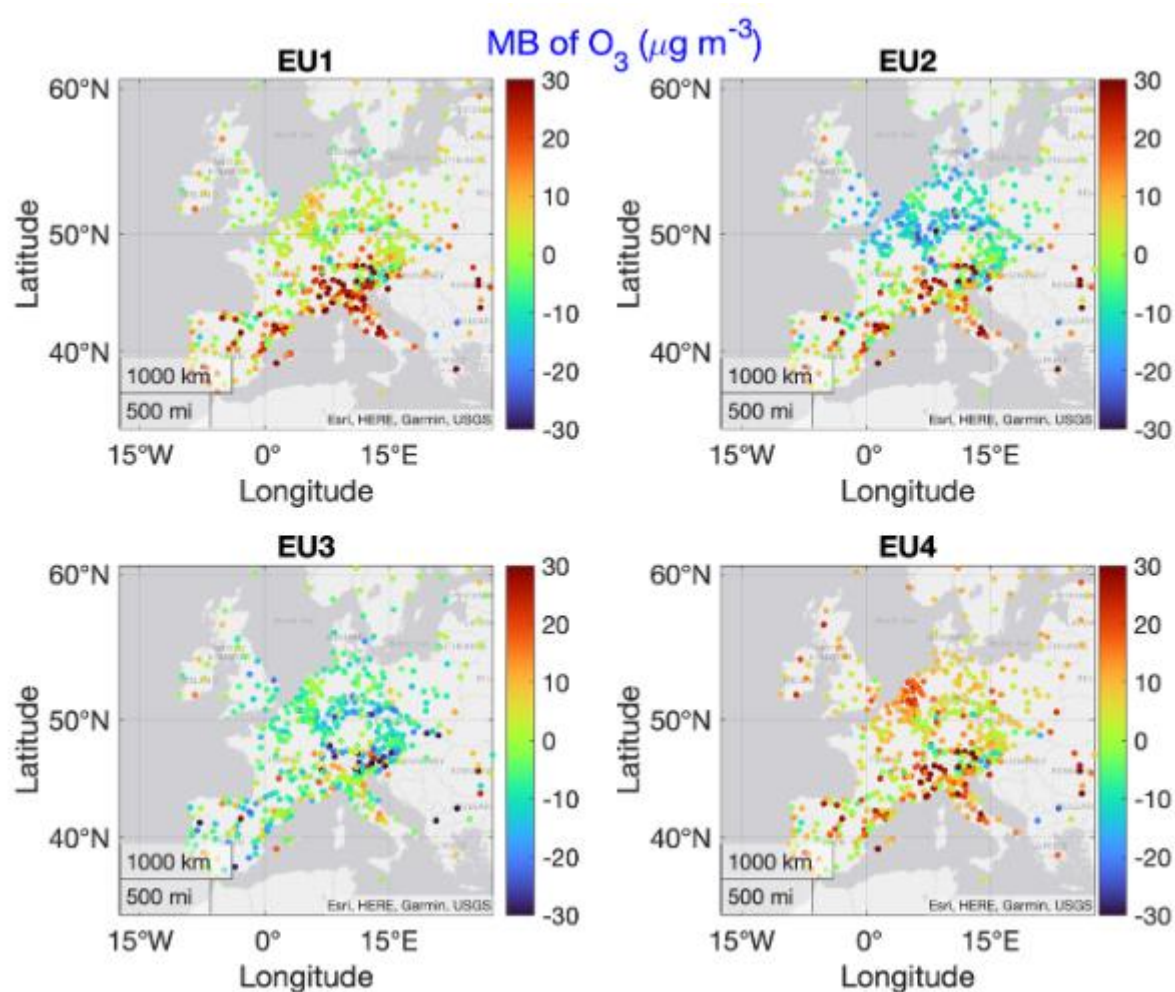


Figure 4: Individual model ozone RMSE calculated over the whole year (2010) over EU. From EU1 through EU4: WRF-Chem (RIFS), WRF-Chem (UPM), LOTOS/EUROS, WRF/CMAQ (STAGE). Units are in $\mu g/m^3$. Colour bars are set to twice the range used in Figure 2 to allow for a visual comparison across continents, accounting for the conversion factor of 1.96 between the different units.

1242



1243

1244

1245

1246

1247

1248

1249

1250

1251

1252

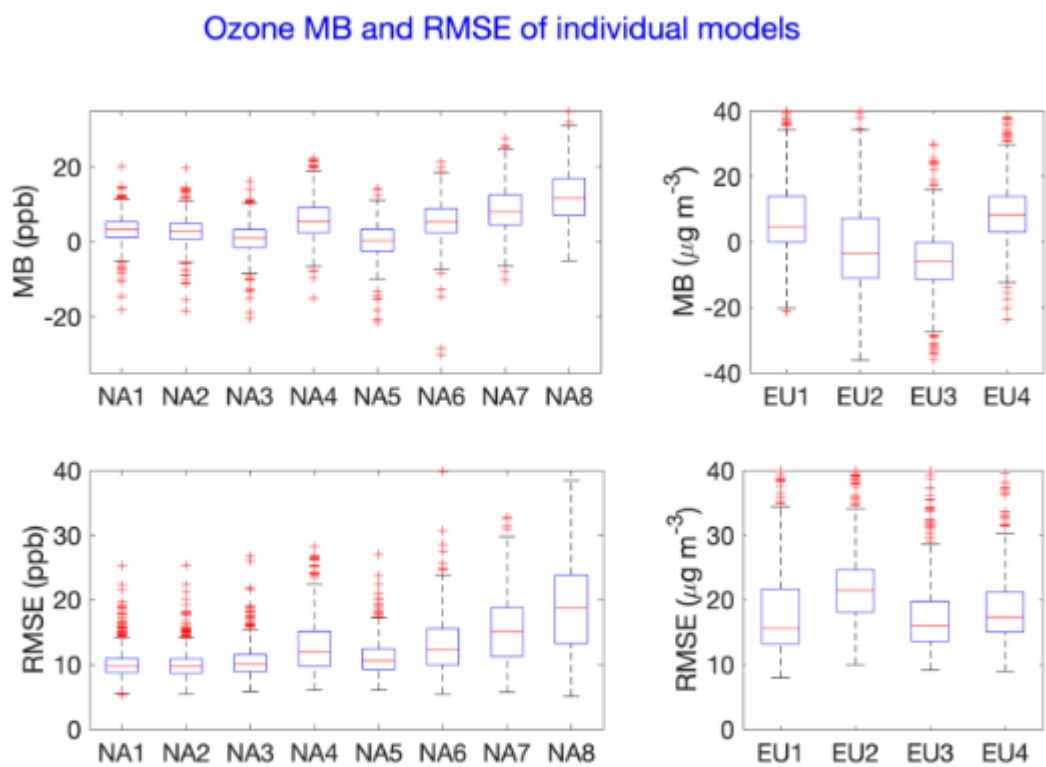
1253

1254

Figure 5: Individual model ozone MB calculated over the whole year (2010) over EU. From EU1 through EU4: WRF-Chem (RIFS), WRF-Chem (UPM), LOTOS/EUROS, WRF/CMAQ (STAGE). Units are in $\mu\text{g}/\text{m}^3$. Colour bars are set to twice the range used in Figure 2b to allow for a visual comparison across continents, accounting for the conversion factor of 1.96 between the different units.

1255

1256



1257

1258

1259

Figure 6: Individual model ozone MB (top panels) and RMSE (bottom panels) calculated over the whole year over NA (left panels) and EU (right panels). NA case units: ppb, EU: $\mu\text{g}/\text{m}^3$

1260

1261

1262

1263

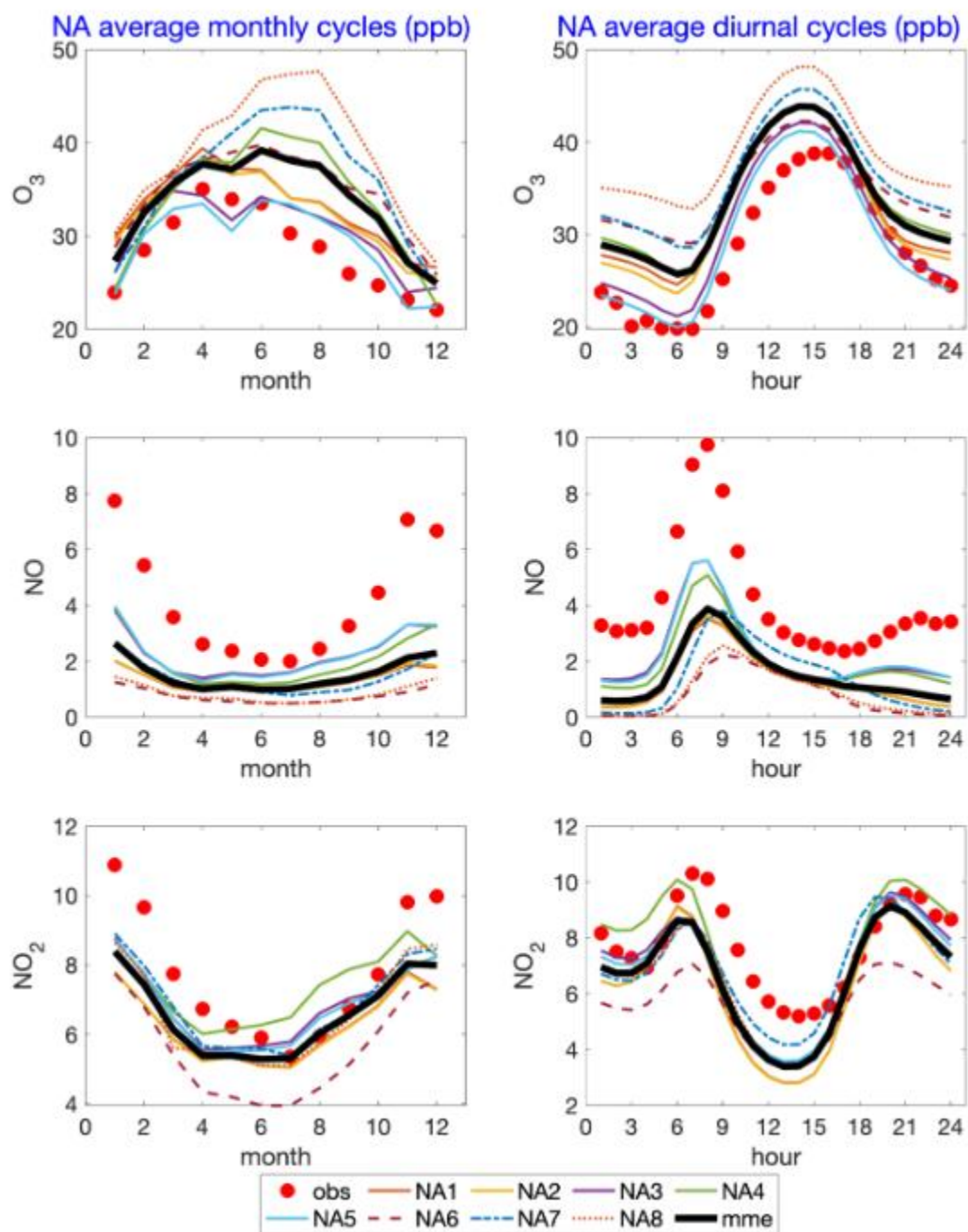


Figure 7: Average monthly (left panels) and diurnal (right panels) cycles of ozone, NO, and NO_2 [ppb] for the 2016 NA case study. Thin coloured lines (solid, dashed, dotted): models; red dots: observations; black line: multi-model mean.

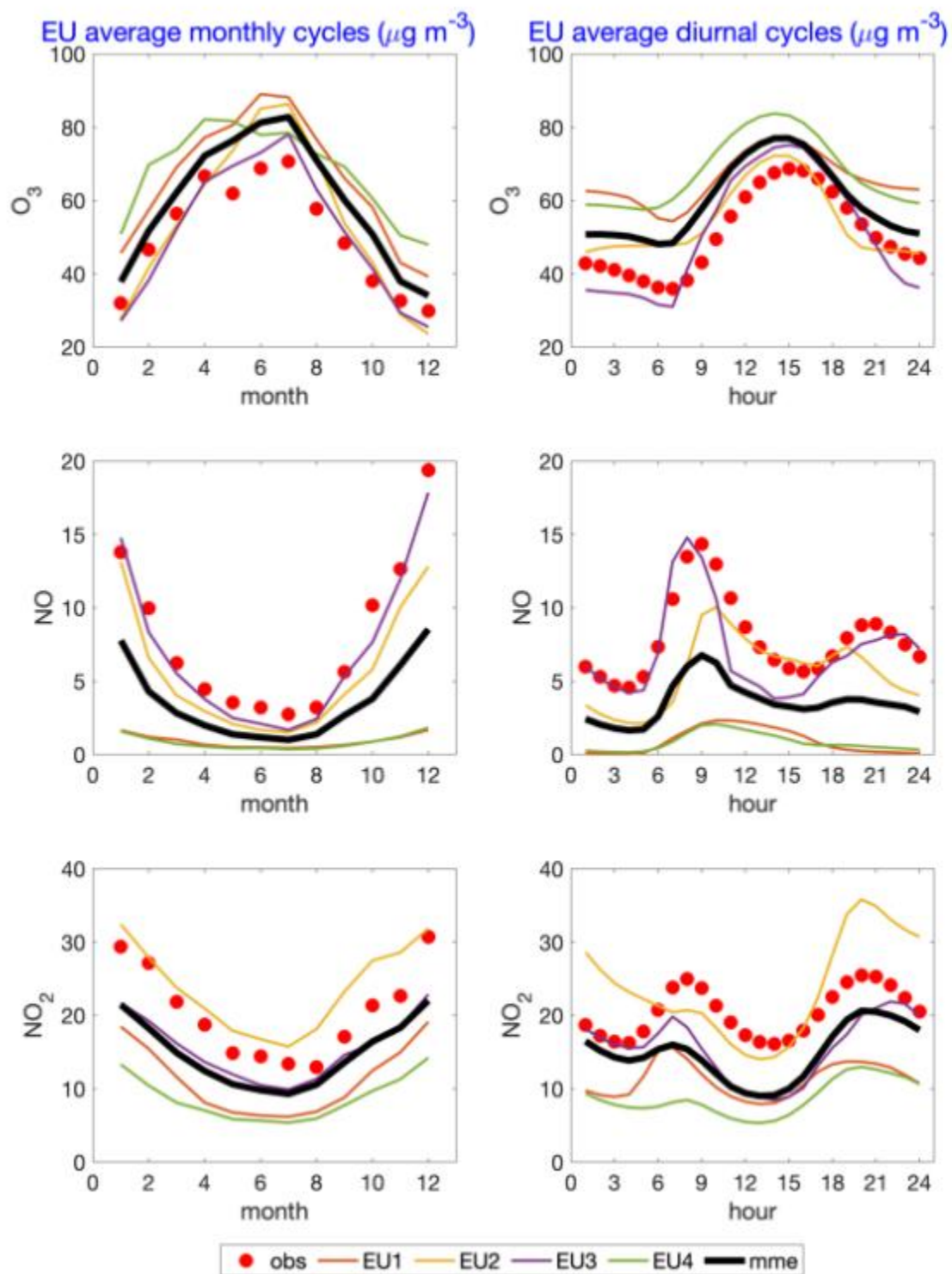


Figure 8: Average monthly (left panels) and diurnal (right panels) cycles of ozone, NO, and NO_2 [$\mu g m^{-3}$] for the 2010 EU case study. Thin coloured lines: models; red dots: observations; black line: multi-model mean.

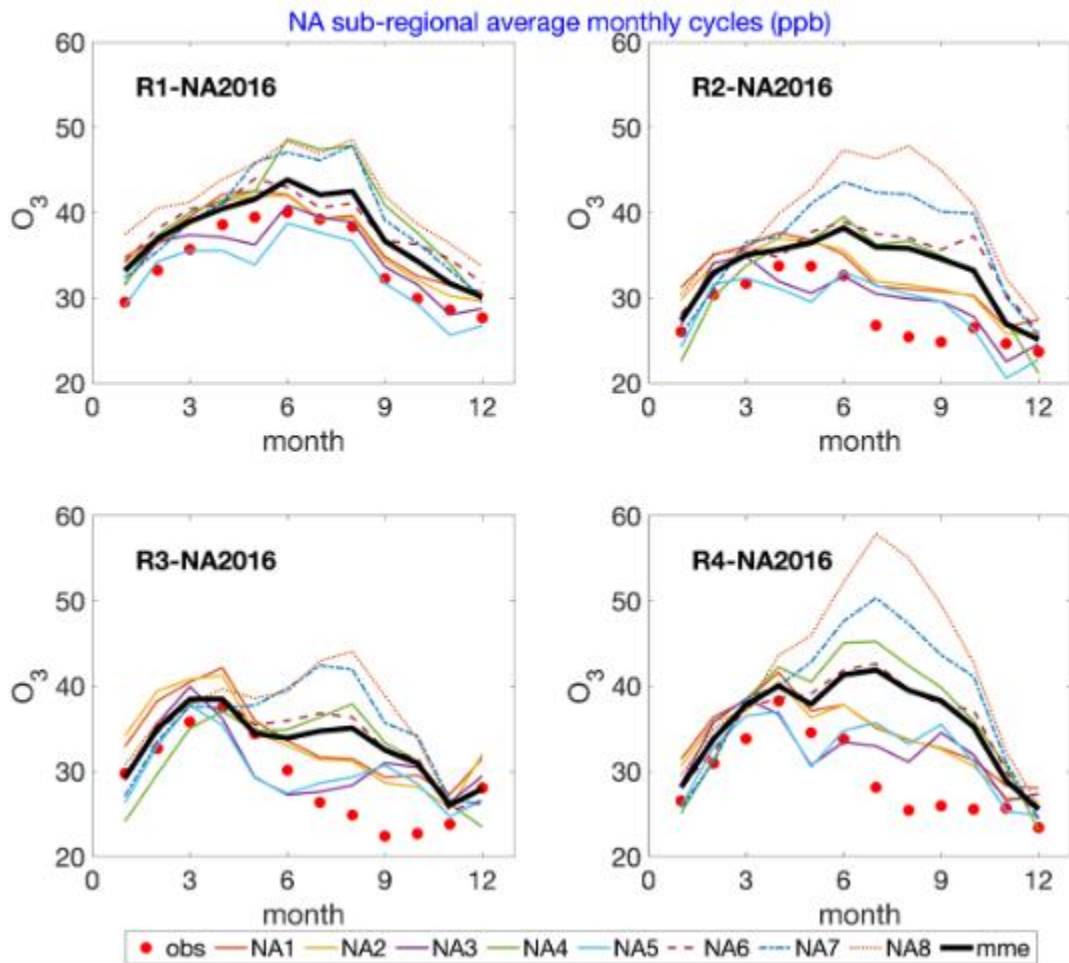


Figure 9: Monthly average cycles of O₃ concentrations in [ppb] as calculated in sub-regions R1-R4 over the NA domain. Thin coloured lines (solid, dashed, dotted): models; red dots: observations; black line: multi-model mean.

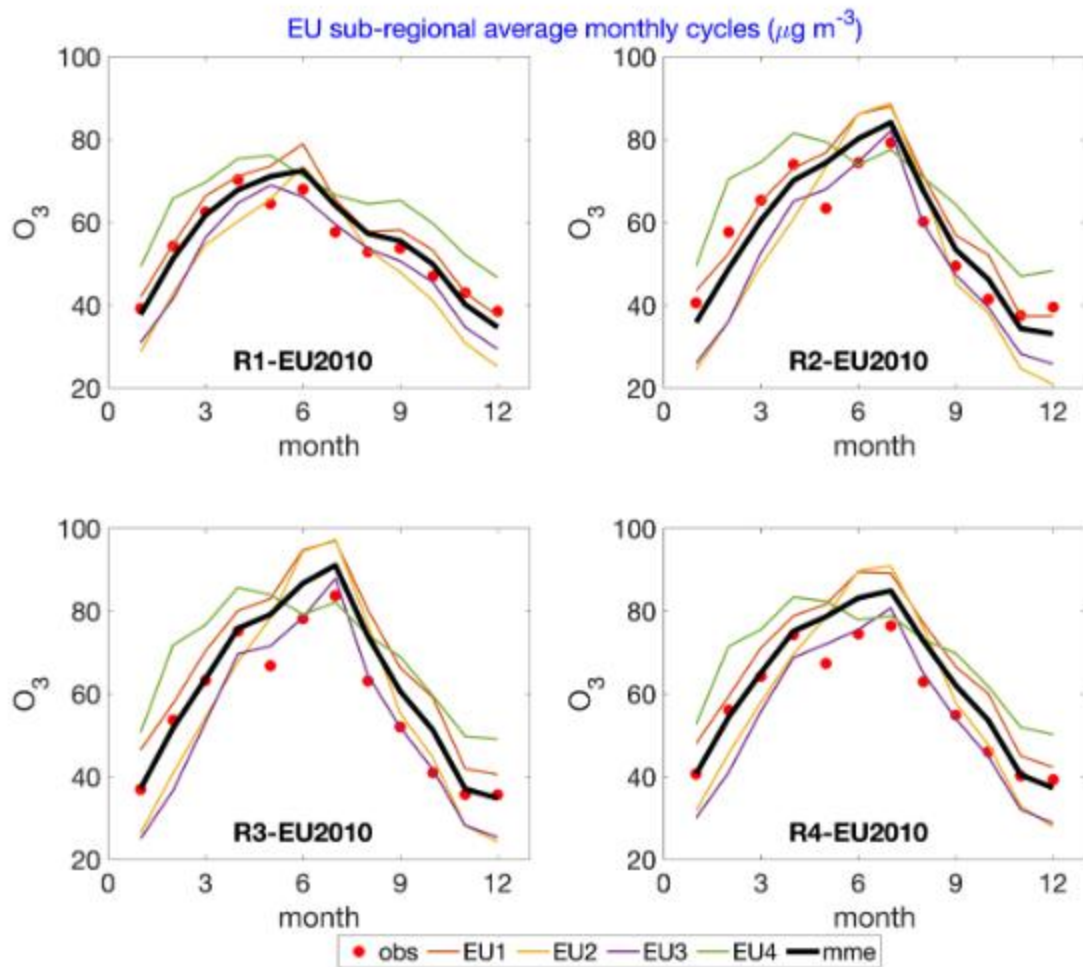


Figure 10: Monthly average cycles of O_3 concentrations in $[\mu\text{g m}^{-3}]$ as calculated in sub-regions R1-R4 over the EU domains. Thin coloured lines: models; red dots: observations; black line: multi-model median.

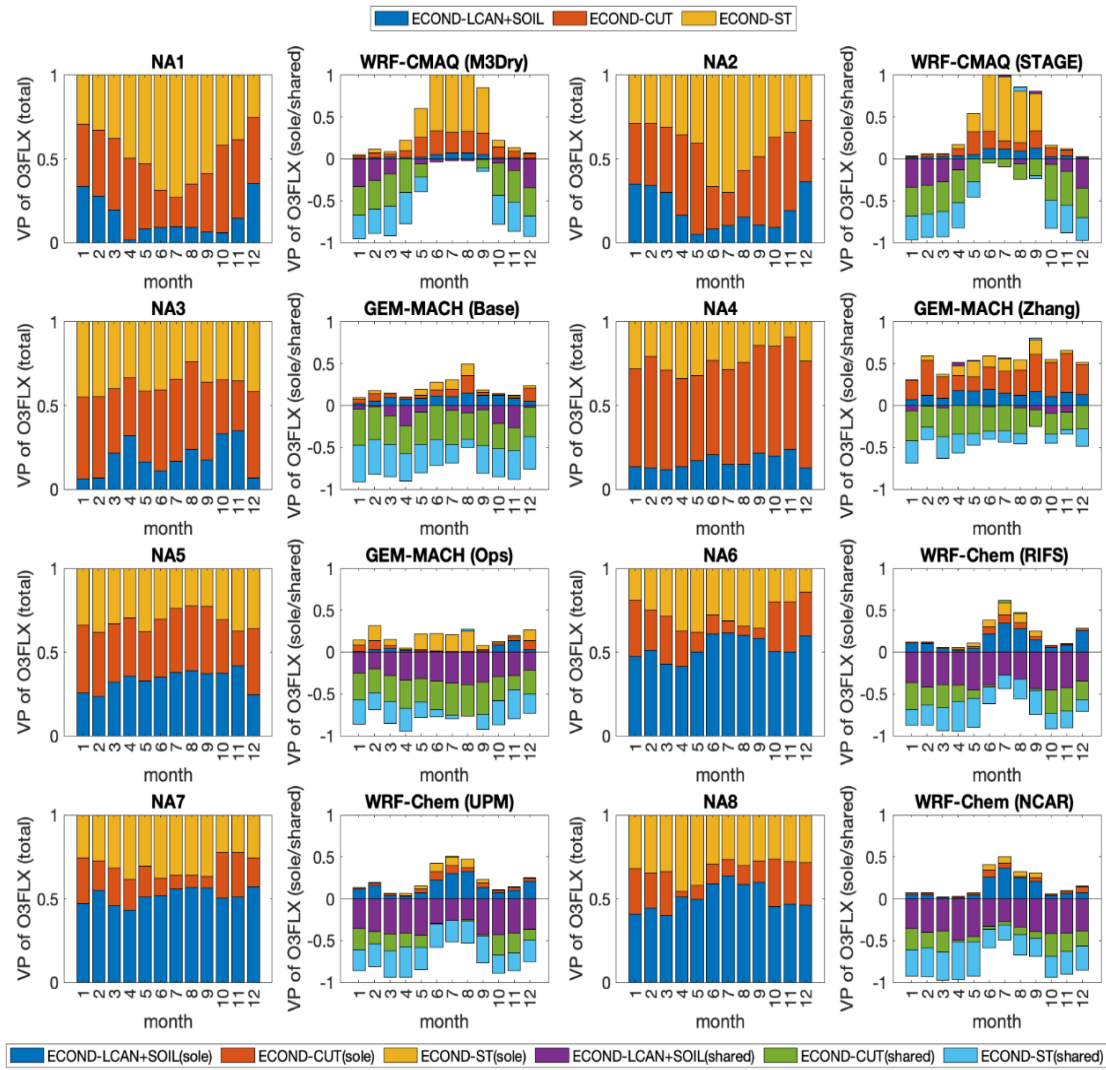


Figure 11: NA case study at 1544 shared cells covered by at least 85% of needle-leaf forest. Panels in 1st and 3rd column: variance partition (VP) of ozone dry deposition flux into the individual importance (i.e. total effect) of (1) lower canopy and soil effective fluxes combined in one factor, (2) cuticular effective flux and (3) stomatal effective flux. Panels in 2nd and 4th column: Split of the individual importance of the effective fluxes into sole and shared contributions. The shared effects are displayed with negative numbers. For the sake of making the pictures easier to read, the explicit names of the modelling systems are reported in the figure.

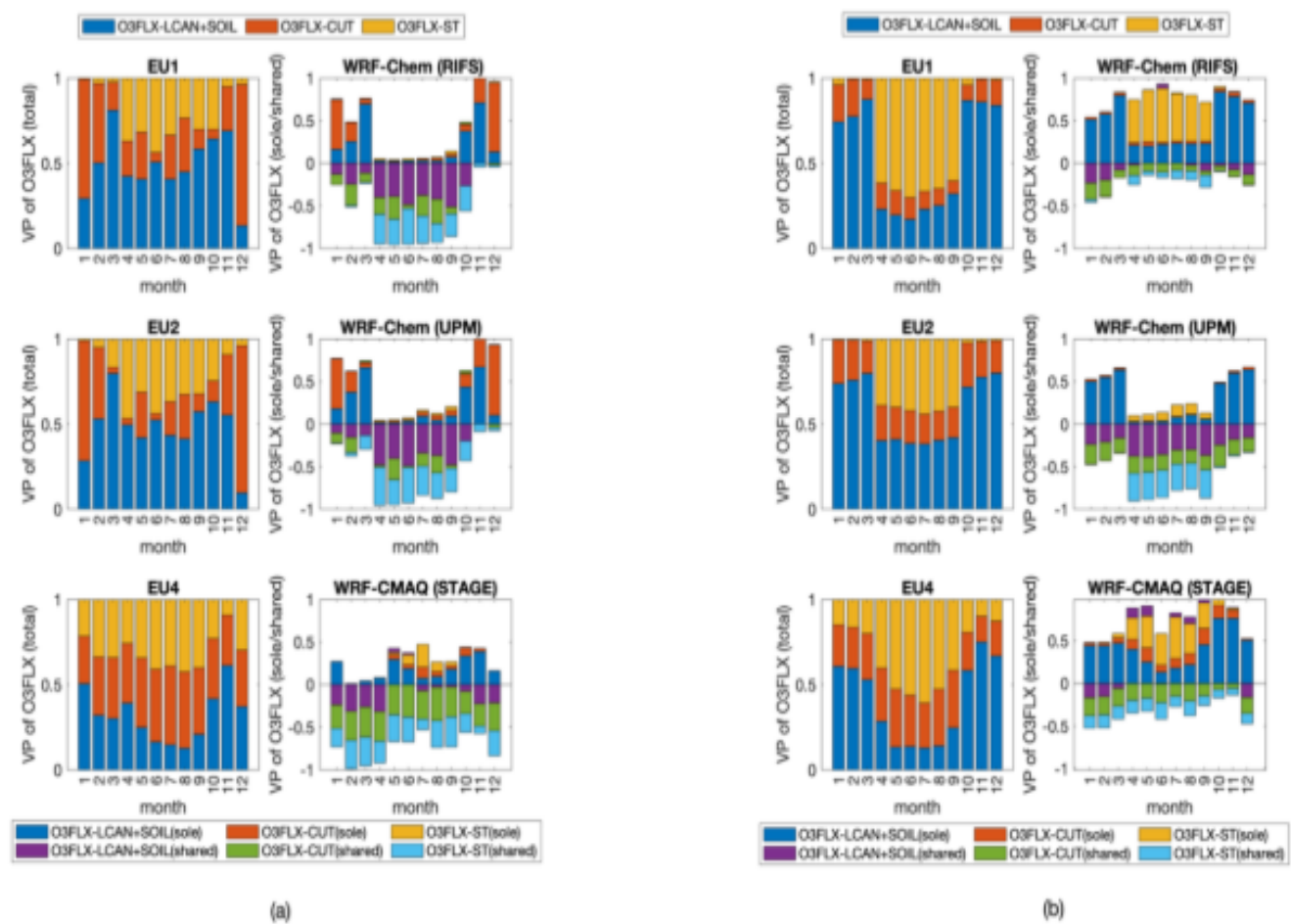


Figure 12: (a) EU case study at 2531 shared cells covered by at least 85% of needle-leaf forest. Panels in 1st column: variance partition (VP) of ozone dry deposition flux into the individual importance (i.e. total effect) of (1) lower canopy and soil effective fluxes combined in one factor, (2) cuticular effective flux and (3) stomatal effective flux. Panels in 2nd column: Split of the individual importance of the effective fluxes into sole and shared contributions. The shared effects are displayed with negative numbers. For the sake of making the pictures easier to read, the explicit names of the modelling systems are reported in the figure. (b) Same as a) but at the location of ozone receptors in EU (1551 shared cells).

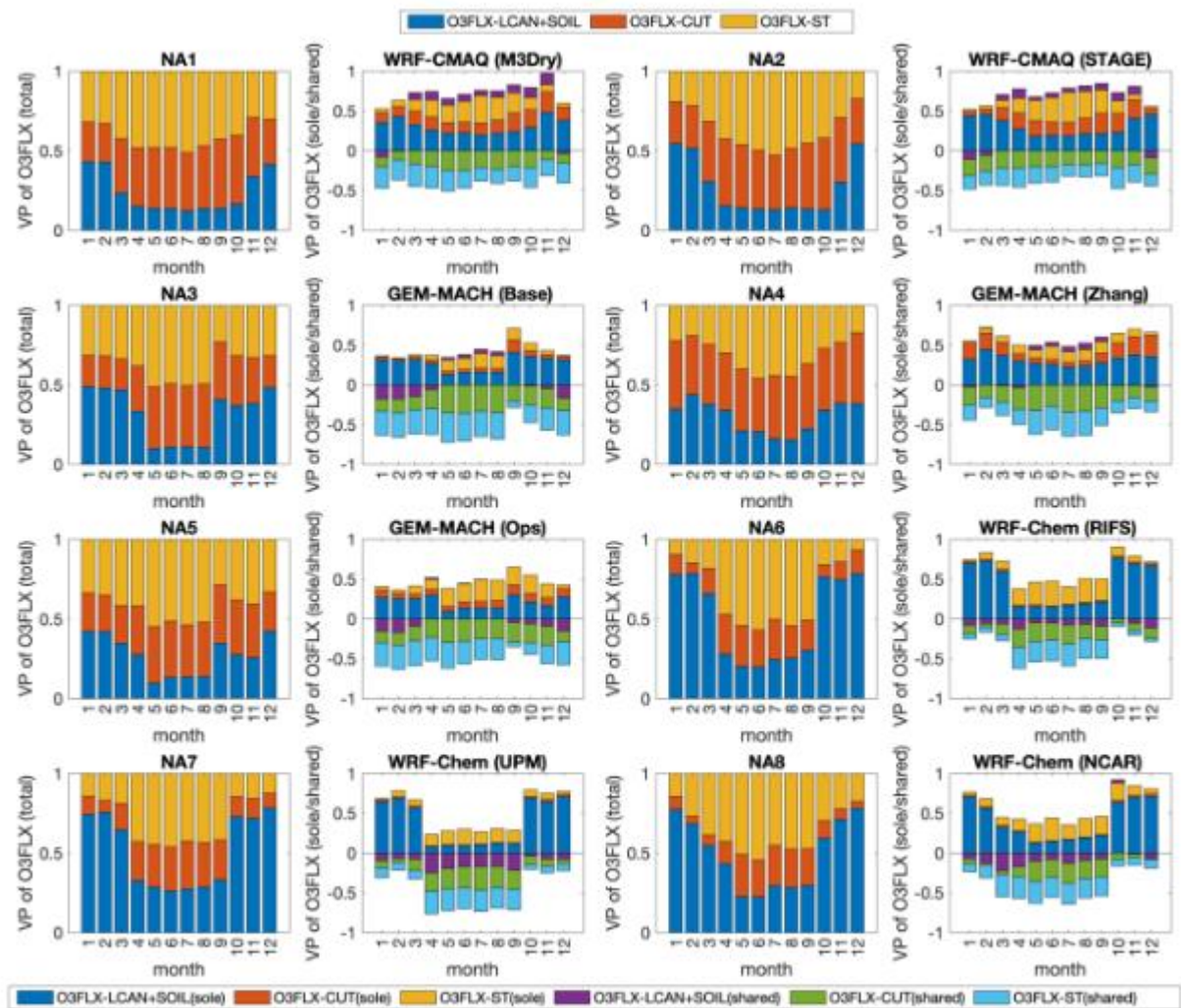


Figure 13: Same as 11 but at the location of ozone Receptors in NA (1551 shared cells).

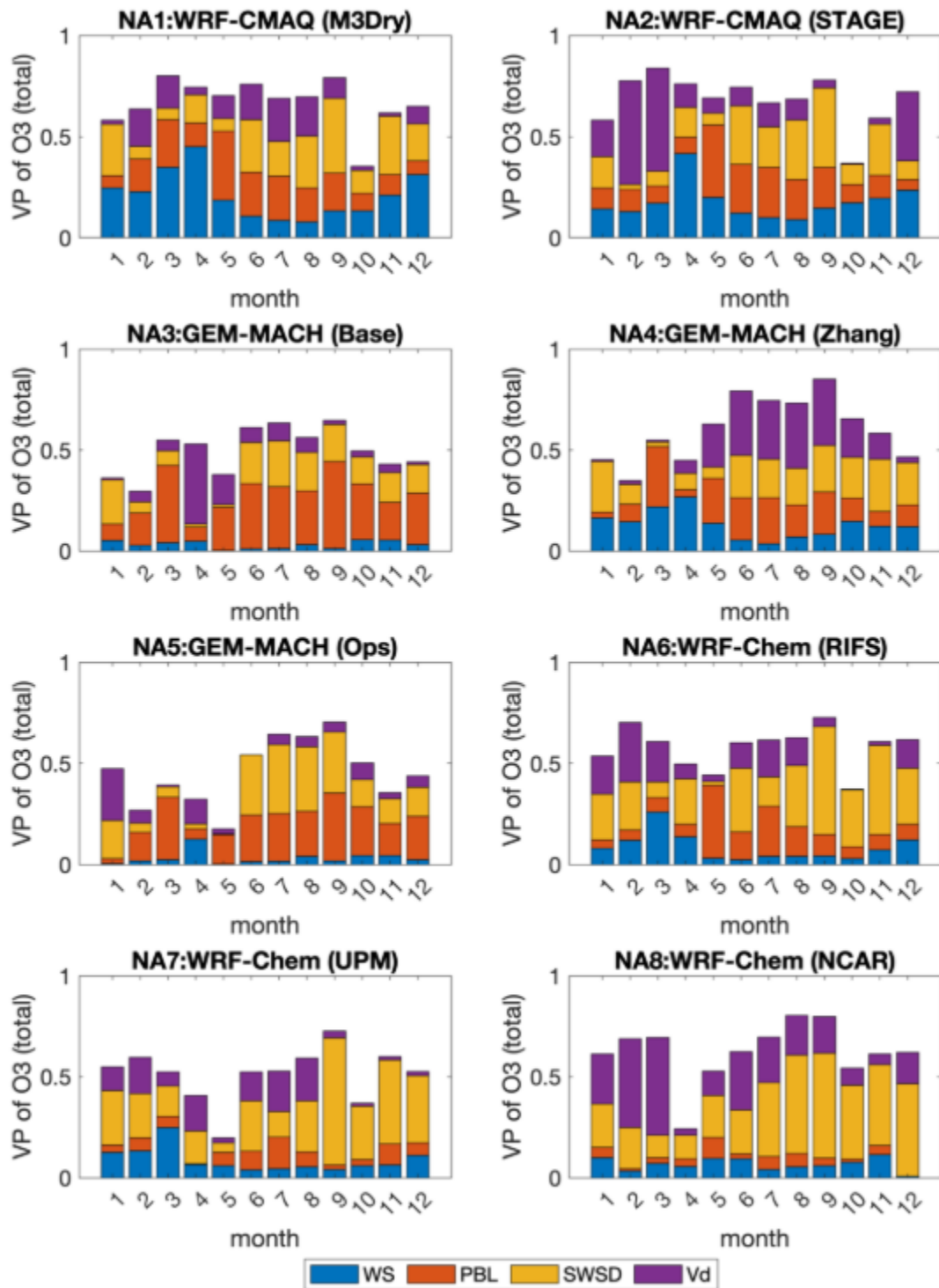


Figure 14: NA case study at 1544 shared cells covered by at least 85% of needle-leaf forest. Variance partition (VP) of ozone concentration for each model into the individual importance (i.e. total effect) of wind speed, PBL height, solar radiation, and dry deposition velocity. For the sake of making the pictures easier to read, the explicit names of the modelling systems are reported in the figure.

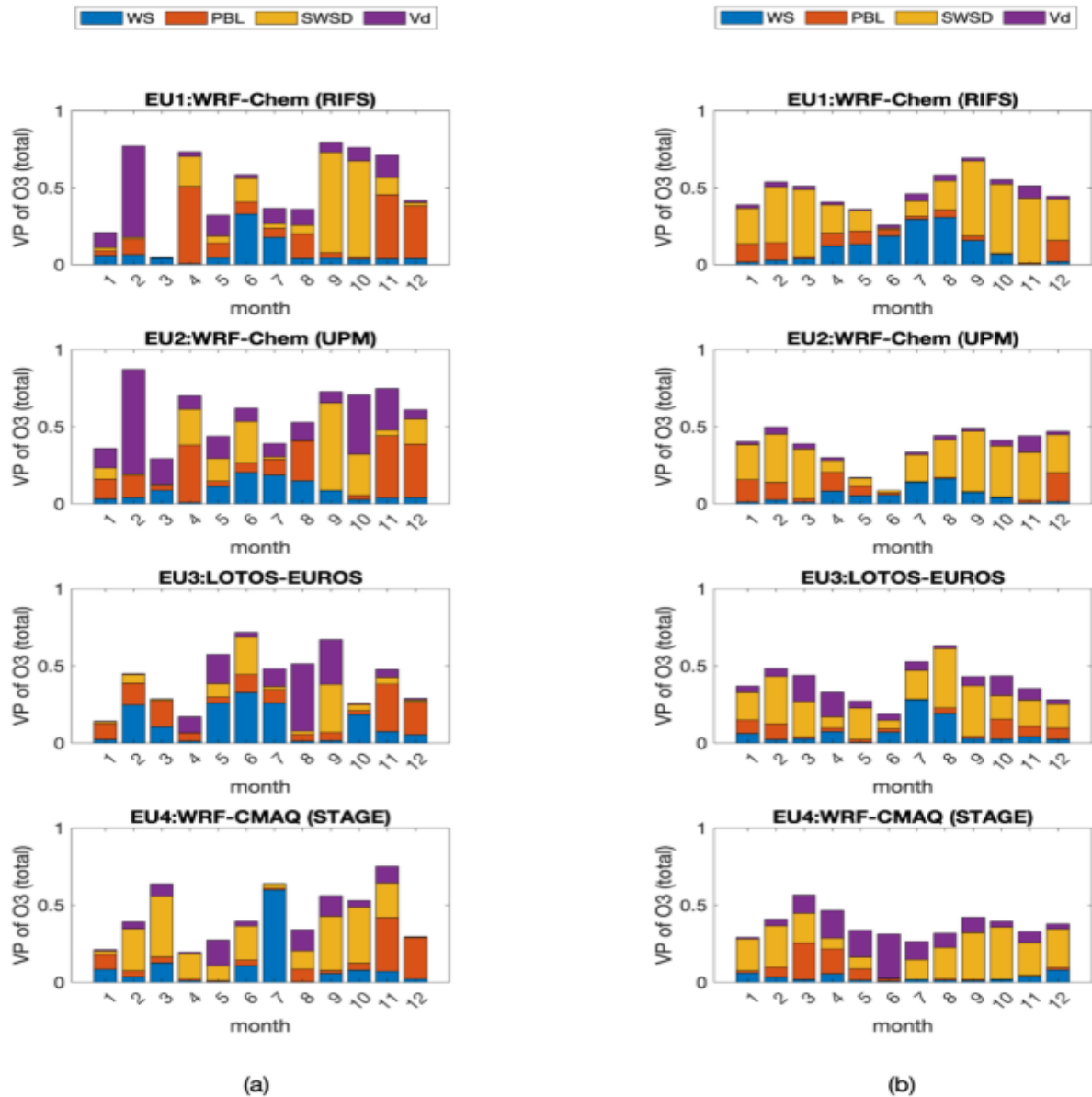


Figure 15: (a) EU case study at 2531 shared cells covered by at least 85% of needle-leaf forest. Variance partition (VP) of ozone concentration for each model into the individual importance (i.e. total effect) of wind speed, PBL height, solar radiation, and dry deposition velocity. For the sake of making the pictures easier to read, the explicit names of the modeling systems are reported in the figure. (b) Same as a) but at the location of the ozone receptors in EU (1551 shared cells).

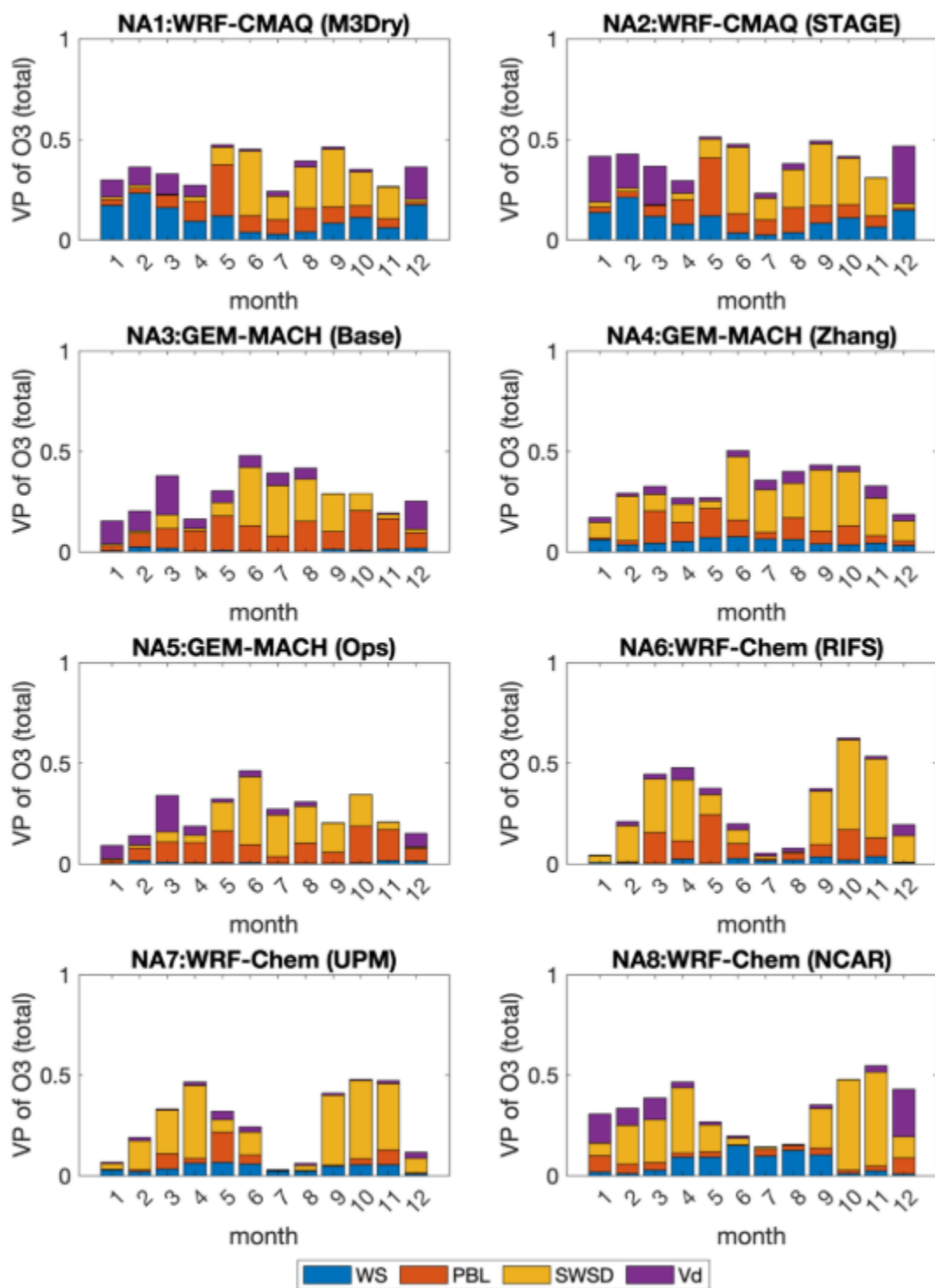
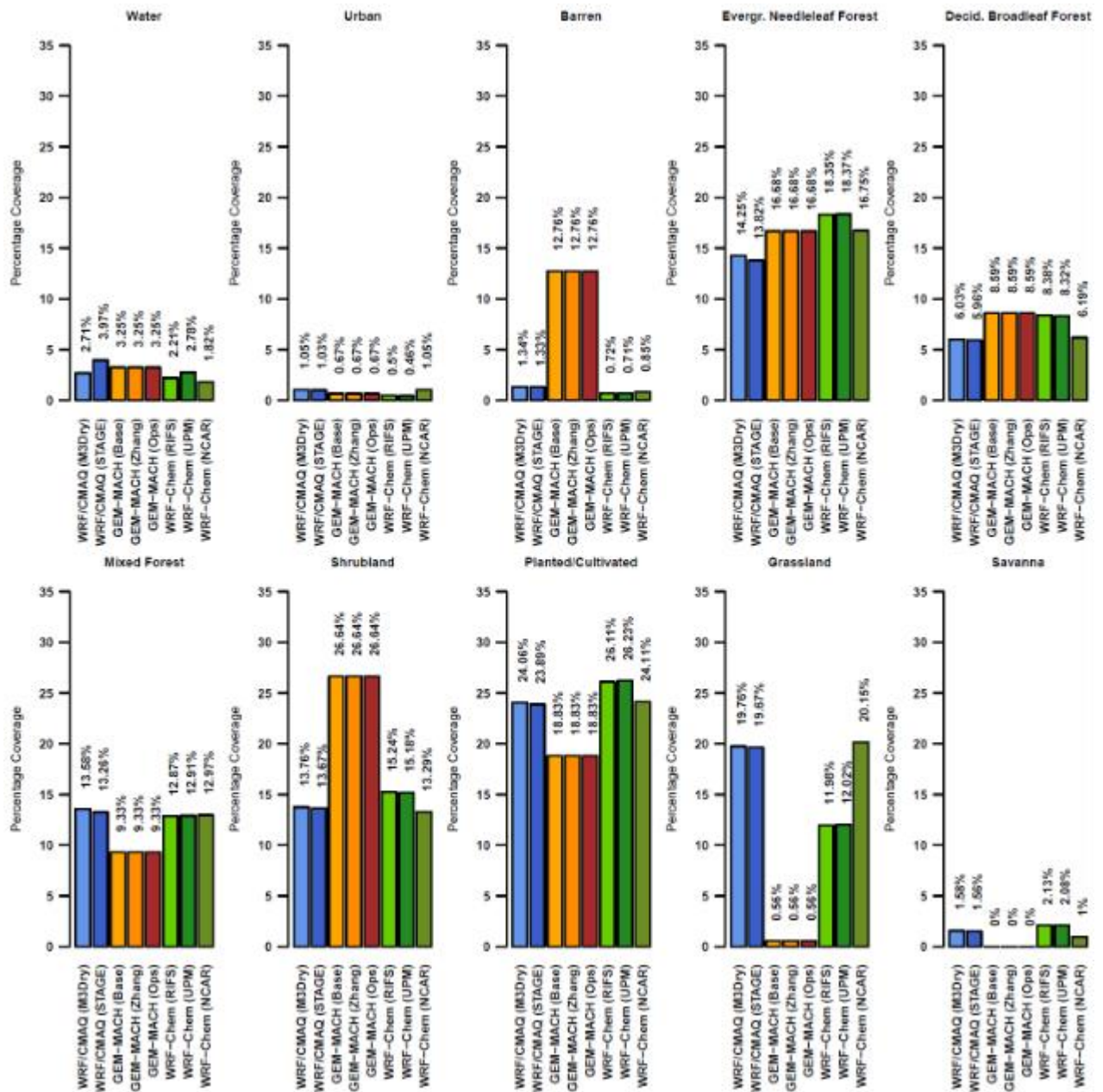


Figure 16: Same as 14 but at the location of ozone receptors in NA (1551 shared cells).

1333
1334

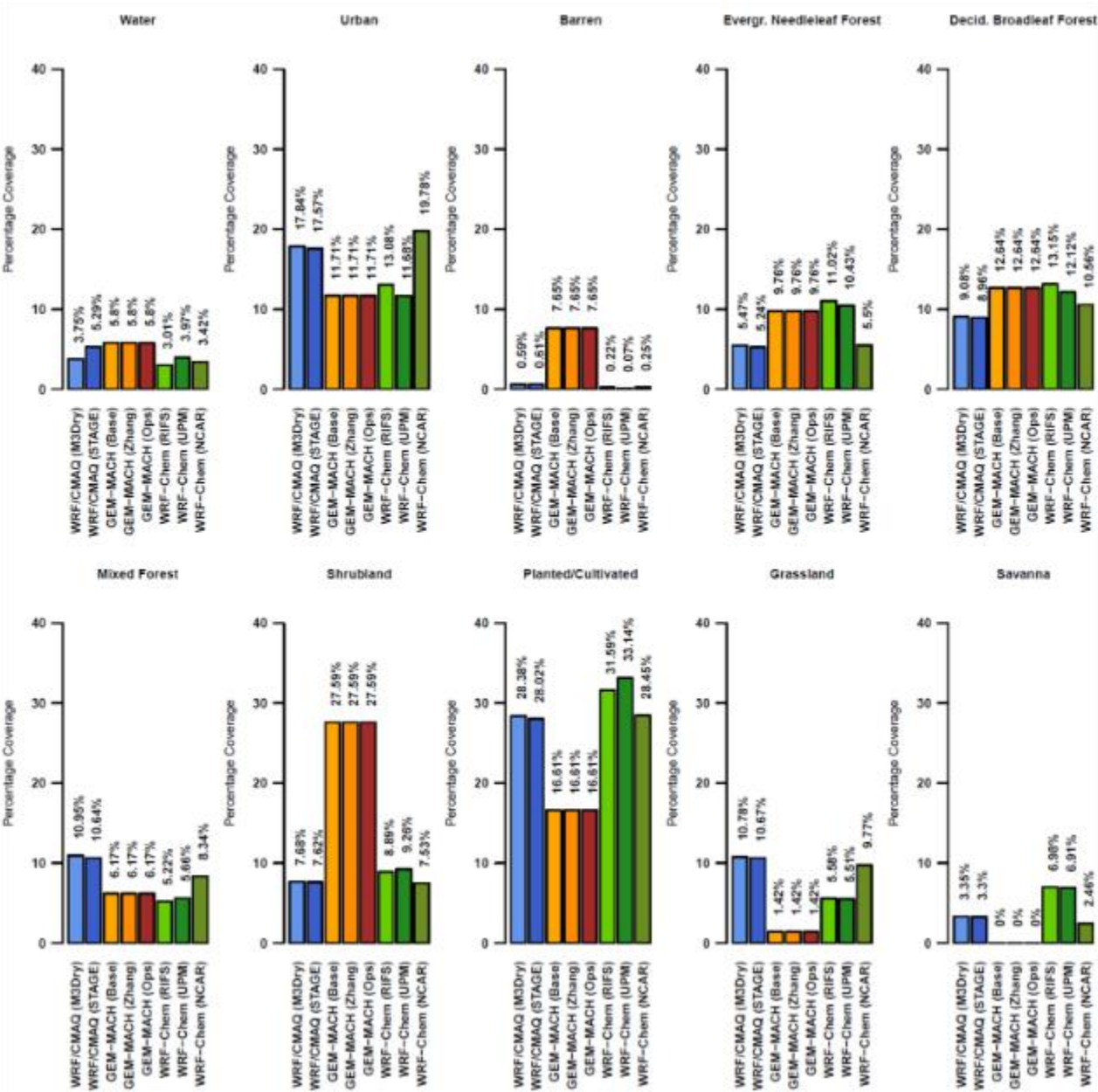
(a)



1335
1336
1337
1338
1339
1340
1341

1342

1343



1344

1345

1346

1347

1348

1349

1350

1351

(b)

Figure 17: (a) Fraction of entire NA common domain (excl. grid cells dominated by water, i.e. water fraction > 0.5) covered by each LU type. (b) Fraction of all grid cells corresponding to O3 receptor locations covered by each LU type.