Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII

Efisio Solazzo[1,] ,Roberto Bianconi[2],Robert Vautard[3], K. Wyat Appel[9], Bertrand Bessagnet[6], Jørgen Brandt[16], Jesper H. Christensen[16], Charles Chemel[11,12], Isabelle Coll[15], Hugo Denier van der Gon[20], Joana Ferreira[8], Renate Forkel[10], Xavier V. Francis[12], George Grell[18], Paola Grossi[2], Ayoe B. Hansen[22], Amela Jeričević[17], Lukša Kraljević[17], Ana Isabel Miranda[8], Michael D. Moran[14], Uarporn Nopmongcol[4], Guido Pirovano[6,7], Marje Prank[19], Angelo Riccio[21], Karine N. Sartelet[5], Martijn Schaap[20], Jeremy D. Silver[16], Ranjeet S. Sokhi[12], Julius Vira[19], Johannes Werhahn[10], Ralf Wolke[13], Greg Yarwood[4], Junhua Zhang[14], S.Trivikrama Rao[9], Stefano Galmarini[1]*

[1]Joint Research Centre, European Commission, ISPRA, Italy;
[2]Enviroware srl, via Dante 142, 20863 Concorezzo (MB), Italy
[3]IPSL/LSCE Laboratoire CEA/CNRS/UVSQ
[4]Environ International Corporation, Novato CA, USA
[5]CEREA, Joint Laboratory Ecole des Ponts ParisTech/ EDF R & D, Université Paris-Est, France
[6]Ineris, Parc Technologique Halatte, France
[7] Ricerca sistema energetico (RSE), Italy
[8]CESAM & Department of Environment and Planning, University of Aveiro, Aveiro, Portugal
[9]Atmospheric Modelling and Analysis Division, Environmental Protection Agency, NC, USA
[10]IMK-IFU, Institute for Meteorology and Climate Research-Atmospheric Environmental Division, Germany
[11]National Centre for Atmospheric Science (NCAS), University of Hertfordshire, Hatfield, UK
[12] Centre for Atmospheric & Instrumentation Research (CAIR), University of Hertfordshire, Hatfield, UK
[13] Leibniz Institute for Tropospheric Research, Leipzig, Germany
[14]Air Quality Research Division, Science and Technology Branch, Environment Canada, Toronto, Canada
[15] IPSL/LISA UMR CNRS 7583, Université Paris Est Créteil et Université Paris Diderot
[16]Department of Atmospheric Environment, National Environmental Research Institute, Aarhus University, Denmark
[17] Meteorological and Hydrological Service, Grič 3, Zagreb, Croatia
[18]CIRES-NOAA/ESRL/GSD National Oceanic and Atmospheric Administration Environmental Systems Research Laboratory Global Systems Division Boulder, Colorado USA
[19]Finnish Meteorological Institute, Helsinki, Finland
[20] Netherlands Organization for Applied Scientific Research (TNO), Utrecht, The Netherlands
[21] Department of Applied Science, University of Naples "Parthenope", Naples, Italy
[22] Department of Environmental Science, Faculty of Science and Technology, Aarhus University, Denmark

[*] Author for correspondence: E.Solazzo. Email: efisio.solazzo@jrc.ec.europa.eu

Abstract. More than ten state-of-the-art regional air quality models have been applied as part of the Air Quality Model Evaluation International Initiative (AQMEII). These models were run by twenty independent groups in Europe and North America. Standardised modelling outputs, over a full year (2006), from each group have been shared on the web distributed ENSEMBLE system, which allows for statistical and ensemble analyses to be performed. The simulations of ground-level ozone concentrations issued from the models are collectively examined in an ensemble fashion, and evaluated with a large set of observations in both continents. The scale of the exercise is unprecedented and offers a unique opportunity to investigate methodologies for generating skilful ensembles of models. Despite the remarkable progress of ensemble air quality modelling over the past decade, there still are outstanding questions regarding this technique. Among them, what is the best and most beneficial way to build an ensemble of members? How to determine the optimum size of the ensemble in order to capture data variability as well as keeping the error low? We try to address these questions by looking at optimal ensemble size and quality of the members. The analysis carried out is based on systematic minimization of the model error and it is of direct relevance for diagnostic/probabilistic model evaluation. We show that the most commonly used multi-model approach, namely the average over all available members, can be outperformed by subsets of members optimally selected in terms of bias, error, and correlation. More importantly, this result does not strictly depend on the skills of the individual members, but requires the inclusion of low ranking-skill members. We apply a methodology to discern among members and to build a skilful ensemble based on model association and data clustering, which makes no use of priori knowledge of model skill. Results show that while the methodology needs further refinements, by optimally selecting the cluster distance and association criteria, this approach can be useful for model applications beyond those strictly related to model evaluation, such as air quality forecasting.

Keywords: AQMEII*,* Clustering, Error minimisation, Multi-model ensemble, Ozone

1. Introduction

Regional air quality models have undergone considerable development over the past three decades, mainly driven by the increased concern regarding the impact of air pollution on human health and ecosystems (Rao et al., 2011). This is particularly true for ozone and particulate matter (e.g., Holloway et al., 2003; Rao et al, 2006; Jacob and Winner 2009). Regional air quality models are now widely used for supporting emission control policy, test efficacy of abatement strategies, real-time forecasts, and integrated monitoring strategies. Moreover, ozone estimates have been used in assimilation schemes to provide further information on meteorological variables such as wind speed (e.g., Eskes, 2003). The combination of outcomes by several models (regardless of their field of application), in what is typically defined as *ensemble modelling,* has been proven to enhance skills when compared against individual model realisation (e.g. Galmarini et al. 2004; Delle Monache and Stull, 2003; van Loon et al., 2007). Although ensemble modelling is well established (both from the applied and the theoretical perspectives) and routinely used in weather forecasting, it is only during the last decade that a growing number of air quality modelling communities have joined their model outputs in multi-model (MM) combinations (Galmarini et al., 2001; Carmichael et al., 2003; Rao et al., 2011). The advantages of ensemble modelling vs. individual model reside in *i)* the mean (or median) of the ensemble is, in effect, a new model which is expected to lower the error of individual members due to mutual cancellation of errors, and *ii)* the spread of the ensemble represents a measure of the variabiliy of the model prediction (Galmarini et al., 2004; Mallet and Sportisse, 2006; Vautard et al., 2006, 2009; van Loon et al., 2007). Potempski and Galmarini (2009) also point out the scientific consensus around MM ensemble techniques as a way of extracting information from many sources and synthetically assessing their variability. In particular, mean and median offer enhanced performance, on average, compared with single model (SM) realisations (Delle Monache and Stull, 2003; Galmarini et al., 2004; McKeen et al., 2005 and others).

A MM ensemble can be generated in many ways (see, e.g., Galmarini et al., 2004), for example by varying some SM internal parameters, or by using different input data, or also by applying several different models to the same scenario. This latter case is of direct relevance in the case of model evaluation, and is one of the main focii of the Air Quality Model Evaluation International Initiative (AQMEII) (Rao et al., 2011), an international project aiming at joining the knowledge and the experiences of modelling groups in Europe and North America. Within AQMEII, standardised modelling outputs have been shared on the web distributed ENSEMBLE system, which allows statistical and ensemble analyses to be performed (Bianconi et al., 2004; Galmarini et al., 2012). A common exercise was launched for modelling communities to use their regional air quality models to retrospectively simulate the entire year 2006 for the continents of Europe and North America. Outputs from numerous regional air quality models have been submitted in the form of hourly average concentrations on a grid of points and at specific locations, allowing direct comparison with air quality measurements collected from monitoring networks (details are given in Rao et al., 2011). The focus of the AQMEII project is to test the ability of models to retrospectively simulate and not forecasting air quality. This type of evaluation, with large temporal and spatial scales, is essential to determine model performance and assessing model deficiencies (Dennis et al., 2010; Rao et al., 2011).

In this study, we analyse ozone concentration data produced by over ten state-of-the-art regional air quality models, run with different configuration and versions by twenty independent groups from both continents. Model's data have been made available along with observational data, on the ENSEMBLE system. We examine the ability of the ensemble mean and median to reduce the error and bias of SM outputs, and draw some deductions about the *size* of the ensemble and its *quality*. We quantify the level of repetition brought by individual model to the ensemble by applying a clustering analysis to examine whether the improvement of ensemble of models mean and median in reducing the error is merely due to the increased ensemble size, or if information carried by each model contributes to the MM superiority.

We remark that the aim of this study is to study the statistical properties of ensemble of models in relation to individual model realisation, for a range of cases. Each model has imperfections, and we try not to speculate as to why the bias of each individual model is generated, which is the topic of a number of detailed papers dealing with sensitivity analyses and evaluation. The ensemble represents a collective model perspective in which errors and biases result from a collective misrepresentation of the system (mcKeen et al 2005).

The paper is organised as follows. In Section 2 the air quality models and network data are presented, and the methodology applied is highlighted, with the strategy for the MM ensemble described in Section 3. Section 4 is devoted to the analysis of combinations of ensemble members that minimise the error and bias, and a clustering technique for selecting and generating an ensemble of independent members is then outlined in Section 5. Conclusions are drawn in Section 6. Finally, Appendix A briefly summarises the definition of the statistical indicators adopted.

2. Models and data

2.1 Experiment set up

In order to carry out a comprehensive evaluation of regional-scale air quality models, simulations are compared to observations for the year of 2006, and various modelling groups provided hourly concentration of ozone and other compounds. Surface concentrations were then interpolated to the monitoring locations to provide model evaluation. The focus of this study is on ozone, whilst a companion study is devoted to the particulate matter (Solazzo et al., 2012).

2.2 Participating models

Table 1 summarises the meteorological and air quality models participating to the AQMEII intercomparison exercise , providing ozone concentrations at receptor sites for the European (EU) and North American (NA). In some cases, the same model is used with a different configuration by different research groups (or in some cases the same group). In total, eleven groups for EU and seven for NA participated to analyse ozone. No a-priori screening on the worst performing model has been performed on the participating members (considered that the models have gone through, at least, operational model evaluation as defined in Dennis et al. (2010) in the past); hence, all data providers are participating in this study.

AQMEII participants were provided with a reference simulation for the year 2006, generated with the WRF v3.1 (Skamarock et al., 2008) and the MM5 (Dudhia, 1993) models, for NA and EU respectively, which were applied by the majority of groups. Several other groups conducted separate simulations with other meteorological drivers (Table 1). Skills and shortcoming of the meteorological models within AQMEII are described by Vautard et al. (2012).

The models participating to the exercise, listed below, have been extensively documented in the scientific literature (including sensitivity tests and evaluation studies):

- CMAQ (developed by the U.S. Environmental Protection Agency)
- CAMx (ENVIRON, 2010)
- CHIMERE (Schmidt et al., 2001; Bessagnet et al., 2004)
- MUSCAT (Wolke et al., 2004; Renner and Wolke, 2010)
- DEHM (Brandt et al., 2007)
- POLYPHEMUS (Mallet et al. 2007; Sartelet et al. 2012)
- EUROS (Schaap et al., 2008)
- SILAM (Sofiev et al., 2006)

- AURAMS (Gong et al., 2006; Smyth et al., 2009)
- EMEP (Simpson et al., 2003; Jeričević et al., 2010).

The combination of meteorological and transport models varies for each group (with the only exception of WRF with WRF-Chem in Europe), thus allowing analysis of diversified modelling data, a diversity which is necessary to sample the spectrum of uncertainty within an ensemble.

Emissions and chemical boundary conditions used by the various AQMEII groups are summarised in Table 1. AQMEII provided a set of emissions (referred to as “standard”) for each continent, focusing on the evaluation of the air quality and meteorological models. The EU standard emissions were prepared by TNO, which provided a gridded emission database for the year 2005 and 2006. This dataset has been widely used, and partly developed in the framework of the European MACC project http://www.gmes-atmosphere.eu/) and is the update of an earlier TNO emission database prepared for the GEMS project (http://gems.ecmwf.int). This inventory does not include biogenic emissions, thus models adopted different inventories to supply for this, as summarised in Table 1. The standard emissions for NA are based on the U.S. National Emissions Inventory 2005, Canadian national inventory 2006, Mexican BRAVO inventory 1999, biogenics from BEISv3.14, fire emissions daily estimates from HMS fire detection and SMARTFIRE system (year 2006), point sources (EGUs) from Continuous Emissions Monitoring data for the year 2006. Full details are given in Pouliot et al. (2012), where the standard inventories for EU and NA are described and compared. The standard emission inventories were used by the vast majority of the participating AQMEII groups (Table 1). Data generated with other emission inventories have also been submitted, and will provide useful comparisons in interpreting the results of model estimated ozone mixing ratios.

AQMEII also made available chemical concentrations at boundaries to modelling groups, as obtained from the GEMS re-analysis data provided by European Centre for Medium-range Weather Forecast (see Schere et al. (2012) for details). Different boundary conditions for ozone used by other AQMEII models were based on satellite measurements assimilated within the IFS system. LMDZ-INCA, which couples the general circulation model Laboratoire de Meteorologie Dynamique and the Interaction with Chemistry and Aerosol model (Hauglustaine et al., 2004) was used for CAMx and CHIMERE in one set of simulations (NA simulations), with another CHIMERE simulation using standard boundary conditions (Table 1).

2.3 Observational data for ozone

The European and North-American continental area have been divided into four subregions (EU1 to EU4 and NA1 to NA4). Figure 1 displays the subregions for both continents, along with the spatial coverage of ozone receptors. Overlaid are the contours of "standard" anthropogenic NOx emissions averaged over the summer months of June-July-August (JJA) of 2006. Only rural receptors below an altitude of 1000 m have been examined (dots in Figs. 1), with, at least, 75% data availability over the 2006. The choice of analysing only rural receptors is dictated by the need to provide comparison with spatial scales consistent with models resolution (see e.g., Vautard et al., 2009). Moreover, ozone measured by monitoring stations in urban areas is sensitive to reaction with NOx, which might reduce ozone production.

The selection of the subregions is based on emissions and altitudinal aspects, as well as practical constrains (data coverage, computational time). Extension of EU subregions are similar to those of two other AQMEII analyses of meteorological forcing (Vautard et al., 2012) and particulate matter (Solazzo et al., 2012). Subregion EU1, the British Isles, France and North Spain, was selected for its mid-latitude, mixed maritime-continental climate and large conurbations (London, Paris). Subregion EU2, Central Europe, has a continental climate with marked seasonality, many large cities, and large emissions areas. Subregion EU3, the Po Valley up to the Alps area of Italy and south-eastern France has a mixed climate, poor air quality, and is influenced by the Alpine barrier. The Southern European domain covers the Mediterranean area (south Italy, east coast of Spain and Greece), with typical Mediterranean climate and large cities (Barcelona, Rome). The number of rural receptors for EU subregions is of 201, 225, 77, and 140 and (domain 1 to 4) respectively. For NA, the number of rural receptors in each subregion is between 134 and 150. NA subregions are broadly derived from previous studies (e.g., Eder et al., 1993), also considering the NOx emission intensity, with the additional constraint of a uniform number of receptors. The eastern part of US (Domain 1), with marked high emissions along the coast of California, and milder emission towards the continent, has high solar radiation, low relative humidity, large cities with poor air quality (Los Angeles). NA2 is to the east of the Rocky Mountains and is characterised by a hot, humid climate, with large cities with poor air quality (Houston, Dallas). Subregion NA3, northeastern NA including parts of Canada, has a marked seasonal cycle, three of the North American Great Lakes, among the highest emissions areas in NA, and large cities (New York City, Philadelphia, Toronto, Montreal). Finally, the south-east US (NA4), with high emissions and strong solar radiation.

Ozone data for EU were derived from hourly data collected by the AirBase and EMEP (European Monitoring and Evaluation Programme, http://www.emep.int/) networks, for a total of 1563 stations, of which over 1400 have a percentage of data validity higher than 80%. Ozone data for NA were prepared from hourly data collected by the AIRS (Aerometric Information Retrieval Systems, http://www.epa.gov/air/data/aqsdb.html) and CASTNet (Clean Air Status and Trends Network, http://java.epa.gov/castnet/) networks in USA and the NAPS (National Air Pollution Surveillance, http://www.ec.gc.ca/rnspa-naps/) network in Canada. A total of 1445 stations are available, more than half with a percentage of data validity higher than 80%.

Furthermore, AQMEII participants also provided vertical profiles of ozone at specific airport stations in NA and EU (Galmarini et al., 2012). Measurements for vertical profiles of ozone from the MOZAIC campaign (http://mozaic.aero.obs-mip.fr/web/) were made available on ENSEMBLE system to provide model evaluation.

3. Single model and multi-model ensemble: operational evaluation and general statistics

3.1. Operational and ensemble statistics for the continental-wide domains

van Loon et al. (2007) showed that the ensemble mean ozone daily cycle over Europe, obtained by averaging over all monitoring stations for the entire year of 2001, agrees almost perfectly with the observations, and better than any member of the ensemble. This result provides substantial evidence of the superiority of the MM approach versus the SM approach. Such a result, while encouraging, poses some additional questions, as for example on the role of repeated averaging (in time and space) in smoothing out peaks and hide data variability, and also on the possibility to prove the superior MM skills for any ensemble combinations. Galmarini and Potempski (2004) show that, for the ETEX-1 case study, the MM did not show significant superior skill (and anyway poorer than an air quality case due to the instantaneous character of the prediction), and thus concluded that, in absence of a method for discriminating between members, the MM improved performance might be just coincidental and depends on the 'lucky shot' of having the right collection of models around the measured data. No methodology exists to pre-select nor to discriminate ensemble members, and therefore the result is purely dependant on opportunity.

Let us consider the plots in Fig. 2, where ozone concentrations predicted by AQMEII models for EU (Fig 2a) and NA (Fig 2b) are shown, continent-wide averaged for the full year of 2006. Whisker representation has been adopted to describe the frequency distribution, where the rectangle identifies the interquantile range (25th to 75th percentile), the square is the mean, the continuous horizontal line the median, crosses are the 1st and 99th percentiles, and the whiskers extend between minimum and maximum value. Measurements, mean and median are also shown. The top row displays the distribution of hourly values (i.e., each bar is the distribution over 8760 hourly values), the middle row is the daily average distribution (over 365 values), and the bottom row is the mean diurnal range (each bar reflects the distribution over 24 values), in which same hours of each day of the year are averaged. Depending on the averaging period, ozone concentrations are reduced by a factor of two, for both continents. This translates into a dramatic reduction of the spread (min and max values are within the inter-quantile ranges for the diurnal cycle) and a clustering of the diurnal time series which results in enhanced statistical agreement. (To maintain anonymity, each participating model has been assigned a random model number, Mod 1 to 11 for EU, and Mod 12 to 18 for NA, which do not correspond to the order of models in Table 1). Thus, averaging over extended areas (continent) and periods (year) has a dramatic effect in reducing the spread of the data.

The ability of the ensemble to sample measurements uncertainty in both continents is analysed by means of the rank histograms in Figs. 3, which are a measure of the ensemble reliability (Talagrand et al., 1998; Joliffe and Stephenson, 2003). The rank histogram is a widely adopted diagnostic tool to evaluate the spread an ensemble of members. In a rank histogram, the population of the *k*-th rank is the fraction of time observations falls between the sorted member *k*-1 and *k*. Ideally, the frequency for each bin should be the same, meaning that the ozone estimate from each ensemble member is as probable as any other member, and observations have an equal probability to belong to any bin (Hamill, 2000). In such case, in fact, the observation and the ensemble members are derived from the same probability distribution, and the probability of the observation falling into a particular bin is the same for all bins. In Fig. 3, unbiased hourly ozone data from the full year, continents-wide averaged are used. For EU (Fig 3a), bins population is rather uniform for the first ten bins (between 6 and 11%), with bins 11 and 12 having a frequency of ~18% each, indicating a difficulty of the ensemble mean to simulate high hourly concentration, which is a bias of the ensemble mean (underestimation). The ranked histogram for NA (Fig 3b) shows the intermediate bins more populated than the side bins, indicating the possible presence of outlying members.

It is not straightforward to assess whether the deviation from flatness in both cases (EU and NA) is due to chance (ensemble members and observations derived from the same distribution) or if there is a compensation effect over such large domains and long time scale. These aspects will be further detailed in Section 3.4.

3.2 Subregional SM and MM ensemble analyses

Regional air quality models are often used on limited spatial and temporal scales (months or season over a few hundreds of kilometres wide, Boynard et al., 2011; Hogrefe et al., 2011; Bloomer et al., 2009; Camalier et al., 2007), for which mutual cancelation of models' errors might not be as effective as in the case of continental and yearly scales, as discussed for the results of Fig. 2. The analyses presented in this study focus on the ozone spatial variability of ozone concentration statistics in four distinct subregions of the continental domains of Fig. 1, examining the temporal variability for the summer months JJA, i.e., when the ozone mixing ratios are at maximum and are of major concern for the public health. Analysis and evaluation of single models performance over the whole year are presented elsewhere.

Subregional ozone diurnal cycles are shown in Fig. 4a (EU) and Fig. 4b (NA), including ensemble mean and median (hourly data have been used for the analysis). Examining the observational data trends, there is an ample intra-continental variability of the ozone maxima, with the North Italian and Mediterranean regions (EU3 and EU4) reaching 60 ppb and over, and peaks of ~45-50 ppb occur in the other EU subregions. For EU1,2,3 the maximum occurs at LT 17, while it is detected two hours earlier in EU4, due to the higher insulation. Minimum values, between LT 7 and 8, range between 20 and 30 ppb, with the Mediterranean area having the highest minimum due to the relative abundance of biogenic emissions (see e.g., Sartelet et al., 2012). Maximum values for NA are of the same general magnitude as EU, between 45 and 55 ppb for subregions 1, 2 and 4 and only reaching ~35 ppb for NA3 in the North-East region, and occur at LT 16. Minimum values, at LT 07, range between 20 to 25 ppb for NA1,2,3 and less than 20 ppb for NA4. This latter (South-East region) exhibits a steep rise of ozone mixing ratios that is indicative of strong daytime photochemical activity in this region.

The majority of models (thin lines in Figs. 4) exhibit highly regional-dependant behaviours, although some common patterns can be detected. The Muscat model is severely biased high at all regions (probably due to …), whilst the majority of the other models have a tendency to underestimate (even significantly) the peak concentration (and the time of the peak), as well as to under predict the night hour mixing level (especially for EU2, continental EU), probably due to high daily temperature gradient in this region. The CMAQ model, as emerge from previous works (Herwehe et al., 2011; Smyth et al., 2009), has the tendency of overestimating ozone concentration at night, due to difficulties in dealing with stable conditions. Another source of performance variability for EU is identified in the different biogenic emissions adopted by each model. This hypothesis is confirmed by the performance of the Chimere model with MEGAN biogenic emission (and standard EMEP anthropogenic), which is the best scoring model for all EU subregions. The DEHM model has the highest bias (low) for EU4 and especially during night time. Brandt et al. (2012) suggest DEHM underestimation of summer months ozone in Europe to be due to isoprene emission being too low. Sartelet et al. (2012), in evaluating the impact of biogenic emission in the context of AQMEII also point out the impact of biogenic emission for estimating ozone. In particular they found that the best model performance for ozone is obtained when using the MEGAN biogenic inventory.

Model's results for NA regions exhibit a lower spread throughout the cycle (Fig. 4b), with the exception of a clear outlying model for NA1,2,3, which is consistently biased low, especially at night. However, the night over prediction is, to various degrees, a common feature to the majority of models, indicating the difficulties in dealing with stable conditions, despite the variety of vertical mixing schemes implemented by the models. The case of the south-east area (NA4), with consistent model overestimation, indicates that…

Ensemble mean and median generally underestimate the amplitude of the diurnal ozone cycle in Europe, despite one largely biased high outlying model. By contrast, mean and median accurately mimic the ozone amplitude for NA1,2,3 (whilst largely overestimate for NA4) thanks to the mutual compensation of the biased low outlier and the general tendency of the other member to over predict. It should be observed that the mean and the median curve overlapping is a consequence of the repeated data averaging (spatially and temporally) that has smoothed out the peaks of the distribution, as previously shown in Fig 2.

Figures 5 quantify the error statistics for EU (Fig 5a) and NA (Fig 5b), in form of a "soccer plot" (Appel et al., 2011). The NMSE vs NMB (see Appendix A for definition) is reported for each individual model, together with the ensemble mean and median, for the four subregions (numbered 1 to 4). Models data falling within the dotted lines indicate compliance with the performance criteria set by Russel and Dennis (2000) for ozone (bias within ± 15% and error within ± 30%). The majority of points lie in the left region of the soccer area, indicating undeprediction for EU, with the exception of Mod1, substantially overestimating the mixing ratio for all subregions. Model results for the NA are well within the 15% box (mainly overestimation), with the exception of NA4, where three models show overestimation between 15 and 20%. Points are mainly grouped by model for EU and by region for NA, suggesting a stronger dependence on local conditions for the latter continent. The ensemble mean and median (approximately identical for NA, as already noted for the diurnal cycle) are within the 15% area, indicating compliance of performance criteria for both continents.

3.3 Sources of model bias

Vautard et al. (2012) have shown that there is a clear model overestimation of the 10 m wind speed across the European continent (up to a factor of two), especially in the EU2 (continental Europe). Solazzo et al. (2012) have also shown that, in the case of particulate matter, there is a significant correlation between the bias of wind speed and that of particulate concentration (also underestimated), meaning that the poor model prediction for wind speed are partially responsible for the bias of primary pollutants. Although ozone is not emitted at ground, we can infer that the positive bias for wind has a diluting effect on the primary species (NOx, VOC, etc), help explaining – partially – the underestimation of ozone across Europe.

From investigating Fig. 5, there seems to exist, especially for NA, a direct correlation between bias and error, suggesting that most of the uncertainty is due to biases. Schere et al. (2012); Appel et al., (2012); Nopmongcol et al. (2012) have shown that such bias is likely introduced by the standard GEMS BC for ozone, which are overestimated, especially in winter. NA-Mod13 uses the same chemistry transport model as NA-Mod17, but uses boundary conditions from a global model, which provided ozone mixing ratios along the boundary that are significantly lower than the standard boundary data, a difference which propagates across the entire NA domain (Schere et al., 2012). During summer months, with O3 largely produced by atmospheric chemistry, the influence of the boundary conditions is reduced, but not disappeared. To quantify the bias due to boundary conditions for ozone over the NA continent, in Fig. 6 models predictions of vertical ozone concentration are compared against measurements collected along flight paths during MOZAIC. The focus is put on the west coast of NA, where the oceanic air masses moved by the westerly winds first encounter the land. Concentration of pollutants aloft in this coastal region should also be less influenced by the ground emissions and they should then be comparable with the boundary conditions for ozone. Each MOZAIC flight gives a set of measurements along a trajectory, from which the closets values to the model predictions are used for comparison. Both, MOZAIC measurements and model predictions are then averaged in time to obtain an average value at each model vertical level. Sixteen flights made over Portland airport for the month of August 2006 have been used (lon/lat: -122.75W/45.5N). For comparison, we use outcomes from three models, adopting GEMS boundary conditions. We are not interested in the inter-model comparison here, but rather in the distance between the MOZAIC measurements and the BCs, and to what extent this translates into bias. The black thick line is the observation (with the standard deviation), the grey thick line is the GEMS BCs at the node 120.833W/47.00E, and the thin grey lines are the models. For completeness, we have reported the ground level concentration of ozone over six ground base regional receptors of the AIRS network within the MOZAIC domain defined for Portland (the projection to the ground of all flight trajectories up to 8500 m, see Galmarini et al. (2012) for details). We notice that the independent ground level and the MOZAIC measurements are in perfect agreement, supporting the validity of the observations. In the boundary layer, below 1000 m, there is a difference between GEMS and MOZAIC of ~ 13 ppb, leading to a positive model bias (models to receptors). The bias turns into negative above that height, and thus we might expect a model underestimation above the boundary layer, but this is not investigated here.

3.4. Reliability of the Multi-model Ensemble

Consistently biased rank histograms, displayed for all subregions of the two continents, exhibit a sloped shape, as in Figs 7a (EU) and 7b (NA). Analysis is based on hourly data for the period JJA. EU subregions 1,2, have histograms significantly deviating from flatness, with most populated bins towards the end of the ranks (model underprediction). EU4 shows empty initial and final bins, indicating an excess of variability. The histogram for the entire EU domain, the biased trends for EU1,2 are compensated by those of EU3 and 4, resulting in a flat rank histogram, as observed for the entire year (Fig. 3a). As discussed at the beginning of section 3.2, by looking at long time and large spatial scales, the seasonal and intra-continental variability is hidden by the averaging and mutual compensation. Strong biases are also observed for the NA subregions (Fig 7b), with over-prediction (left bins most populated) for all sub regions, as seen in Fig. 5b. The spread also suffers from deficiencies of the ensemble in all cases, with excess of spread for NA1 (middle ranks more frequent) or insufficient spread, such as in NA2,3,4 (side-bins more populated). This latter case is typically due to not having captured all sources of error properly (Vautard et al., 2009). This might be due to many members of the ensemble using the same meteorological drivers and/or emissions. Comparing the histograms of Fig. 7b for the entire NA domain for JJA and that of Fig. 3b for the entire NA domain for the full year highlights that for the full year the bins were more uniform, with a tendency to a “bell” shape, whereas for JJA the distribution is drastically biased and bin populations uneven. This is due to large underestimation in the winter months by models adopting the GEMS boundary conditions for ozone (Appel et al., 2012) which compensates for the overestimation in the summer.

4. Multimodel analysis: selected vs. unselected model ensembles

4.1 Ensemble size

In this section we evaluate whether an ensemble built with a subset of models can outperform the ensemble mean of all available members, as anticipated by the theoretical analysis of Potempski and Galmarini (2009). The analysis is done for the subregions of EU and NA separately, using hourly ozone data for the period JJA.

Let us consider the distribution of some statistical skills (RMSE, PCC, MB, MGE, defined in the Appendix A) *of the mean of all* possible combinations of available members, *n* (*n* is 11 for EU and 7 for NA). The number of combinations of any *k* members is $\begin{pmatrix} n \\ k \end{pmatrix}_{k=2,\ldots,n-1}$ .For example there are as many as 462 combinations of 5 models for EU, and 35 combinations of 3 models for NA. Results are presented in Figs 8a (EU) and 8b (NA). The first column is the RMSE, the middle column is the MB, and the last is PCC, as function of the number of members of the MM ensemble.

The continuous lines on each plot are the mean and median of the distribution of any *k*-model combinations. Mean and median have similar behaviour decaying as $O(1/k)$, (Potempski and Galmarini, 2009). These curves move toward more skilful model combinations as the number of members *k* increases, which confirms the common practice to average over all available members to obtain enhanced performance with respect to the mean of the available members. For MB, the mean trend is flat, because of the quasi-symmetric error fluctuations about the mean value. Mean RMSE curves decrease steeply from two to four models (all subregions, except NA4). A further striking feature is that the best SM has similar (EU1,2, NA1,3) or lower (EU3,4, NA2,4) RMSE than the ensemble mean with all members. This is most probably due to having included members with large variances in the ensemble (Potempski and Galmarini, 2009).

Analysis of mean RMSE for EU subregions (Fig 8a), for which a large set of members is available, shows a plateau for $k > 5$. This would indicate that there is no advantage, on average, to combine more than six members, as the benefit in minimizing the RMSE is negligible. Investigating the maximum of RMSE, it results $max_k\ RMSE\ (k) > max_k\ RMSE\ (k+1)$. The mean of ensembles with a large number of members have the properties of reducing the maximum error. For example, RMSE of EU3 has a large error span, between 2.5 and 15 ppb for $k = 2$, which reduces between 4 and 7 ppb for k = 10 (Fig 8a). A similar trend is replicated for PCC (all subregions), with a monotonic improvement of minimum PCC values when increasing k.

Values of minimum RMSE (lower bound) exhibit a more complex behaviour. A minimum, among all combinations, systematically emerges for ensembles with number of members $k < n$. Similarly, a maximum of PCC is achieved by combinations of a subset of members. This result tells that ensembles of a few members outperform –systematically– the whole ensemble; further to that, adding new members to such an optimal ensemble (thus moving towards higher value of k in Figs. 8) deteriorates the quality of the ensemble, as the error increases.

4.2 Ensemble combinations of minimum RMSE and MB

In Table 2, the MM combination of minimum RMSE is reported for any k, where models are identified by the RMSE-ranking (for example, 2-5 is the ensemble mean of the second and fifth best SM in terms of RMSE). The SM RMSE ranking is defined by domain. Models may not have the same SM RMSE ranking over the different subregions. In bold face are the global minimum, i.e. the minimum among all possible combinations (a five and a six-model combination gave the same minimum RMSE for EU4).

The main aspect worth noticing is that the RMSE-ranking shows that the optimum is, in some cases, achieved by MM ensemble containing low ranking members. This indicates that all members are needed to build a skilful ensemble. This means that an ensemble of top ranking model results can be worse than an ensemble of top ranking and low ranking ones. Also outliers need to be included in the ensemble to obtain the best performance.

It can be argued that large ensembles are needed to capture extreme events with high concentrations. Figure 9 displays a scatter plot of the 1-hour daily maximum ozone concentration for the EU subregions (analysis for NA with fewer members produced similar results and is not reported). On the x-axes is the 1-hour maximum of the ensemble of all available members, and the y-axes the 1-hour maximum of the ensemble of the selected members with minimum RMSE (bold-face combinations of Table 2). Data distribution along the diagonal line for each region shows that ensembles of selected models and full ensembles have the same probability to capture the extreme concentrations. In particular for EU1 and EU3, maximum predicted are higher with the small ensemble. This is because poor performing SM added to an ensemble can improve RMSE and even lead to optimal RMSE. This is due to biases compensating one another.

As an example, let us consider the case of Fig. 10, in which ozone concentrations of: observations (Fig 10a), the ensemble of ranked models 1 and 5, (Fig 10b), ranked model 2 (Fig 10c), and ranked model 11 (Fig 10d), are displayed at receptor positions. With reference to Table 2, the ranked combination 1-5-11 is that of minimum RMSE for EU1. One could wonder as to why the least ranked model (11) gives a better contribution to the ensemble that a high ranking one, for example the second best. Looking at the British Isles and France receptors (Domain 1 of Fig 1a), the MM mean of Fig 10b clearly underestimates the observations in the south of France. When the 11th ranked model (Fig. 10d) is added to the ensemble of Fig 10b, it allows error compensation and RMSE to be lower than the combination with the 2nd best ranked model of Fig. 10c. This is because, the 2nd best model has a performance very similar to the best one (the 1st already included in the ensemble), and thus brings no new information to the existing ensemble, while the 11th model, whilts scoring bad at all domain, matched the high concentration of southern France, i.e. the only place where the best models are failing. RMSE penalises large errors, and model ranked 11 is the model that, on a spatially averaged sense, allows the ensemble to match the concentration at a larger number of stations and larger number of hour with minimum error.

Statistical results and whisked plots for the full and for the selected members ensemble are reported in Table 3 and Fig. 11, for each subregion. The first row of Table 3 is the full members ensemble mean, the second row is the MM of minimum RMSE (bold face combinations of Table 2). RMSE is, of course, lower for the latter at all subregions. PCC varies only slightly, indicating that the association between observation and ensemble is not strictly related to model's error. The minimum RMSE combinations also improve the estimation of the spread (the standard deviation of the MM ensemble, σ), in almost all regions (Table 3) and especially for NA regions. Thus, having reduced the number of the members does not degrade the ensemble variability, which actually becomes closer to the spread of the observations. This is most probably due to the reduced variability induced by members sharing similar emissions and boundary conditions. Graphical representation in Fig. 11 also shows the how the selected member compares against the full member ensemble in terms of spread, maximum and minimum, and percentile distribution. The improvement of the selected ensemble to model the spread and the high concentrations are most visible, particularly for EU regions.

5. Reduce data complexity: a clustering approach

Results of previous section 4 have shown that a skilful ensemble is built with an optimal number of members between two (NA) and five (EU), also including low ranking skill score members. Is there any methodology that allows discerning among members? To this scope, we adopted here a methodology for clustering highly associated models and discard redundant information, using the PCC as metric (we note that PCC is independent to model bias, therefore the analysis would be the same for unbiased models). The most representative models of each cluster, chosen based on a distance metric, are then used to generate a reduced ensemble. In this way, the information each member brings to the ensemble is "unique" to the maximum possible degree.

The metric used to calculate the distance between the PCC of any two models, and between clusters, is the squared Euclidean distance. First the points which are furthest apart are identified, and used as the initial cluster centres. Then, the other cases are allocated to the closest centre by Euclidean distance from each centre. Results for this procedure are displayed in Figure 12 (EU) and Fig 13 (NA). The height of each inverted U-shaped line represents the distance between the two clusters being connected. Independent clusters are identified by different colours. Sensitivity analysis to other distance metrics (not shown), have shown that the clustering of models is independent on the metric adopted for the distance (leaving the group associations unaltered). The distance itself, however, changes, but this does not affect the results of this study. On the vertical axis of Figs 12 and 13, models are identified by their number and by their RMSE-ranking (discussed in Section 4.2). The ranking information allows to track the models positioning and analysing whether aggregation results from difference in the models themselves (transport model, meteorological drivers, emission, etc.), or if the models' performance (RMSE) have an influence.

For EU (Fig 12), the maximum PCC distance (degree of model's disassociation) varies between 0.12 (EU4) to 0.28 (EU2). By contrast, analysis of NA regions (Fig. 13) shows the maximum distance is of 0.08 for all regions, except for NA2 (~0.03). Association between models is thus stronger for NA, indicating a lower degree of independence. This can be explained by considering that four out of seven models used the same meteorological driver, and six models shared the same emissions.

Moreover, for EU it is possible to isolate two repeating groups of models whose PCC distance is very small: Mod6 and Mod7, Mod11 and Mod3. Models of the former group are essentially the same, as they share the transport and meteorological model (WRF-Chem), as well as emission and boundary conditions. They also have similar RMSE-ranking. Mod11 and Mod3 differ in the air quality model but share the MM5 input and the standard anthropogenic emissions. The NA analysis, with fewer members, shows repeated association of groups of two models: Mod15 and16 (same meteorological driver WRF, anthropogenic emission and boundary conditions), Mod13 and Mod17 (share the meteorological and air quality models), and Mod14 and Mod18 (same meteorological driver). Mod12 is associated with Mod14 and Mod18, with the exception of NA3.

In order to find an optimal set of clusters, we can define a threshold at which models are said to be independent (imagine cutting vertically the dendrograms). The selection of the cutting height is partially arbitrary. The common practice suggests cutting the dendrogram at the height where the distance from next clustered groups is relatively large, and the retained number of clusters is small compared to the original number of models (Riccio et al., 2012). Members of ensemble generated with higher threshold are more distant, thus more independent. Cluster representative and ensembles are summarised in Table 4, for both continents and for different PCC distance. For clusters composed by two members only, and for cluster with a symmetric structure (same mutual distance among all members such as the third cluster of EU2, Fig 12b), it was no possible to identify a model having minimum distance or whose distance from the centre of the cluster was minimum in terms of RMSE. In these cases more than one model, in turn, is selected to represent the cluster.

The number of independent members varies between 3 and 6 for EU, and between 2 and 4 for NA (this difference is probably due to the smaller number of models for NA). It is remarkable that the number of independent clusters matches the number of models to generate MM ensemble of minimum RMSE of Figs. 8, for both continents. The two investigations are in fact independent, as the clustering analysis makes no use of observational data. Looking at the bold-face combinations of Table 2 (minimum RMSE combination), it can be deduced that the ensembles of minimum RMSE have two or more members belonging to the same cluster, and for NA4 all members from the same cluster. This result is driven by not having many independent members, due to models sharing of boundary conditions, meteorology, and emissions.

We have also compared the RMSE of MM ensembles of Table 4 with the curves of RMSE discussed in Fig. 8, and reported the results in Fig 14a (EU) and Fig 14b (NA). Imagine connecting, for any number of models, the minimum and the maximum RMSE of Figs. 8. The curves obtained are those of Figs. 14 (thick lines: minimum; dotted lines: maximum). RMSE of combinations of Table 4 (obtained with the clustering technique) are the short lines in Figs. 14, reported along with the ranked combination. In the cases of clusters with two members only (symmetric clusters), it was not possible to identify the cluster representative, and thus two members have been retained for the analysis. What is interesting to compare in Figs. 14 is the position of the cluster's combination against the RMSE of the full member ensemble, to infer whether the new methodology is able to produce reduced ensembles more skilful than the full ensemble mean. In fact, we can notice that independent model combinations have, in most cases, lower RMSE than the full ensemble, and that for all subregions there are ensembles that clearly outperform it. For example, the combinations 1-2-3-8-11, 2-6-7-8, 1-3-6-9-11, 1-2-4-8-9-11 for EU1,2,3, and 4, respectively, have lower RMSE than the mean of all ensemble members and close to the minimum curve. There are, on the other hand, situations in which the ambiguous definition of cluster representatives leads to high-RMSE MM combinations, as for the four-member combination of EU4 (1-2-4/5-11) and NA1 (2-4-5). Further work needs to be devoted to remove such ambiguity.

6. Conclusions

This study collectively evaluates and analyses the performance of over ten regional air quality models and their ensembles in the context of the AQMEII intercomparison exercise. The scale of the exercise is unprecedented, with two continent-wide domains being modelled, for a full year. Focus of this paper is on the *collective* analysis of surface ozone concentration, rather than on inter-comparing metrics for each individual model. We start with analysing time series for ozone in subregions of the continental Europe and North America, and proceed with the interpreting uncertainties for individual model and ensemble. Analysis of model bias and error in each subregion demonstrate that most of the model's error is introduced by bias, from emissions, boundary conditions and meteorological drivers.

We then show that there are ensemble combinations with a reduced number of members whose mean produces an error smaller then the full member ensemble mean. Thus, we have shown that a skilful ensemble is not necessarily generated by all available members, but rather by selecting members that can contribute to minimise the error.

We find that an ensemble of top ranking model results can be worse than an ensemble of top ranking and low ranking ones. Until now it was assumed that, as long as a large set of results were treated statistically in one ensemble, then the ensemble would have been better than any individual member. Furthermore, it was supposed that the better the results the better the ensemble. What we demonstrate here is that such hypothesis is not necessarily true, as also outliers need to be included in the ensemble to enhance performance. Further to that, we have shown that the score does not improve with just increasing the number of models in the ensemble. By contrast, the level of dependence of model results may lead to a deterioration of the results and to an overall worsening of performance. Despite the remarkable progress of ensemble air quality modelling over the past decade and the effort spent to build a theoretical foundation, there still are many outstanding questions regarding this technique. Among them, what is the best and most beneficial way to build an ensemble of members? How to determine the optimum size of the ensemble in order to capture data variability as well as keeping the error low?

To try addressing these questions, we adopt a methodology for reducing data complexity, known as *clustering technique*, which has the purpose of simplifying information provided by the large amounts of data (such as air quality model outputs) by classifying, or clustering, the data into groups based on a select metric, where there is no prior knowledge of grouping. Results show that, whilst this methodology needs further refinements, by opportunely selecting the cluster distance and association criteria, we can generate ensemble of selected members with errors significantly lower than the full members ensemble mean. The results of the clustering analysis have a general character and impact, and are of direct relevance for applications not only strictly related to ensemble model evaluation, but also to other ensemble communities, for example air quality forecasting, climate, oceanography, and others.

Aknowlegments

The authors acknowledge for the strong support of the European Commission, Airbus, and the Airlines (Lufthansa, Austrian, Air France) who carry free of charge the MOZAIC equipment and perform the maintenance since 1994. MOZAIC is presently funded by INSUCNRS (France), Meteo-France, and Forschungszentrum (FZJ, Julich, Germany). The MOZAIC data based is supported by ETHER (CNES and INSU-CNRS). It is also acknowledged the Center for Energy, Environment and Health (CEEH), financed by The Danish Strategic Research Program on Sustainable Energy under contract no 2104-06-0027. Homepage: www.ceeh.dk.

Appendix A: Statistical Measures

Defining y the vector of model output and obs the vector of observations (n-component both), having mean value $\overline{y}$ and $\overline{obs}$, respectively.

Mean bias:

$$MB = \frac{\sum_i (y_i - obs_i)}{n} \tag{A1}$$

Root mean square error:

$$RMSE = \sqrt{\frac{\sum_i (y_i - obs_i)^2}{n}} \qquad \text{(A2)}$$

Mean Gross Error:

$$MGE = \frac{\sum_i | y_i - obs_i |}{n} \qquad \text{(A3)}$$

Normalised mean square error

$$NMSE = \frac{\sum_i (y_i - obs_i)^2}{n\, \overline{y}\, \overline{obs}} \qquad \text{(A4)}$$

Fractional Bias

$$FB = 2\frac{\overline{obs} - \overline{y}}{\overline{obs} + \overline{y}} \qquad \text{(A5)}$$

Normalised Mean Bias:

$$NMB = \frac{\sum_i (y_i - obs_i)}{n\, \overline{y}\, \overline{obs}} \qquad \text{(A6)}$$

Pearson correlation coefficient:

$$PCC = \frac{\sum_i (y_i - \overline{y})(obs_i - \overline{obs})}{\sum_i (y_i - \overline{y})^2 \sum_i (obs_i - \overline{obs})^2} \qquad \text{(A7)}$$

References

Appel, K.W., Chemel, C., and et al, 2012. Examination of the Community Multiscale Air Quality (CMAQ) model performance for North America and Europe for the AQMEII project. Atmospheric Environment 10.1016/j.atmosenv.2011.11.016

Appel, K.W., Gilliam, R.C., Davis, N., Zubrov, A., Howard, S.C., 2011. Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models. Environmental Modelling & Software 26, 434-443.

Beaver, S., Palazoglu, A., 2006. A cluster aggregation scheme for ozone episode selection in the San Francisco, CA Bay Area. Atmospheric Environment 40, 713-725

Bessagnet, B., Hodzic, A., Vautard, R., Beekmann, M., Cheinet, S., Honoré, C., Liousse, C., Rouil, L., 2004. Aerosol modeling with CHIMERE: preliminary evaluation at the continental scale. Atmospheric Environment 38, 2803-2817.

Bianconi, R., Galmarini, S., Bellasio, R., 2004. Web-based system for decision support in case of emergency: ensemble modelling of long-range atmospheric dispersion of radionuclides. Environmental Modelling and Software 19, 401-411.

Bloomer, B.J., Stehr, J.W., Piety, C.A., Salawitch, R.J., Dickerson,R.R., 2009. Observed relationships of ozone air pollution with temperature and emission. Geophysical Research letter 36, L09803.

Boynard, A., Beekman, M., Foret., G., Ung, A., Szopa, S., Schmechtig, C., Coman, A., 2011. An ensemble of regional ozone model uncertainty with an explicit error representation. Atmospheric Environment 45, 784-793.

Brandt, J., Silver, J. D., and et al., 2012. An integrated model study for Europe and North America using the Danish Eulerian Hemispheric Model with focus on intercontinental transport of air pollution. Atmospheric Environment

Brandt, J., J. H. Christensen, L. M. Frohn, C. Geels, K. M. Hansen, G. B. Hedegaard, M. Hvidberg and C. A. Skjøth, 2007. THOR – an operational and integrated model system for air pollution forecasting and management from regional to local scale. Proceedings of the 2nd ACCENT Symposium, Urbino (Italy), July 23-27, 2007

Camalier, L., Cox,W., Dolwick, P., 2007. The effects of meteorology on ozone in urban areas and their use in assessing ozone trends. Atmospheric Environment 41, 7127-7137.

Carmichael, G. R., Ferm, M., Thongboonchoo, N., Woo, J.-H., Chan, L. Y., Murano, K., Viet, P.H., Mossberg, C., Bala, R., Boonjawat, J., Upatum, P., Mohan, M., Adhikary, S. P., Shrestha, A. B., Pienaar, J. J., Brunke, E. B., Chen, T., Jie, T., Guoan, D., Peng, L. C., Dhiharto, S., Harjanto, H., Jose, A. M., Kimani, W., Kirouane, A., Lacaux, J., Richard, S., Barturen, O., Cerda, J. C., Athayde, A., Tavares, T., Cotrina, J. S., and Bilici, E., 2003. Measurements of sulfur dioxide, ozone and ammonia concentrations in Asia, Africa, and South America using passive samplers, Atmos. Environ., 37, 1293–1308.

Ciaramella, A., Giunta, G., Riccio, A., Galamrini, S., 2009. Independent model selection for ensemble dispersion forecasting, Applications of Supervised and Unsupervised Ensemble Methods, Oleg Okun and Giorgio Valentini (eds), Springer, pp 213-231

Delle Monache, L., Stull, R., 2003. An ensemble air quality forecast over western Europe during an ozone episode. Atmospheric Environment 37, 3469-3474.

Dennis et al., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems. Environ Fluid Mech DOI 10.1007/s10652-009-9163-2

Dudhia, J. 1993 A nonhydrostatic version of the PennState/NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. Monthly Weather Review 121, 1493-1513.

Eder, B.K., Davis, J.M., Bloomfield, P., 1993. A characterization of the spatiotemporal variability of non-urban ozone concentrations over the eastern United States. Atmospheric Environment 27, 2645-2668.

ENVIRON, 2010. User's guide to the Comprehensive Air Quality model with extensions (CAMx) version 5.20 (March, 2010), http://www.camx.com.

Eskes, H., 2003. Stratospheric ozone: Satellite Observations, Data Assimilation and Forecasts, Proceedings of the Seminar on Recent Developments in Data Assimilation for Atmosphere and Ocean, 8-12 September 2003, ECMWF, Reading, UK, pp. 341-360

Galmarini et al., 2001

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., and et al., 2004. Ensemble dispersion forecasting. Part I: concept, approach and indicators. Atmospheric Environment 38, 4607-4617.

Galmarini, S., Bianconi, R., Appel, W., Solazzo, E., and et al., 2012. ENSEMBLE and AMET: two systems and approaches to a harmonised, simplified and efficient assistance to air quality model developments and evaluation. Atmospheric Environment doi:10.1016/j.atmosenv.2011.08.076

Gong, W., A.P. Dastoor, V.S. Bouchet, S. Gong, P.A. Makar, M.D. Moran, B. Pabla, S. Ménard, L-P. Crevier, S. Cousineau, and S. Venkatesh, 2006. Cloud processing of gases and aerosols in a regional air quality model (AURAMS). Atmos. Res. 82, 248-275.

Guenther, A., Zimmerman, P., Wildermuth, M., 1994. Natural volatile organic compound emission rate estimates for US woodland landscapes. Atmos. Environ., 28, 1197–1210.

Hamill, T.H., 2000. Interpretation of Rank Histograms for verifying ensemble forecasts. Monthly Weather Review 129, 550-560

Hauglustaine, D.A., Hourdin, F., Walters, S., Jourdain, L., Filiberti, M.-A., Larmarque,, J.-F., Holland, E. A., 2004. Interactive chemistry in the Laboratoire de Météorologie Dynamique general circulation model: description and background tropospheric chemistry evaluation. J. Geophys. Res., 109, D04314, doi:10.1029/3JD003957.

Henne, S., Brunner, D., Folini, D., Solberg, S., Klausen, J., Buchmann, B., 2010. Assessment of parameters describing representativeness of air quality in-situ measurement sites. Atmos. Chem. Phys. 10, 3561–3581.

Herwehe, J.A., Otte, T.L., Mathur, R., Rao, S.T., 2011. Diagnostic analysis of ozone concentrations simulated by two regional-scale air quality models. Atmospheric Environment 45, 5957-5969.

Hogrefe, C., Hao, W., Zalewsky, E.E., Ku, J.-Y, Lynn, B., Rosenzweig, C., Schultz, M.G., Rast, S., Newchurch, M.J., Wang, L., Kinney, P.L., Sistla, G., 2011. An analysis of Long-term regional scale ozone simulations over the north-eastern Unites States: Variability and trends. Atmos. Chem. Phys., 11, 567–582.

Holloway, T., Fiore, A., Hastings, M.G., 2003. Intercontinental transport of air pollution: will emerging science lead to a new hemispheric treaty? *Env. Sc. Tech.*, 37, 4535-4542.

Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. Atmospheric Environment 43, 51-63.

Jeričević, A., Kraljević, L. Grisogono, B., Fagerli, H., and Večenaj, Ž., 2010. Parameterization of vertical diffusion and the atmospheric boundary layer height determination in the EMEP model. Atmos. Chem. Phys., 10, 341-364. DOI:10.5194/acp-10-341-2010

Jollife, I. T., and D. B. Stephenson (Eds.) (2003), Forecast Verification: A Practitioner's Guide in Atmospheric Science, 240 pp., John Wiley, Hoboken, N. J.

Kumar, A., Barnston, A.G., Hoerling, M.P., 2001. Seasonal prediction, probability verifications, and ensemble size. J. Climate 14, 1671-1676.

Mallet V., Quélo D., Sportisse B., Ahmed de Biasi M., Debry E., Korsakissok I., Wu L., Roustan Y., Sartelet K., Tombette M., and Foudhil H., 2007. Technical Note: The air quality modeling system Polyphemus. Atmos. Chem. Phys., 7, 5479-5487.

Mallet, V., Sportisse, B., 2006. Ensemble-based air quality forecasts: A multimodel approach applied to ozone. Journal of Geophysical Research 111, D18302.

McKeen, S., Grell, G., Peckham, S., Wilczak, J., Djalalova, I., Hsie, E.-Y., Frost, G., Peischl, J., Schwarz, J., Spackman, R., Holloway, J., De Gouw, J., Warneke, C., Gong, W., Bouchet, V., Gaudreault, S., Racine, J., McHenry, J., McQueen, J., Lee, P., Tang, Y., Carmichael, G.R., Mathur, R., 2009. An evaluation of real-time air quality forecasts and their urban emissions over eastern Texas during the summer of 2006 Second Texas Air Quality Study field study. Journal of Geophysical Research 114, D00F11, doi:10.1029/2008JD011697.

McKeen, S., Wilczak, J., Grell, G., Djalalova, I.,Peckham, S., Hsie, E.-Y., Gong, W., Bouchet, V., Menard, S., Moffett, R., McHenry, J., McQueen, J., Tang, Y., Carmichael, G., Pagowski, M., Chan, A.C., Dye, T.S., Frost G., Lee, P., Mathur, R., 2005. Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. J. Geophys. Res., 110, D21307.

McKeen, S.A., Chung, S.H., Wilczak, J., Grell, G., Djalalova, I., Peckham, S., Gong, W., Bouchet, V., Moffet, R., Tang, Y., Carmichael, G.R., Mathur, R., Yu, S., 2007. Evaluation of several PM2.5 forecast models using data collected during the ICARTT/NEAQS 2004 field study. Journal of Geophysical Research 112, D10S20, doi:10.1029/2006JD007608.

Nopmongcol, U., Koo, B., Tai, E., Jung, J., Piyachaturawat, P. Modeling Europe with CAMx for the Air Quality Model Evaluation International Initiative (AQMEII). Atmospheric Environment 10.1016/j.atmosenv.2011.11.023

Potempski, S., Galmarini, S., 2009. Est Modus in Rebus: analytical properties of multi-model ensembles. Atmospheric Chemistry and Physics 9, 9471-9489.

Pouliot, G., Pierce, T, Denier van der Gon, H. ,., Schaap, M., Moran, M., Nopmongcol, U., Comparing Emissions Inventories and Model-Ready Emissions Datasets between Europe and North America for the AQMEII Project. *Submitted for publication to Atmospheric Environment (this issue)*

Rao, S.T., Galmarini, S., Puckett, K., 2011. Air quality model evaluation international initiative (AQMEII). Bulletin of the American Meteorological Society 92, 23-30. DOI:10.1175/2010BAMS3069.1

Renner, E., Wolke, R., 2010. Modelling the formation and atmospheric transport of secondary inorganic aerosols with special attention to regions with high ammonia emissions. Atmos. Environ. 41, 1904–1912.

Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., Potempski, S., 2012. On the systematic reduction of data complexity in multi-model ensemble atmospheric dispersion modelling. Journal Geophysical Research, *in press.*

Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modelling. Atmospheric Environment 34, 2283 – 2324.

Sartelet K., Couvidat F., Seigneur C., Roustan Y., 2012. Impact of biogenic emissions on air quality over Europe and North America. Atmospheric Environment doi:10.1016/j.atmosenv.2011.10.046

Schaap, M., Timmermans, R.M.A., Sauter, F.J., Roemer, M., Velders, G.J.M., and et al. 2008. The LOTOS-EUROS model: description, validation and latest developments. Int. J. environ. Pollut. 32, 270-290.

Schere, K., Flemming, J., Vautard, R., Chemel, C., et al. Trace Gas/Aerosol concentrations and their impacts on continental-scale AQMEII modelling subregions. Atmospheric Environment doi:10.1016/j.atmosenv.2011.09.043

Schmidt, H., Derognat, C., Vautard, R., and Beekmann, M., 2001. A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe. Atmospheric Environment, 36, 6277-6297.

Simpson, D., A. Guenther, C. N. Hewitt, R. Steinbrecher, 1995. Biogenic emissions in Europe. 1. Estimates and uncertainties. J. Geophys. Res., 100D, 22875–22890.

Simpson, D., Fagerli, H., Jonson, J.E., Tsyro, S., Wind, P., and Tuovinen, J.-P., 2003. The EMEP Unified Eulerian Model. Model Description. Technical Report EMEP MSC-W Report 1/2003.The Norwegian Meteorological Institute, Oslo, Norway.

Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D. O., Barker, D. M.,Duda, M. G., Huang, X.-Y., Wang, W., and Powers, J. G. (2008) A description of the Advanced Research WRF Version 3, National Center for Atmospheric Research, Tech. Note, NCAR/TN-475+STR, 113 pp.

Smyth, S.C., W. Jiang, H. Roth, M.D. Moran, P.A. Makar, F. Yang, V.S. Bouchet, and H. Landry, 2009: A comparative performance evaluation of the AURAMS and CMAQ air quality modelling systems. Atmos. Environ., 43, 1059-1070.

Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., Kukkonen, J., 2006. A dispersion modeling system SILAM and its evaluation against ETEX data. Atmospheric Environment 40, 674-685.

Solazzo, E., Bianconi, R., Pirovano, G., Volker, M., Vautard, R., and et al., 2012 .Operational model evaluation for particulate matter in Europe and North America in the context of the AQMEII project. Atmospheric Environment

Talagrand, O., Vautard, R., Strauss, B., 1998. Evaluation of probabilistic prediction systems, paper presented at Seminar on Predictability, Eur. Cent. for Medium Weather Forecasting, Reading, UK.

van Loon, M., Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., and et al. 2007. Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble average. Atmos. Environ. 41, 2083-2097.

Vautard, R., M. Schaap, R Bergström, B. Bessagnet, J. Brandt, P.J.H. Builtjes, J.H. Christensen, C. Cuvelier, V. Foltescu, A. Graf, A. Kerschbaumer, M. Krol, P. Roberts, L. Rouïl, R. Stern, L. Tarrason, P. Thunis, E. Vignati, P. Wind, 2009, Skill and uncertainty of a regional air quality model ensemble. Atmos. Environ., 43, 4822-4832.

Vautard, R., Moran, M.D., Solazzo, E., Gilliam, R.C., Matthias, V., et al. Evaluation of the meteorological forcing used for AQMEII air quality simulations. Atmospheric Environment 10.1016/j.atmosenv.2011.10.065

Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., and et al. 2009. Skill and uncertainty of a regional air quality model ensemble. Atmos. Environ., 43, 4822-4832.

Vautard, R., van Loon, M., Schaap, M., Bergström, R., Bessagnet, B., et al. 2006. Is regional air quality model diversity representative of uncertainty for ozone simulation ? Geophys. Res. Lett., 33, L24818, doi:10.1029/2006GL027610.

Wolke, R., Knoth, O., Hellmuth, O., Schröder, W., Renner, E. 2004. The parallel model system LM-MUSCAT for chemistry-transport simulations: Coupling scheme, parallelization and applications. Parallel Computing, 363–370.

Captions

Table 1. Participating models and features

Table 2. RMSE-ranked combinations of models that give minimum RMSE. In bold the minimum of all combinations

Table 3. Statistical skills for all members ensemble (first row of each domain), ensemble of minimum RMSE (second row).σ is the standard deviation in μg m^{-3} for EU and ppb for NA.

Table 4. Ranking of cluster representatives for EU and NA subregions for varying PCC distance.

Figure 1 – Continental maps of *a)* Europe and *b)* North America and subregions by colours. The dots are the positions of ozone receptors used in the analysis. The contours are the anthropogenic NOx emissions (in kg km^{-2}) using the standard inventories.

Figure 2. Ozone ranges at receptors, averaged in space over *a)* EU domain and *b)* NA domain and in time for the whole 2006 year.

Figure 3. Ranked histogram for the whole subregions of *a)* EU and *b)* NA, full model ensemble, hourly data for the whole 2006 year.

Figure 4. Time series (JJA) of diurnal ozone cycle for *a)* EU and *b)* NA subregions

Figure 5. Mean Bias vs Normalised Mean Square Error for *a)* EU and *b)* NA. Subregions 1 to 4 represented by number, coloured by model or ensemble. Mean and median for each subregion are highlighted

Figure 6. Average vertical profiles of ozone in the vicinity of Portland airport. Measurements are the monthly mean MOZAIC for August with the standard deviation; the thick grey line is the GEMS BC at a node close to Portland; thin grey lines are the ozone predictions by three models employing GEMS BCs; the brown square is the ozone concentration at surface receptors in the vicinity of Portland.

Figure 7. Ranked histogram for *a)* EU and *b)* NA by subregions, full model ensemble, hourly data for the period JJA.

Figure 8. RMSE, MGE, MB, PCC of the ensemble mean of any possible combination of members for *a)* EU, and *b)* NA. Continuous lines are the mean and the median of the distributions

Figure 9. Daily maximum concentrations for EU subregions, for the period JJA. Horizontal axis: ensemble maximum of all available members. Vertical axis: ensemble maximum of model combinations with minimum RMSE.

Figure 10. Ozone concentrations (μg m^{-3}) for the period JJA at receptors position. *a)* observations, *b)* ensemble of ranked models 1 and 5; models ranked *c)* 5th and *d)*11th.

Figure 11. Box-plot of observed ozone concentration, full model ensemble and selected (combinations with minimum RMSE) model ensemble. Top row: EU subregions; bottom row: NA subregions.

Figure 12 Dendrograms of models clustering as function of mutual PCC distance for EU subregions

Figure 13 Dendrograms of models clustering as function of mutual PCC distance for NA subregions

Figure 14 Curves of minimum (thick lines) and maximum (dotted lines) RMSE obtained by connecting min and max of Fig. 8. The short lines are the RMSE of MM ensembles from clustering analysis (combinations of Table 4). The labels are the individual RMSE-ranking of MM members. Different colours correspond to different subregions for *a)* EU and *b)* NA.

Tables

Table 1

	Model		**Res**	**n.Vertical**	**Emission**	**Chemical BC**
	Met	**AQ**	**(km)**	**layers**		
European Domain	MM5	DEHM	50	29	Global emission databases, EMEP	Satellite measurements
	MM5	Polyphemus	24	9	Standard[§]	Standard
	PARLAM-PS	EMEP	50	20	EMEP model[§]	From ECMWF and forecasts
	WRF	CMAQ	18	34	Standard[§]	Standard
	WRF	WRF-Chem	22.5	36	Standard[§]	Standard
	WRF	WRF-Chem	22.5	36	Standard[§]	Standard
	ECMWF	SILAM	24	9	Standard anthropogenic In-house biogenic	Standard
	MM5	Chimere	25	9	MEGAN, Standard	Standard
	LOTOS	EUROS	25	4	Standard[§]	Standard
	COSMO	Muscat	24	40	Standard[§]	Standard
	MM5	CAMx	15	20	MEGAN, Standard	Standard
North American Domain*	GEM	AURAMS	45	28	Standard[+]	Climatology
	WRF	Chimere	36	9	Standard	LMDZ-INCA
	MM5	CAMx	24	15	Standard	LMDZ-INCA
	WRF	CMAQ	12	34	Standard	Standard
	WRF	CAMx	12	26	Standard	Standard
	WRF	Chimere	36	9	Standard	standard
	MM5	DEHM	50	29	global emission databases, EMEP	Satellite measurements

[§] Standard anthropogenic emission and biogenic emission derived from meteorology (temperature and solar radiation) and land use distribution implemented in the meteorological driver (Guenther et al., 1994; Simpson et al., 1995).
*Standard inventory for NA includes biogenic emissions (see text).
[+]Standard anthropogenic inventory but independent emission processing, excluding wildfire and different version of BEIS (v3.09) used.

Table 2

		Number of Models					
		2	3	4	5	6	7
EU	dom1	1-2	**1-5-11**	1-2-7-11	1-2-5-7-11	1-2-4-5-6-11	1-2-3-4-5-6-11
	dom2	3-8	2-3-8	**2-3-5-8**	1-2-3-5-8	1-2-3-4-5-8	1-2-3-4-5-6-8
	dom3	2-3	**2-3-5**	1-2-3-5	1-2-3-9-11	1-2-3-5-8-11	1-2-3-5-8-9-11
	dom4	5-9	1-6-9	2-6-7-9	**1-6-9-10-11**	**1-2-6-9-10-11**	1-2-3-6-9-10-11
NA	dom1	1-2	1-2-3	**1-2-4-6**	1-2-3-4-6	1-2-3-4-6-7	
	dom2	1-2	**1-3-4**	1-2-3-4	1-2-3-4-5	1-2-3-4-5-6	
	dom3	2-3	**1-2-3**	1-2-3-4	1-2-3-4-5	1-2-3-4-5-6	
	dom4	**1-2**	1-2-3	1-2-3-4	1-2-3-4-5	1-2-3-4-5-6	

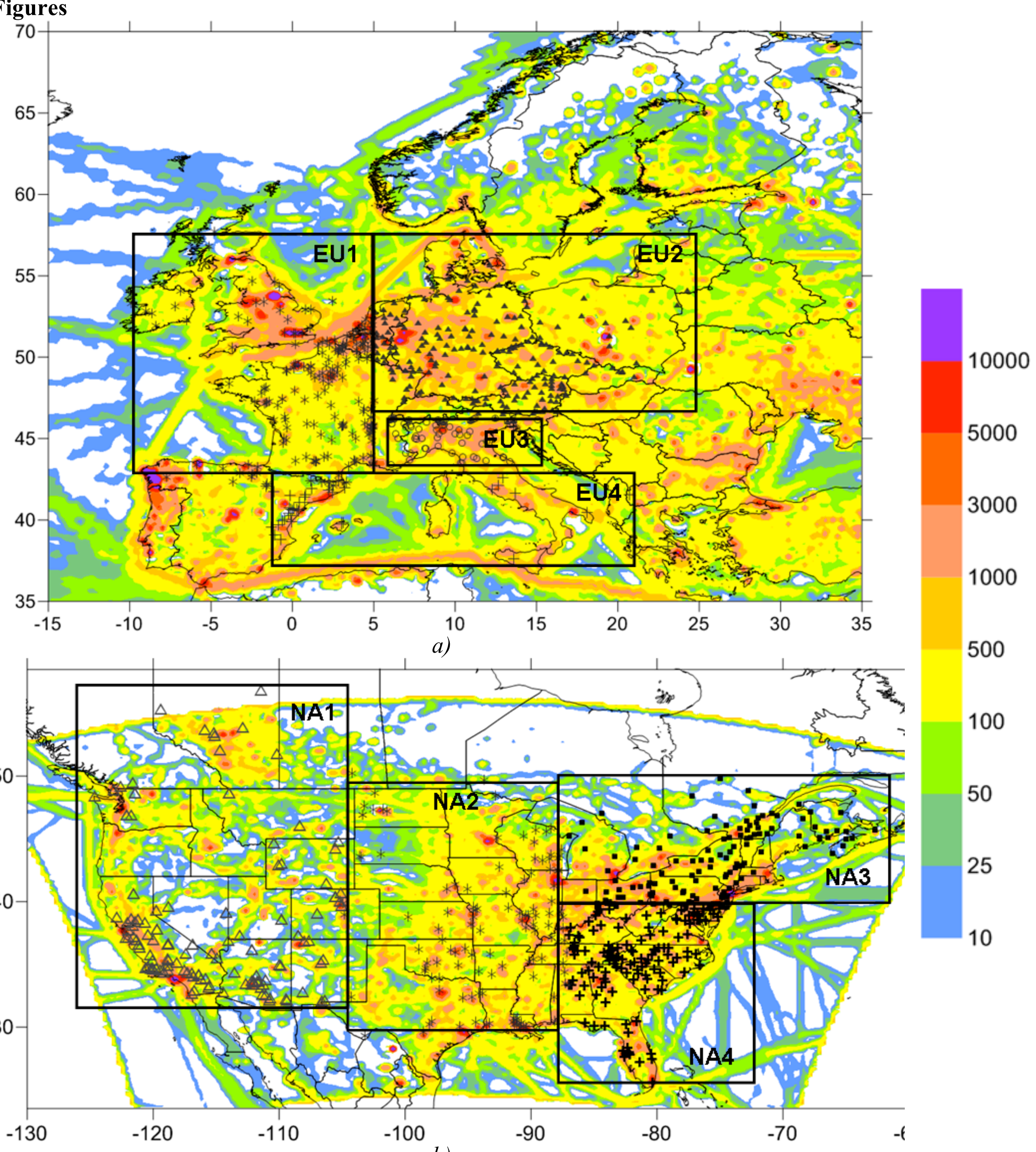
Table 3.

		Bias	FBias	RMSE	PCC	σ
EU	**Dom 1**	-5.11	-0.08	12.01	0.97	18.29
	σ_{obs}=27.4	-0.82	-0.01	8.49	0.96	22.24
	Dom 2	-8.77	-0.11	13.50	0.93	17.59
	σ_{obs}=24.5	1.35	0.02	7.78	0.95	22.29
	Dom 3	-4.87	-0.06	17.38	0.89	20.25
	σ_{obs}=31.7	-2.34	-0.03	14.90	0.90	24.17
	Dom 4	-1.11	-0.013	8.27	0.92	17.25
	σ_{obs}=20.7	-1.25	-0.014	7.27	0.94	18.34
NA	**Dom 1**	0.66	0.02	3.63	0.94	12.3
	σ_{obs}=10.13	-0.11	-0.003	3.45	0.94	12.1
	Dom 2	3.90	0.10	6.40	0.92	11.80
	σ_{obs}=12.83	2.05	0.05	4.82	0.92	12.6
	Dom 3	4.51	0.13	7.34	0.85	12.5
	σ_{obs}=10.36	2.55	0.07	5.8	0.87	10.5
	Dom 4	10.55	0.26	12.35	0.90	12.3
	σ_{obs}=14.50	5.10	0.13	7.98	0.91	14.2

Table 4

	distance	Number of members	Ranking of cluster representatives
EU1	PCC > 0.06	3	6-2-8/9
	PCC =0.05	4	3-2-8/9-11
	PCC = 0.03	5	3-2-8/9- 11-1/10
EU2	PCC>0.045	4	6-1/8-2-7/9
EU3	PCC>0.08	3	3-6/7- 1
	PCC=0.06	5	3-11- 6/7- 1-9
EU4	PCC > 0.04	4	1-4/5-2-11
	PCC =0.02	6	1-4/5-2/10-9-11/7-8
NA1	PCC>0.04	3	3/4-1/5- 2
NA2	PCC>0.012	3	3/5-6/7-1
NA3	PCC>0.035	2	2/4-6
	PCC=0.03	4	4-2-3-6/7
NA4	PCC>0.025	3	4/7- 5/6-3

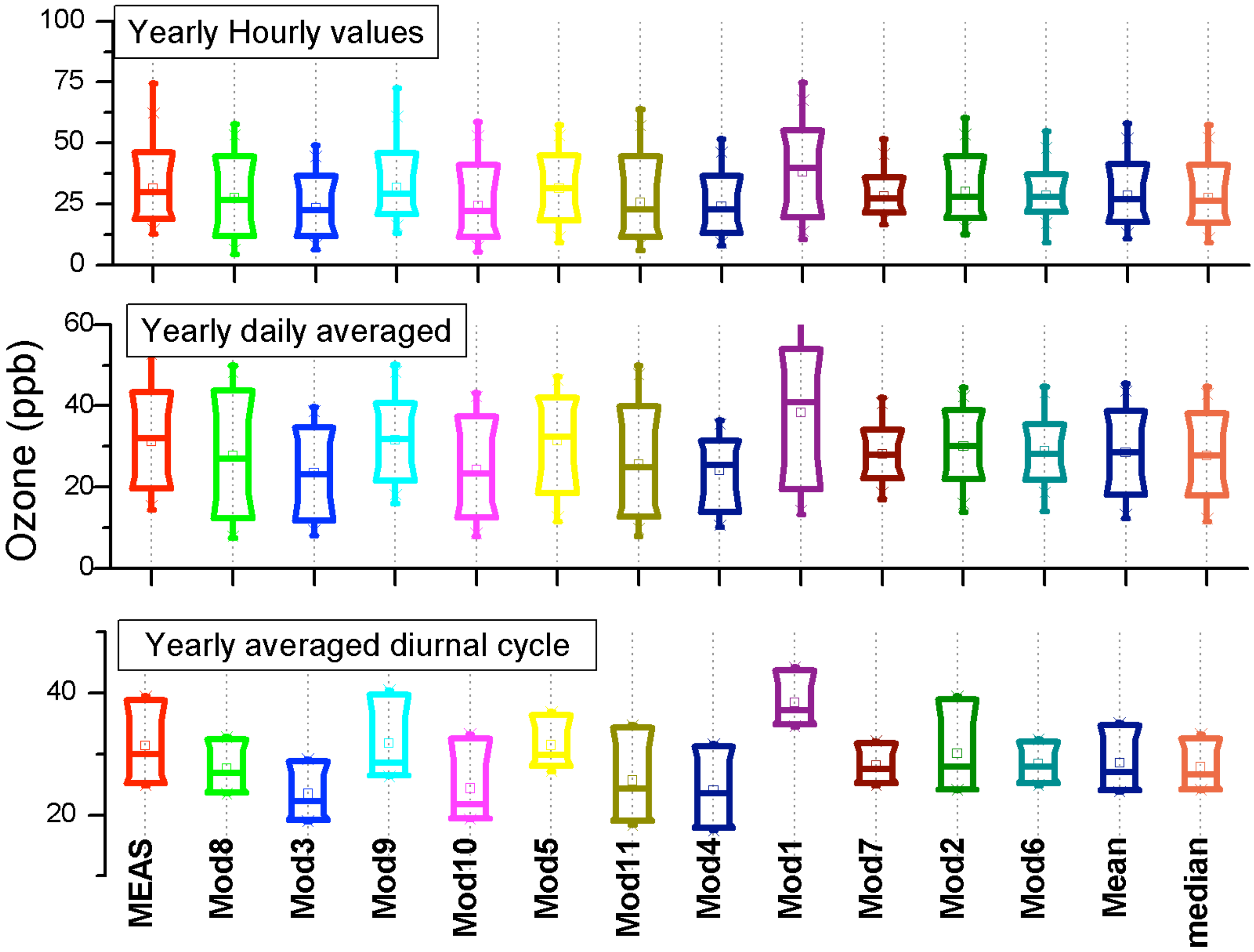
Figures



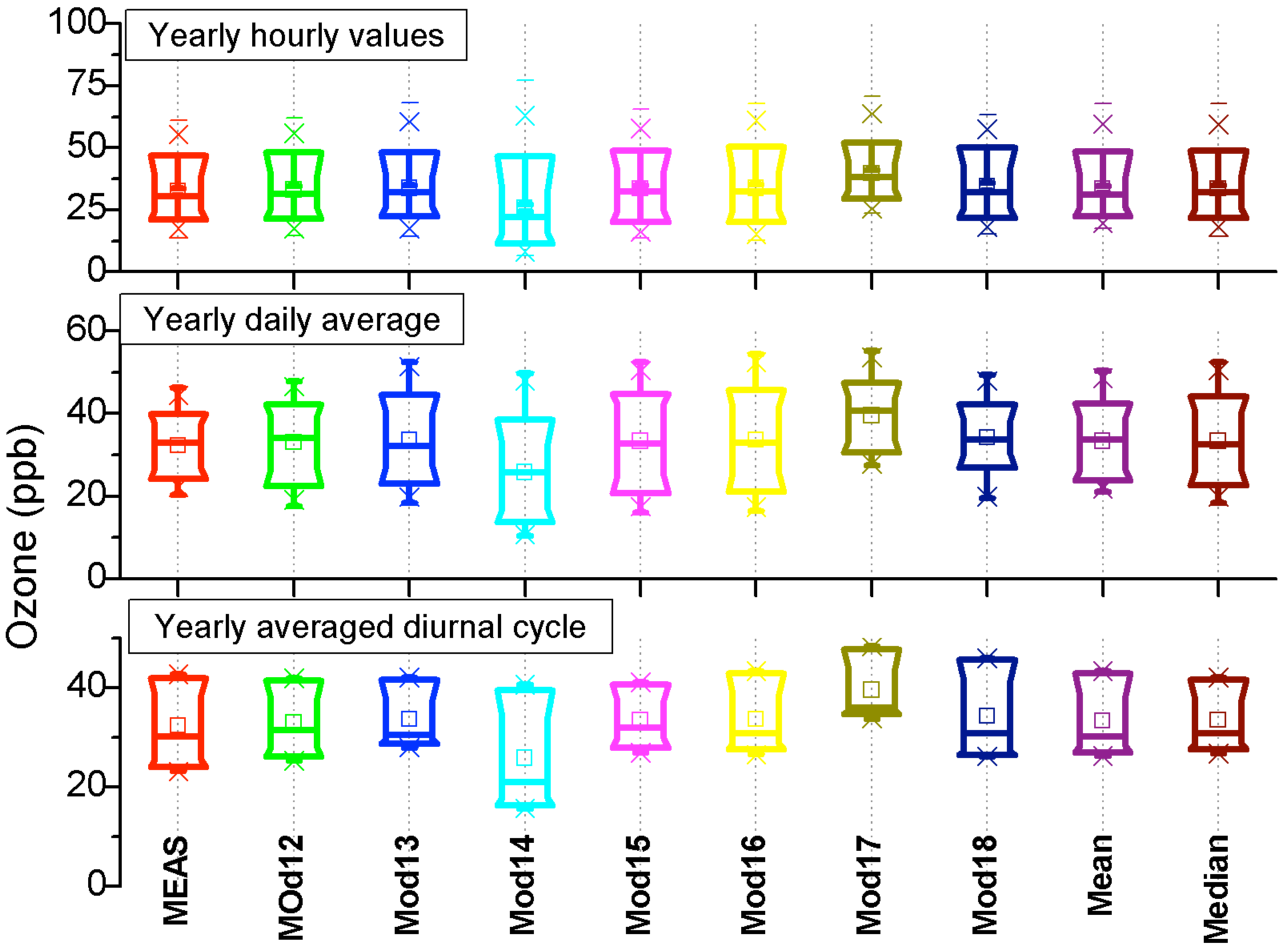
a)

b)

Figure 1

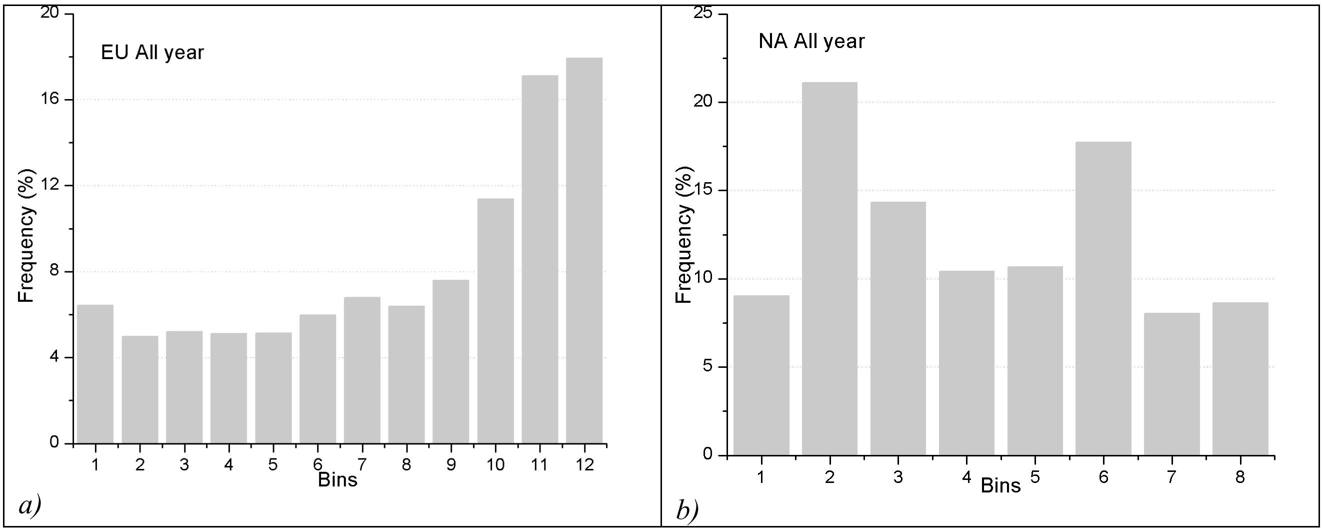

Yearly Hourly values
100
75
50
25
0
Yearly daily averaged
60
40
20
0
Ozone (ppb)
Yearly averaged diurnal cycle
40
20
MEAS
Mod8
Mod3
Mod9
Mod10
Mod5
Mod11
Mod4
Mod1
Mod7
Mod2
Mod6
Mean
median

a)

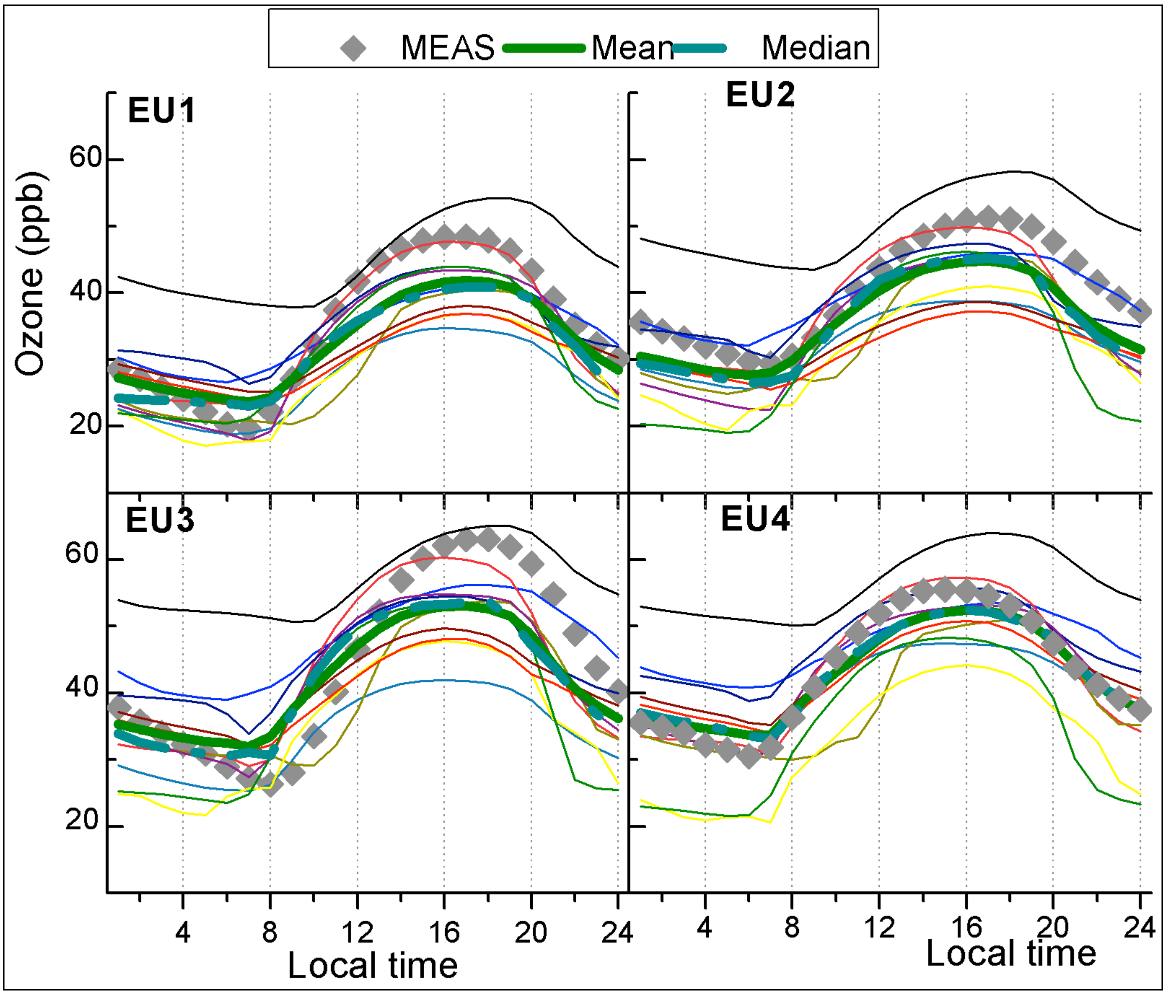


b)

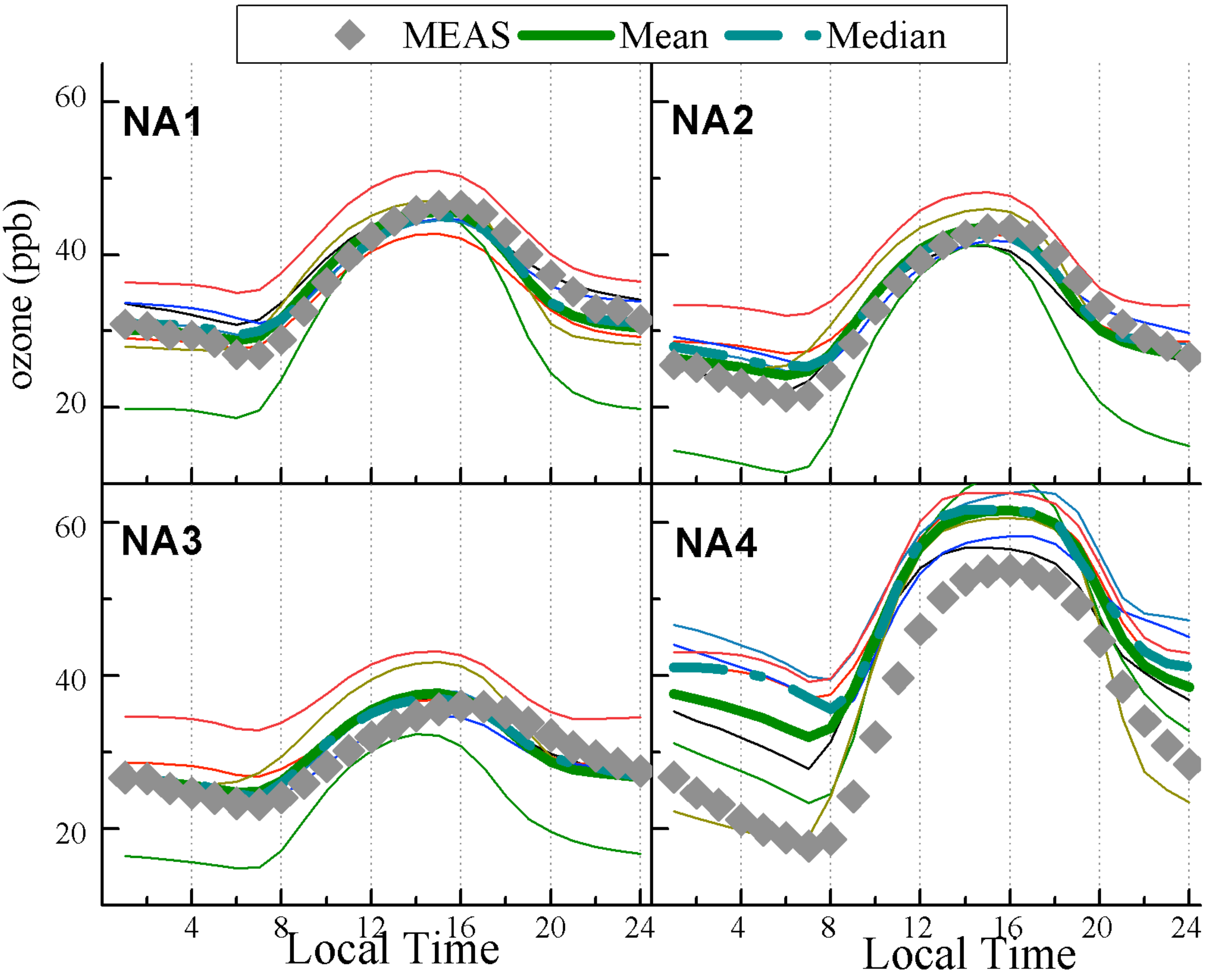
958 **Figure 2**



960 **Figure 3**


MEAS
Mean
Median
EU1
EU2
EU3
EU4
Ozone (ppb)
60
40
20
4
8
12
16
20
24
Local time

a)

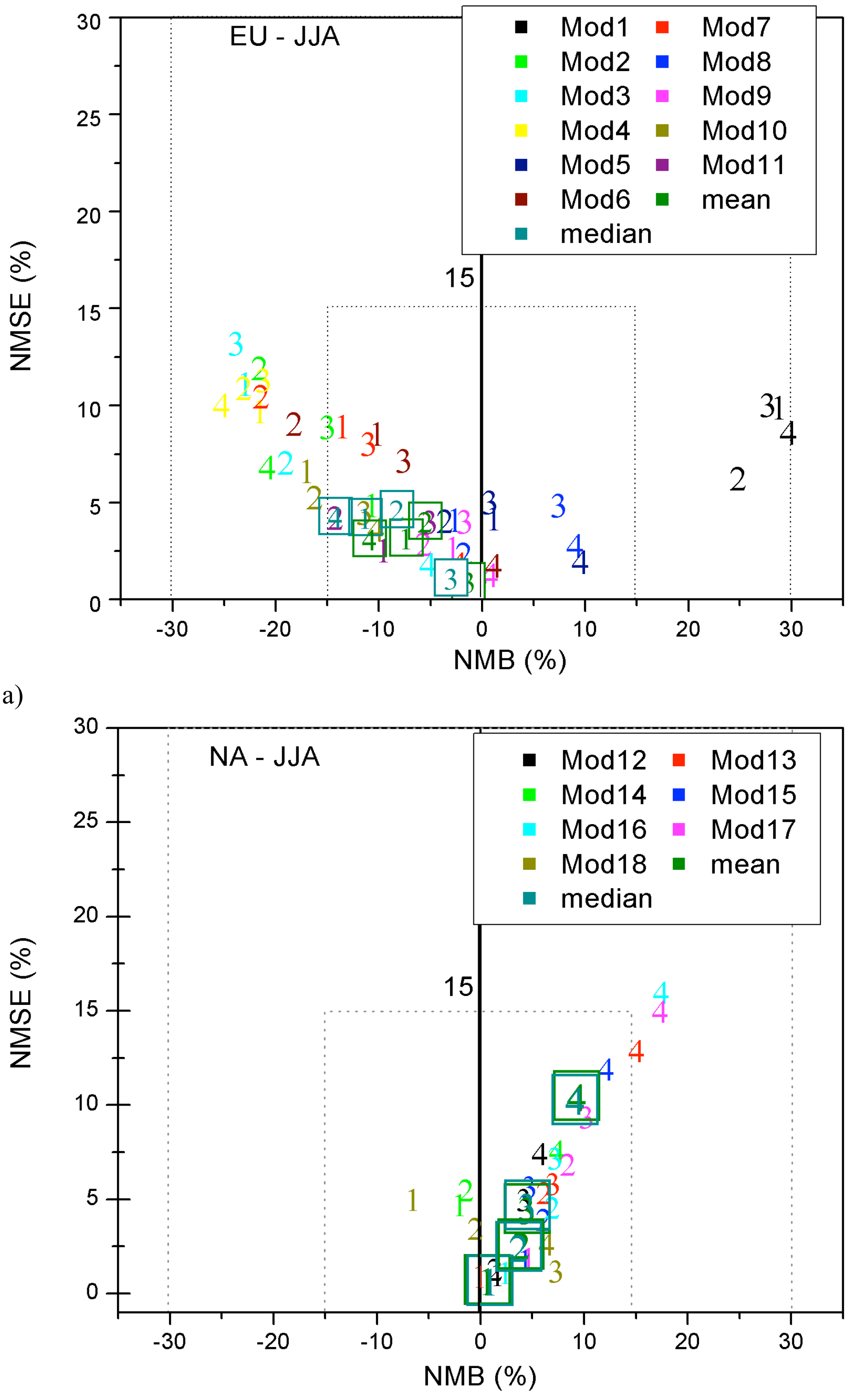


b)

962 **Figure 4.**

964

966


EU - JJA
Mod1
Mod2
Mod3
Mod4
Mod5
Mod6
median
Mod7
Mod8
Mod9
Mod10
Mod11
mean
15
NMSE (%)
NMB (%)
30
25
20
15
10
5
0
-30
-20
-10
0
10
20
30
NA - JJA
Mod12
Mod14
Mod16
Mod18
median
Mod13
Mod15
Mod17
mean
15
NMSE (%)
NMB (%)
30
25
20
15
10
5
0
-30
-20
-10
0
10
20
30

a)

b)

Figure 5.

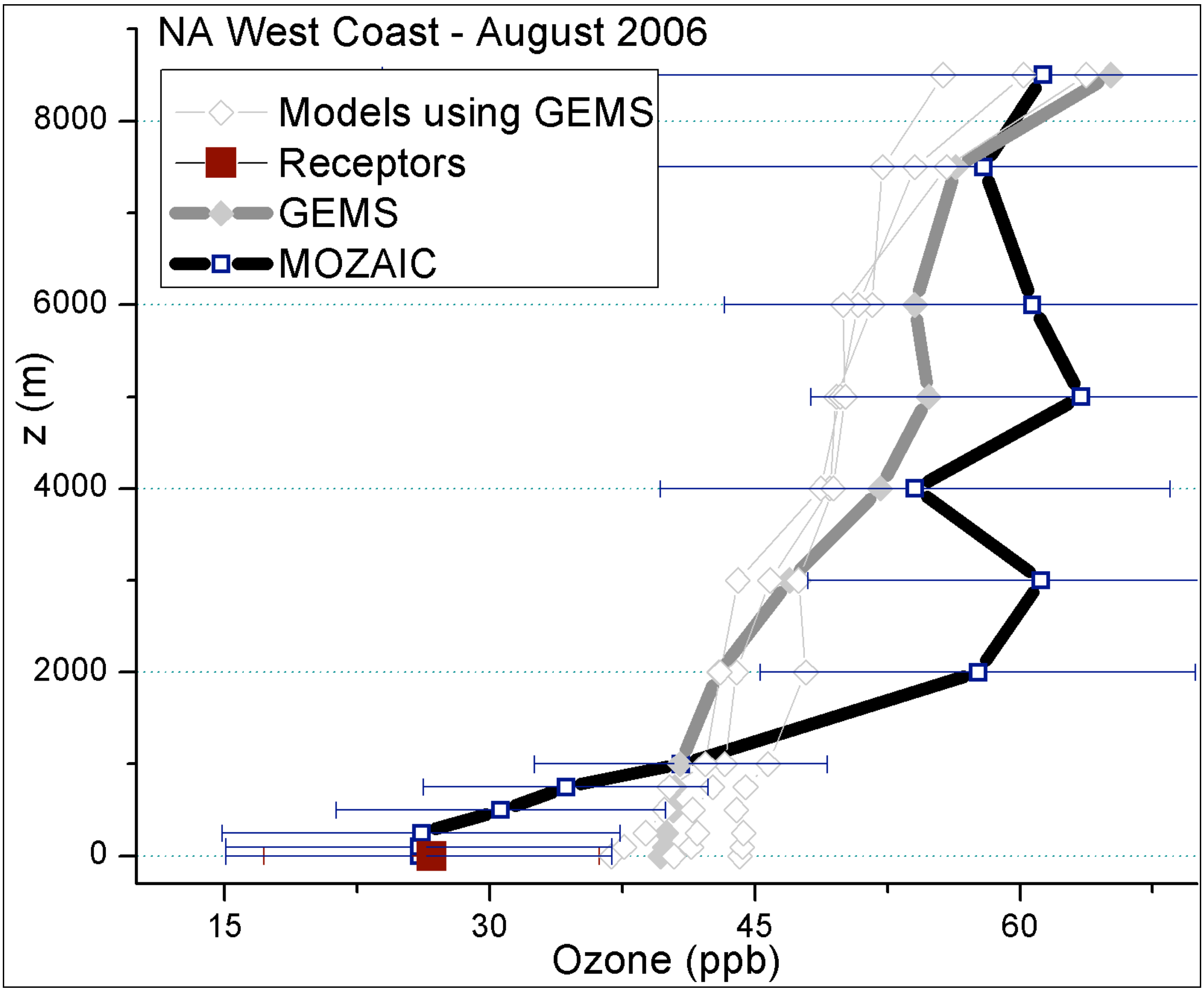

NA West Coast - August 2006
Models using GEMS
Receptors
GEMS
MOZAIC
z (m)
8000
6000
4000
2000
0
15
30
45
60
Ozone (ppb)

Figure 6

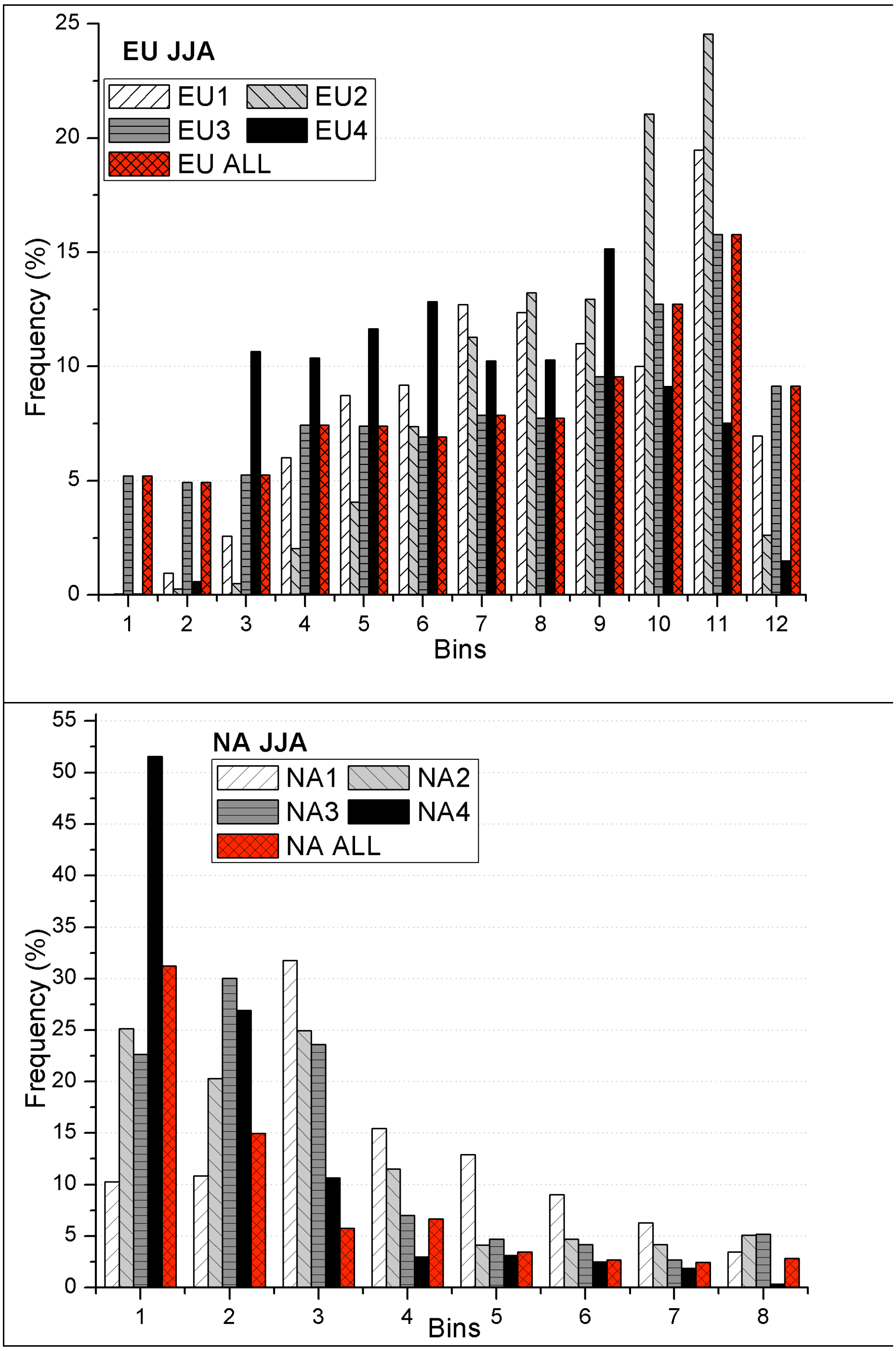
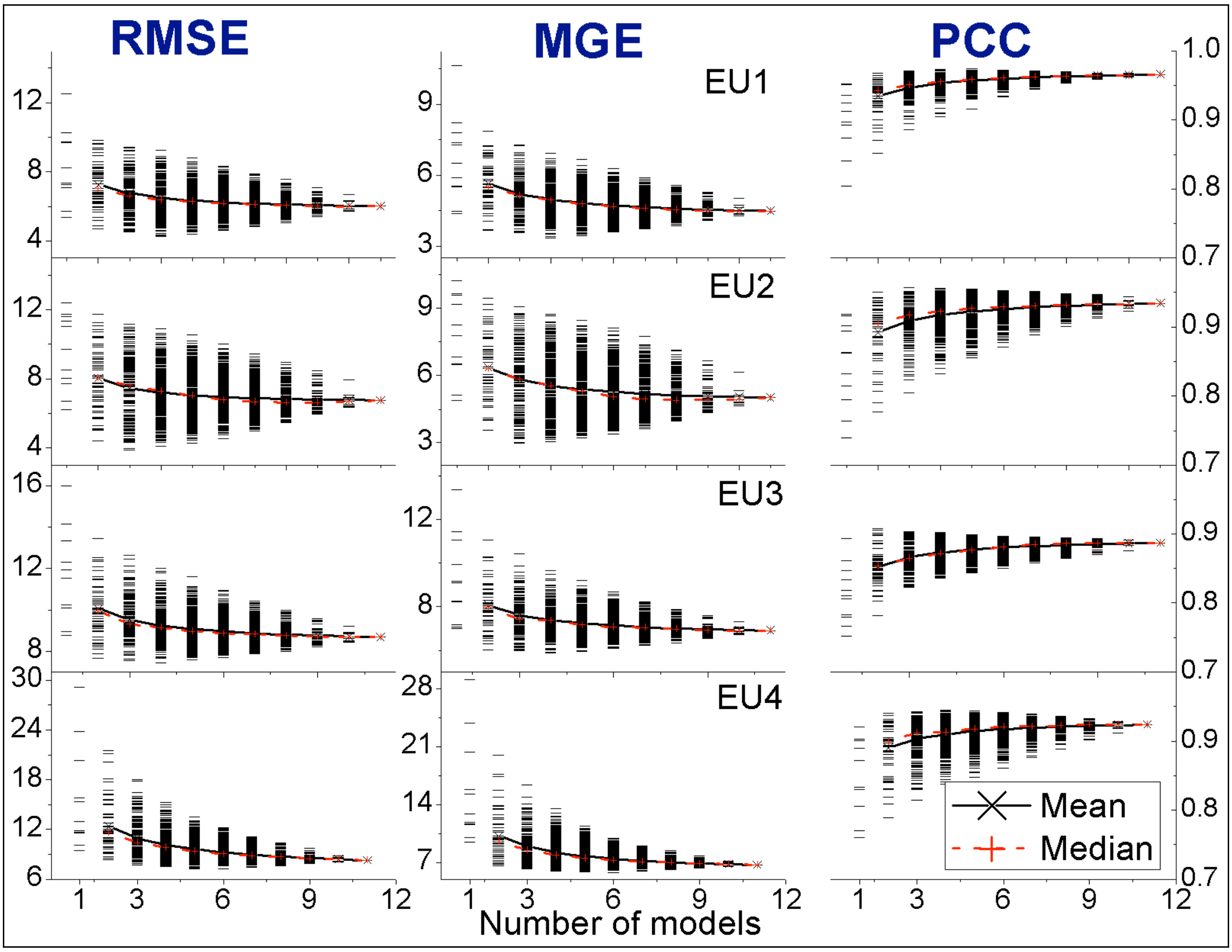
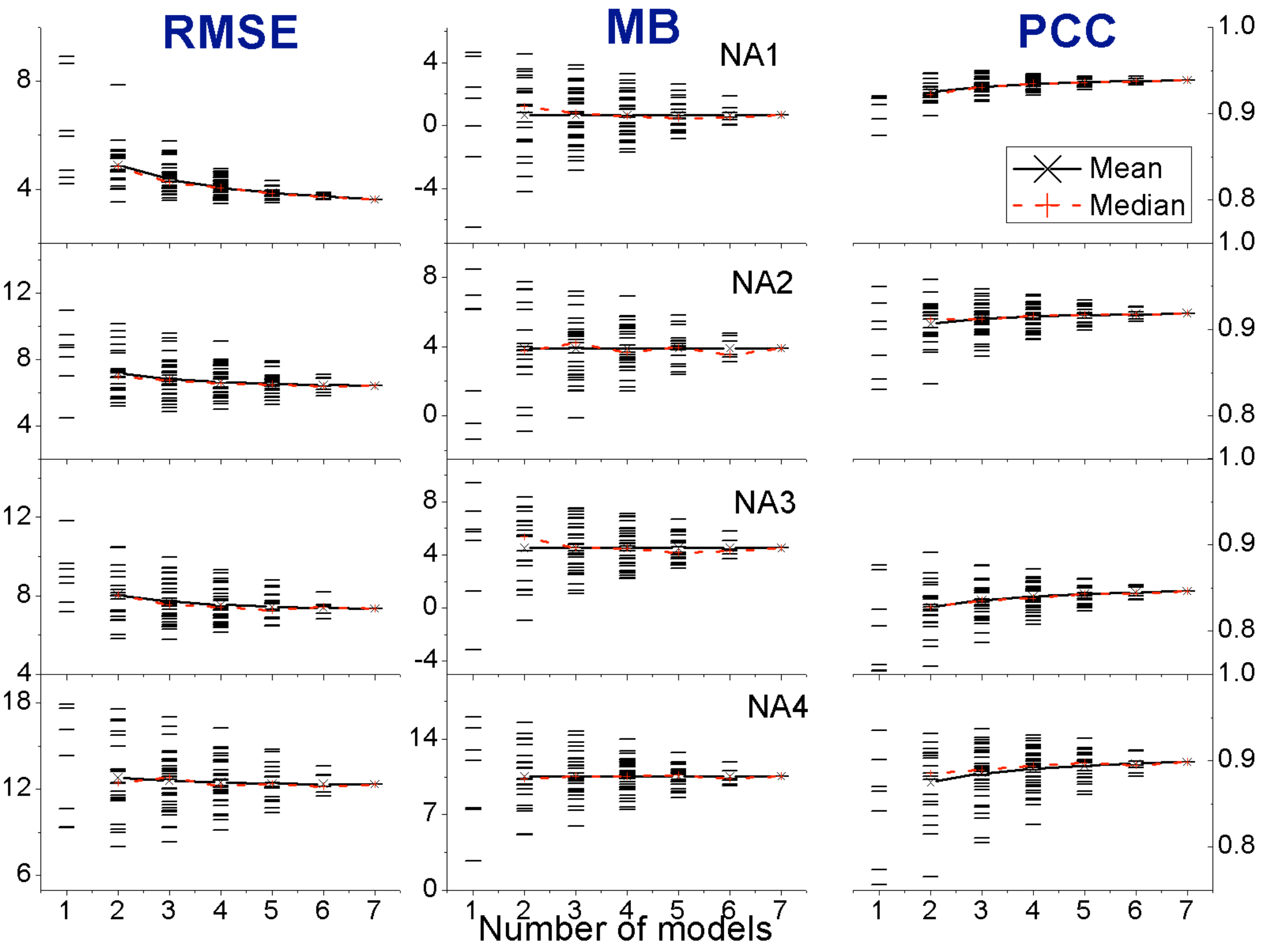

EU JJA
EU1
EU2
EU3
EU4
EU ALL
Frequency (%)
0
5
10
15
20
25
1
2
3
4
5
6
7
8
9
10
11
12
Bins
NA JJA
NA1
NA2
NA3
NA4
NA ALL
Frequency (%)
0
5
10
15
20
25
30
35
40
45
50
55
1
2
3
4
5
6
7
8
Bins

Figure 7

986



988 **Figure 8a**



990 **Figure 8b**

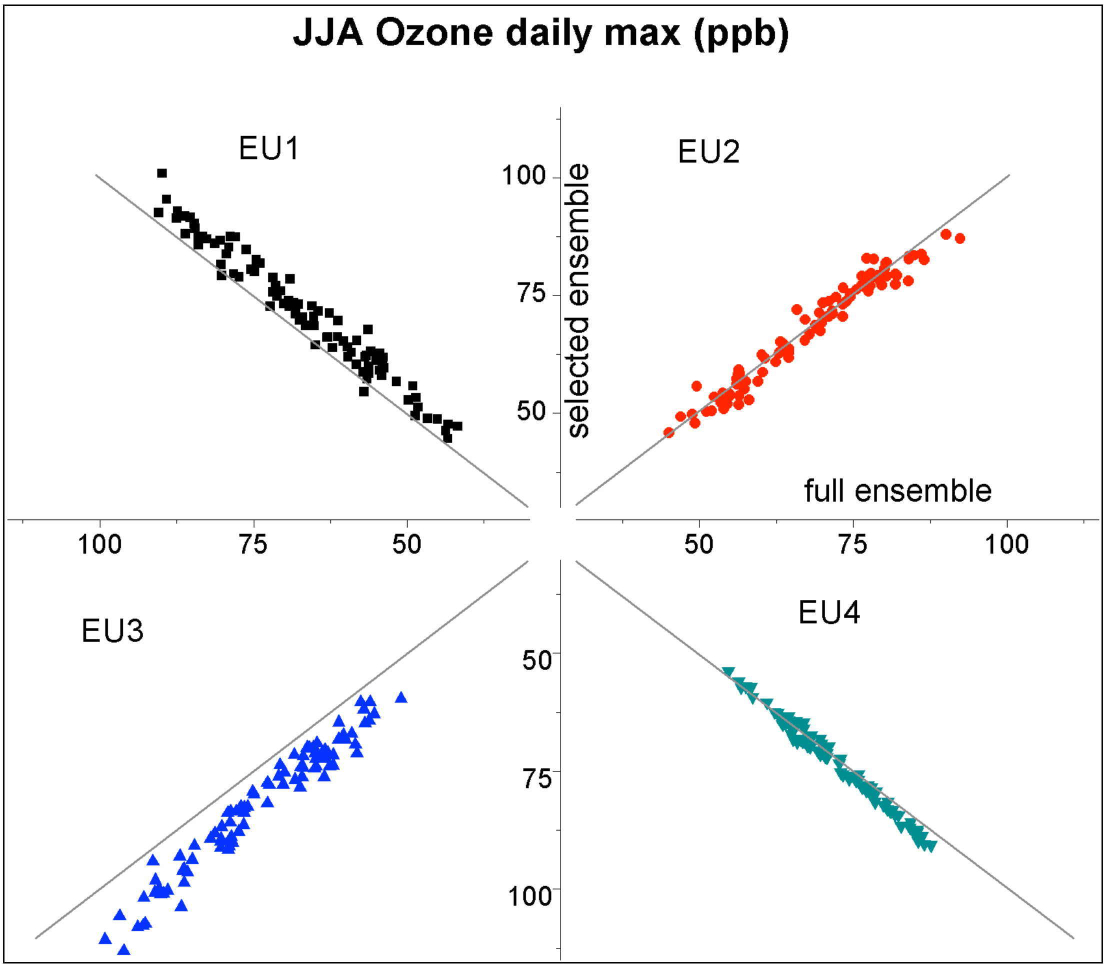

JJA Ozone daily max (ppb)
EU1
EU2
EU3
EU4
selected ensemble
full ensemble
100
75
50
50
75
100
100
75
50
50
75
100

Figure 9

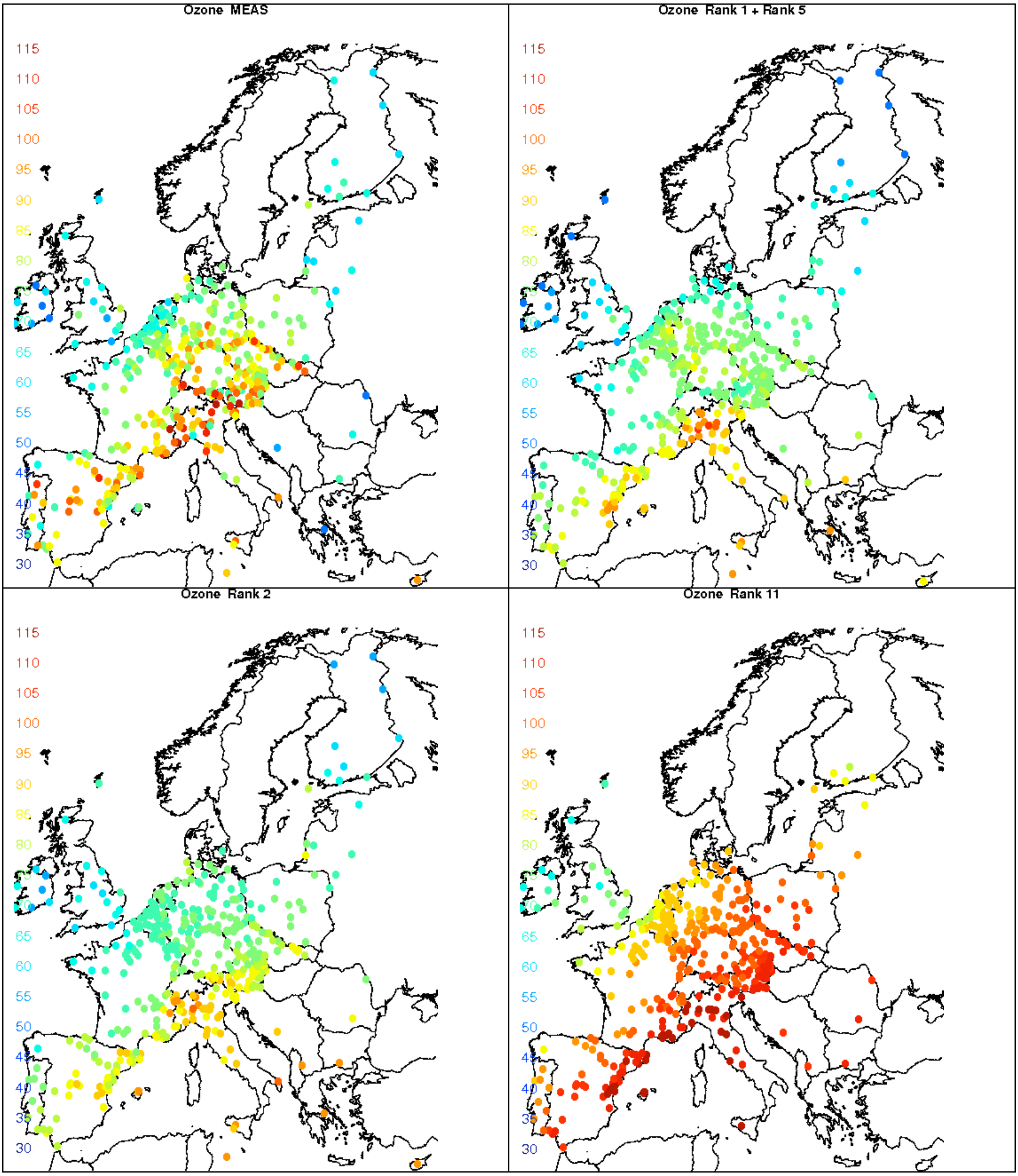

Ozone MEAS
Ozone Rank 1 + Rank 5
Ozone Rank 2
Ozone Rank 11
115
110
105
100
95
90
85
80
65
60
55
50
45
40
35
30

Figure 10

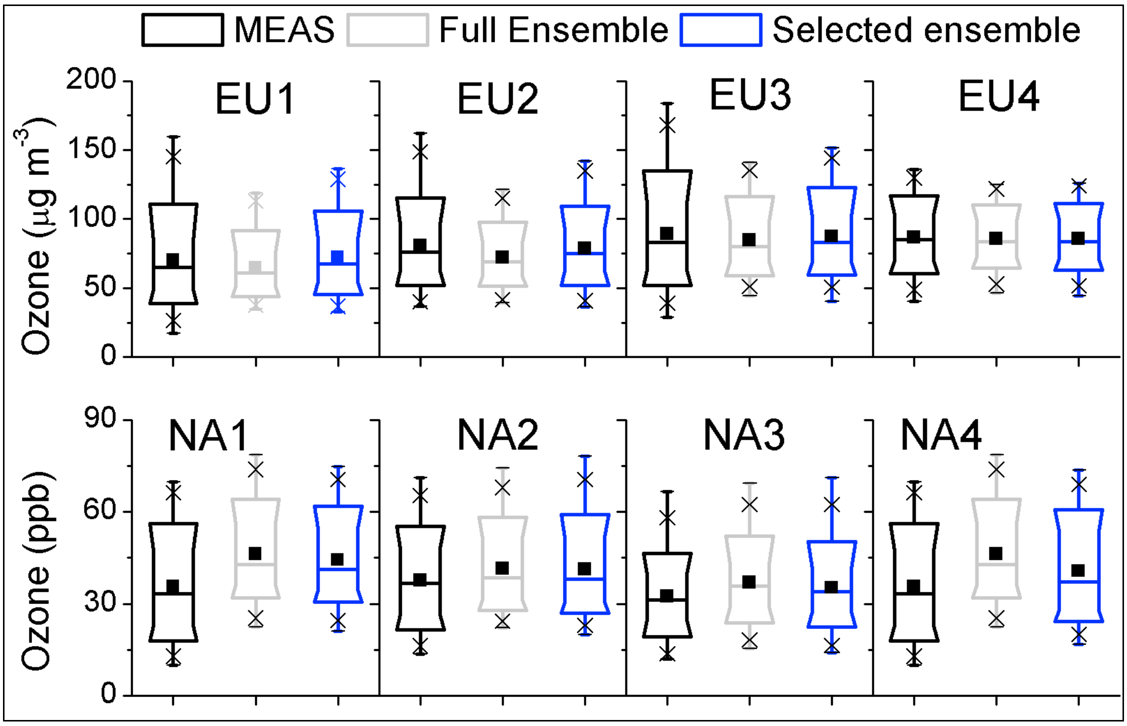

MEAS
Full Ensemble
Selected ensemble
Ozone (µg m^{-3})
200
150
100
50
0
EU1
EU2
EU3
EU4
Ozone (ppb)
90
60
30
0
NA1
NA2
NA3
NA4

Figure 11

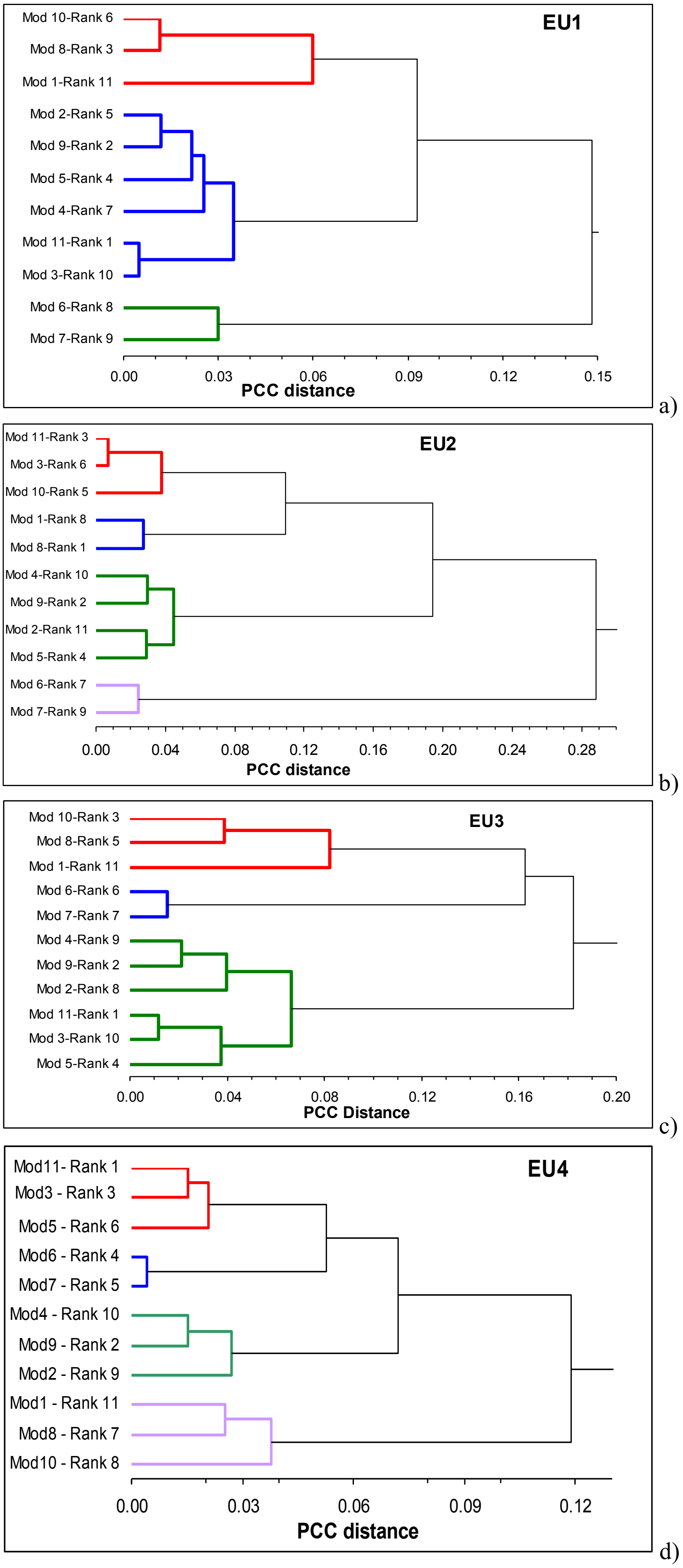

EU1
Mod 10-Rank 6
Mod 8-Rank 3
Mod 1-Rank 11
Mod 2-Rank 5
Mod 9-Rank 2
Mod 5-Rank 4
Mod 4-Rank 7
Mod 11-Rank 1
Mod 3-Rank 10
Mod 6-Rank 8
Mod 7-Rank 9
0.00
0.03
0.06
0.09
0.12
0.15
PCC distance
a)
EU2
Mod 11-Rank 3
Mod 3-Rank 6
Mod 10-Rank 5
Mod 1-Rank 8
Mod 8-Rank 1
Mod 4-Rank 10
Mod 9-Rank 2
Mod 2-Rank 11
Mod 5-Rank 4
Mod 6-Rank 7
Mod 7-Rank 9
0.00
0.04
0.08
0.12
0.16
0.20
0.24
0.28
PCC distance
b)
EU3
Mod 10-Rank 3
Mod 8-Rank 5
Mod 1-Rank 11
Mod 6-Rank 6
Mod 7-Rank 7
Mod 4-Rank 9
Mod 9-Rank 2
Mod 2-Rank 8
Mod 11-Rank 1
Mod 3-Rank 10
Mod 5-Rank 4
0.00
0.04
0.08
0.12
0.16
0.20
PCC Distance
c)
EU4
Mod11- Rank 1
Mod3 - Rank 3
Mod5 - Rank 6
Mod6 - Rank 4
Mod7 - Rank 5
Mod4 - Rank 10
Mod9 - Rank 2
Mod2 - Rank 9
Mod1 - Rank 11
Mod8 - Rank 7
Mod10 - Rank 8
0.00
0.03
0.06
0.09
0.12
PCC distance
d)

Figure 12

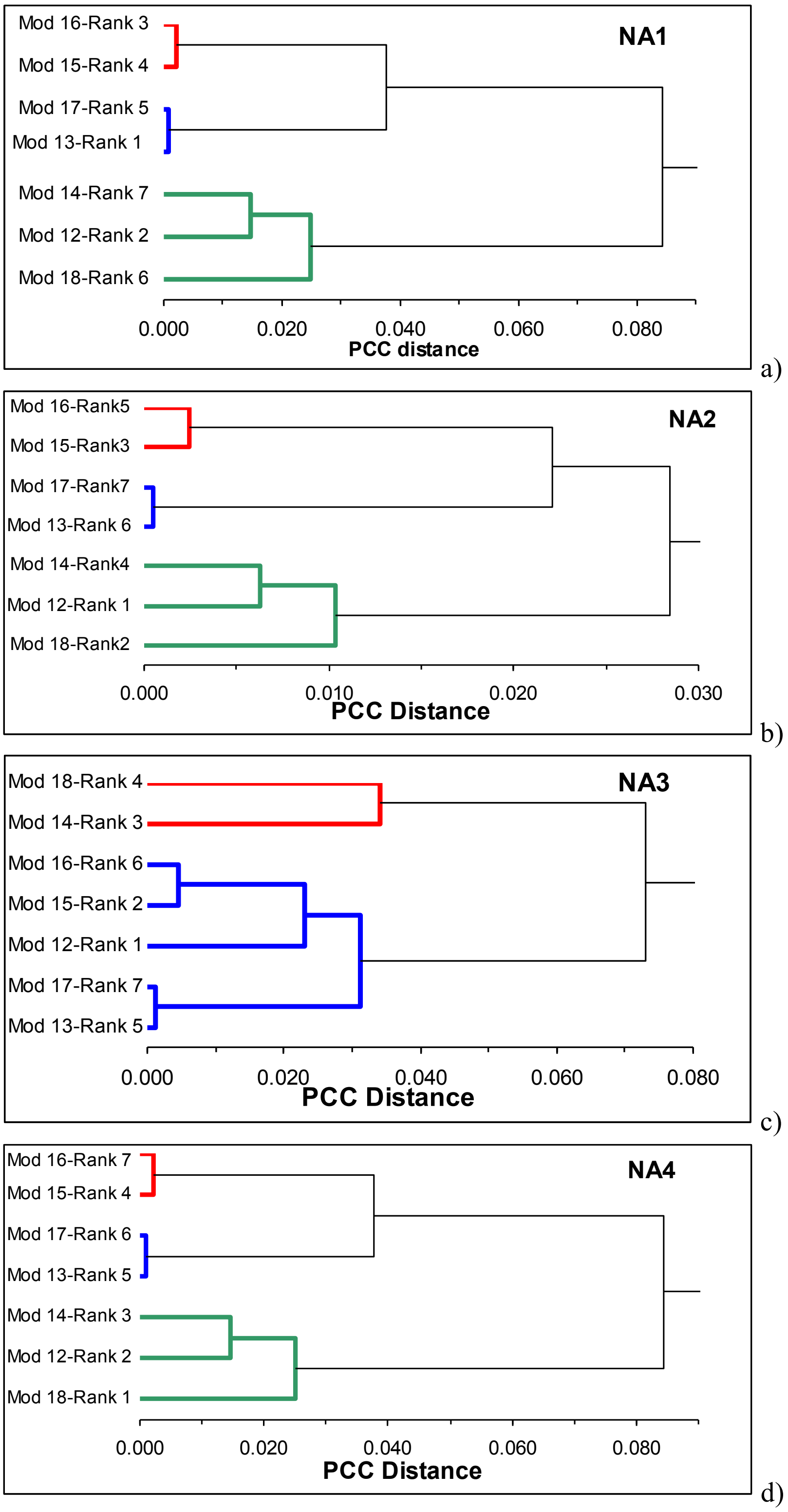

NA1
Mod 16-Rank 3
Mod 15-Rank 4
Mod 17-Rank 5
Mod 13-Rank 1
Mod 14-Rank 7
Mod 12-Rank 2
Mod 18-Rank 6
0.000
0.020
0.040
0.060
0.080
PCC distance
a)
NA2
Mod 16-Rank5
Mod 15-Rank3
Mod 17-Rank7
Mod 13-Rank 6
Mod 14-Rank4
Mod 12-Rank 1
Mod 18-Rank2
0.000
0.010
0.020
0.030
PCC Distance
b)
NA3
Mod 18-Rank 4
Mod 14-Rank 3
Mod 16-Rank 6
Mod 15-Rank 2
Mod 12-Rank 1
Mod 17-Rank 7
Mod 13-Rank 5
0.000
0.020
0.040
0.060
0.080
PCC Distance
c)
NA4
Mod 16-Rank 7
Mod 15-Rank 4
Mod 17-Rank 6
Mod 13-Rank 5
Mod 14-Rank 3
Mod 12-Rank 2
Mod 18-Rank 1
0.000
0.020
0.040
0.060
0.080
PCC Distance
d)

Figure 13

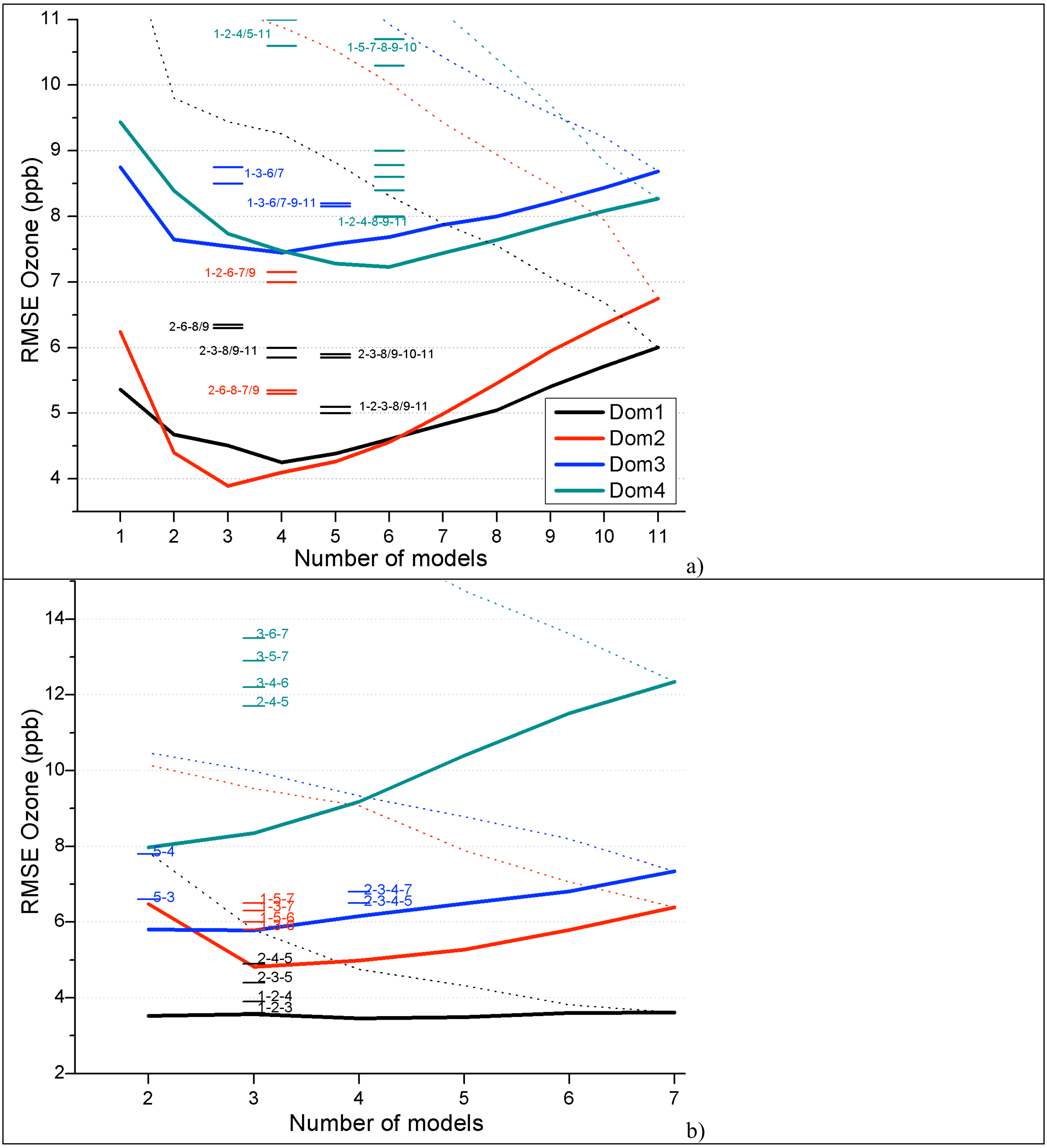

RMSE Ozone (ppb)
Number of models
1-2-4/5-11
1-5-7-8-9-10
1-3-6/7
1-3-6/7-9-11
1-2-4-8-9-11
1-2-6-7/9
2-6-8/9
2-3-8/9-11
2-3-8/9-10-11
2-6-8-7/9
1-2-3-8/9-11
Dom1
Dom2
Dom3
Dom4
a)
3-6-7
3-5-7
3-4-6
2-4-5
5-4
5-3
1-5-7
1-3-7
1-5-6
1-3-6
2-3-4-7
2-3-4-5
2-4-5
2-3-5
1-2-4
1-2-3
b)

Figure 14