

Model evaluation and ensemble modelling of surface-level ozone in

Europe and North America in the context of AQMEII

Ef시오 Solazzo^{1,}, Roberto Bianconi², Robert Vautard³, K. Wyatt Appel⁹, Michael D. Moran¹⁴, Christian Hogrefe⁹, Bertrand Bessagnet⁶, Jørgen Brandt¹⁶, Jesper H. Christensen¹⁶, Charles Chemel^{11,12}, Isabelle Coll¹⁵, Hugo Denier van der Gon²⁰, Joana Ferreira⁸, Renate Forkel¹⁰, Xavier V. Francis¹², George Grell¹⁸, Paola Grossi², Ayo B. Hansen²², Amela Jeričević¹⁷, Lukša Kraljević¹⁷, Ana Isabel Miranda⁸, Uarporn Nopmongkol⁴, Guido Pirovano^{6,7}, Marje Prank¹⁹, Angelo Riccio²¹, Karine N. Sartelet⁵, Martijn Schaap²⁰, Jeremy D. Silver¹⁶, Ranjeet S. Sokhi¹², Julius Vira¹⁹, Johannes Werhahn¹⁰, Ralf Wolke¹³, Greg Yarwood⁴, Junhua Zhang¹⁴, S.Trivikrama Rao⁹, Stefano Galmarini¹*

¹Joint Research Centre, European Commission, ISPRA, Italy;

²Enviroware srl, via Dante 142, 20863 Concorezzo (MB), Italy

³IPSL/LSCE Laboratoire CEA/CNRS/UVSQ

⁴Environ International Corporation, Novato CA, USA

⁵CEREA, Joint Laboratory Ecole des Ponts ParisTech/ EDF R & D, Université Paris-Est, France

⁶Ineris, Parc Technologique Halatte, France

⁷Ricerca sistema energetico (RSE), Italy

⁸CESAM & Department of Environment and Planning, University of Aveiro, Aveiro, Portugal

⁹Atmospheric Modelling and Analysis Division, Environmental Protection Agency, NC, USA

¹⁰IMK-IFU, Institute for Meteorology and Climate Research-Atmospheric Environmental Division, Germany

¹¹National Centre for Atmospheric Science (NCAS), University of Hertfordshire, Hatfield, UK

¹²Centre for Atmospheric & Instrumentation Research (CAIR), University of Hertfordshire, Hatfield, UK

¹³Leibniz Institute for Tropospheric Research, Leipzig, Germany

¹⁴Air Quality Research Division, Science and Technology Branch, Environment Canada, Toronto, Canada

¹⁵IPSL/LISA UMR CNRS 7583, Université Paris Est Créteil et Université Paris Diderot

¹⁶Department of Atmospheric Environment, National Environmental Research Institute, Aarhus University, Denmark

¹⁷Meteorological and Hydrological Service, Grič 3, Zagreb, Croatia

¹⁸CIRES-NOAA/ESRL/GSD National Oceanic and Atmospheric Administration Environmental Systems Research Laboratory Global Systems Division Boulder, Colorado USA

¹⁹Finnish Meteorological Institute, Helsinki, Finland

²⁰Netherlands Organization for Applied Scientific Research (TNO), Utrecht, The Netherlands

²¹Department of Applied Science, University of Naples "Parthenope", Naples, Italy

²²Department of Environmental Science, Faculty of Science and Technology, Aarhus University, Denmark

* Author for correspondence: E.Solazzo. Email: ef시오.solazzo@jrc.ec.europa.eu

40 **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
42 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
44 allows for statistical and ensemble analyses to be performed by each group. The estimated ground-
level ozone mixing ratios from the models are collectively examined in an ensemble fashion and
46 evaluated against a large set of observations from both continents. The scale of the exercise is
unprecedented and offers a unique opportunity to investigate methodologies for generating skilful
48 ensembles of regional air quality models outputs. Despite the remarkable progress of ensemble air
quality modelling over the past decade, there are still outstanding questions regarding this
50 technique. Among them, what is the best and most beneficial way to build an ensemble of
members? And how should the optimum size of the ensemble be determined in order to capture data
52 variability as well as keeping the error low? These questions are addressed here by looking at
optimal ensemble size and quality of the members. The analysis carried out is based on systematic
54 minimization of the model error and is important for performing diagnostic/probabilistic model
evaluation. It is shown that the most commonly used multi-model approach, namely the average
56 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 skill of
58 the individual members, but may require the inclusion of low ranking skill-score members. A
clustering methodology is applied to discern among members and to build a skilful ensemble based
60 on model association and data clustering, which makes no use of priori knowledge of model skill.
Results show that, while the methodology needs further refinement, by optimally selecting the
62 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.

64

Keywords: AQMEII, clustering, error minimisation, multi-model ensemble, ozone, model
66 evaluation

68

70

72

74 1. Introduction

Regional air quality (AQ) models have undergone considerable development over the past three
76 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
78 (e.g., Holloway et al., 2003; Rao et al., 2006; Jacob and Winner 2009). Regional AQ models are
now widely used for supporting emissions control policy formulation, testing the efficacy of
80 abatement strategies, performing real-time AQ forecasts, and evaluating integrated monitoring
strategies. Moreover, ozone estimates have been used in assimilation schemes to provide further
82 information on meteorological variables such as wind speed (e.g., Eskes, 2003). The combination of
outcomes predicted by several models (regardless of their field of application), in what is typically
84 defined as ensemble modelling, has been shown to enhance skill when compared against an
individual model realisation (e.g., Delle Monache and Stull, 2003; Galmarini et al. 2004; van Loon
86 et al., 2007). Although ensemble modelling is well established (both from the applied and
theoretical perspectives) and is now routinely used in weather forecasting, it is only during the last
88 decade that a growing number of AQ modelling communities have joined their model outputs in
multi-model (MM) combinations (Galmarini et al., 2001; Carmichael et al., 2003; Rao et al., 2011).
90 The advantages of ensemble modelling versus an individual model are at least twofold: (i) the mean
(or median) of the ensemble is, in effect, a new model that is expected to lower the error of the
92 individual members due to mutual cancellation of errors; and (ii) the spread of the ensemble
represents a measure of the variability of the model predictions (Galmarini et al., 2004; Mallet and
94 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
96 information from many sources and synthetically assessing their variability. In particular, the mean
and median offer enhanced performance, on average, compared with single-model (SM) realisations
98 (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), including by
100 varying some internal parameters for multiple simulations with an SM, by using different input data
(e.g., emissions) for multiple simulations with an SM, or by applying several different models to the
102 same scenario. This latter approach is the main focus of the Air Quality Model Evaluation
International Initiative (AQMEII) (Rao et al., 2011), an international project aimed at joining the
104 knowledge and experiences of AQ modelling groups in Europe and North America. Within
AQMEII, standardised modelling outputs have been shared on the web-distributed ENSEMBLE
106 system, which allows statistical and ensemble analyses to be performed by multiple groups

(Bianconi et al., 2004; Galmarini et al., 2012). A joint exercise was launched for European and
108 North American AQ modelling communities to use their own regional AQ models to simulate the
entire year 2006 for the continents of Europe and North America, retrospectively. Outputs from
110 numerous regional AQ models have been submitted in the form of both gridded, hourly
concentration fields and values at specific locations, allowing for direct comparison with air quality
112 measurements available from monitoring networks across North America and Europe (see Rao et
al., 2011 for additional details). This type of evaluation, with large temporal and spatial scales, is
114 essential to assess model performance and identify model deficiencies (Dennis et al., 2010; Rao et
al., 2011).

116

In this study, we analyse ozone mixing ratios provided by simulations from eleven state-of-the-art
118 regional AQ models run by eighteen independent groups from North America (NA) and Europe
(EU) (while a companion study is devoted to the examination of particulate matter, Solazzo et al.,
120 2012). Model predictions have been made available, along with observational data, to the
ENSEMBLE system. The ability of the ensemble mean and median to reduce the error and bias of
122 SM outputs is examined, and conclusions regarding the size of the ensemble and its quality are
made. The level of repetition provided by each individual model to the ensemble is quantified by
124 applying a clustering analysis to examine whether the improvement in error using the mean or
median of the model ensemble is due to the increased ensemble size, or if information carried by
126 each model contributes to the MM superiority.

128 The main objective of this study is to assess the statistical properties of the ensemble of models in
relation to the individual model realisations for a range of air quality cases. Each model has
130 imperfections, and it is beyond the scope of this analysis to identify the causes of model bias for
each ensemble member. Several other papers examining the performance of the individual model
132 simulation are available in the AQMEII special issue.

134 **2. Models and data**

2.1 Experimental set up

136 In order to carry out a comprehensive evaluation of the participating regional-scale AQ models, the
model estimates are compared to observations for the year of 2006, with the various modelling
138 groups providing hourly ozone mixing ratios and concentrations of other compounds. Surface
concentrations were then interpolated to the monitoring locations for the purposes of model
140 evaluation.

142 *2.2 Participating models*

144 Table 1 summarises the meteorological and AQ models participating in the AQMEII
intercomparison exercise and providing ozone mixing ratios at European or North American
receptor sites, or both. In some cases the same model is used with a different configuration by
146 different research groups (or in a few cases by the same research group). In total, eleven groups
submitted ozone predictions for EU and seven submitted ozone predictions for NA. No a-priori
148 screening on the worst performing models has been performed on the participating members;
however, it is assumed that the models have at least previously gone through an operational model
150 evaluation, as defined in Dennis et al. (2010).

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

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

- 160 - CMAQ (Byun and Schere, 2006)
- CAMx (ENVIRON, 2010)
- 162 - CHIMERE (Schmidt et al., 2001; Bessagnet et al., 2004)
- MUSCAT (Wolke et al., 2004; Renner and Wolke, 2010)
- 164 - DEHM (Brandt et al., 2007)
- POLYPHEMUS (Mallet et al. 2007; Sartelet et al. 2012)
- 166 - EUROS (Schaap et al., 2008)
- SILAM (Sofiev et al., 2006)
- 168 - AURAMS (Gong et al., 2006; Smyth et al., 2009)
- EMEP (Simpson et al., 2003; Jeričević et al., 2010)
- 170 - WRF/Chem (<http://www.acd.ucar.edu/wrf-chem/>)

The combination of meteorological and chemical transport models varies for each group (with the
172 only exception being the WRF model with the WRF/Chem model, which was used twice for EU),
thus allowing analysis of a diversified set of model output, which is necessary to sample the
174 spectrum of uncertainty within an ensemble.

176 Emissions and chemical boundary conditions used by the various AQMEII groups are summarised
in Table 1. AQMEII provided a set of time-varying gridded emissions (referred to as “standard”
178 emissions) for each continent, focusing on the evaluation of the AQ and meteorological models.
The EU standard emissions were prepared by TNO (Netherlands Organization for Applied
180 Scientific Research), which provided a gridded emissions database for the years 2005 and 2006.
This dataset was partly developed in the framework of the European MACC project
182 (<http://www.gmes-atmosphere.eu/>), and is an update of an earlier TNO emissions database prepared
for the GEMS project (<http://gems.ecmwf.int>). This inventory does not include biogenic emissions,
184 and therefore different approaches were used by different groups to provide biogenic emissions, as
summarised in Table 1. The NA standard emissions are based on the 2005 U.S. National Emissions
186 Inventory (NEI), 2006 Canadian national inventory, and 1999 Mexican BRAVO inventory.
Biogenic emissions were provided by the BEISv3.14 model, while daily estimates of fire emissions
188 were provided by the HMS fire detection and SMARTFIRE system (year 2006). In-stack emissions
measurements for many U.S. power plants were provided by Continuous Emissions Monitoring
190 data for the year 2006. Full details regarding the standard emissions used for EU and NA are given
in Pouliot et al. (2012). The standard emissions were used by the vast majority of the participating
192 AQMEII groups (Table 1). Model results generated with other emissions inventories have also been
submitted, however, which provides a useful comparison in interpreting the results of model-
194 estimated ozone mixing ratios.

196 AQMEII also made available a set of chemical concentrations at the lateral boundaries of the EU
and NA domains, which were provided by the Global and regional Earth-system Monitoring using
198 satellite and in-situ data (GEMS) re-analysis product provided by European Centre for Medium-
range Weather Forecast (see Schere et al., 2012 for more details). Other boundary conditions for
200 ozone used by several AQMEII modeling groups were based on satellite measurements assimilated
within the Integrated Forecast System (IFS). LMDZ-INCA, which couples the general circulation
202 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
204 (NA simulations), with another CHIMERE model simulation using the standard AQMEII boundary
conditions (Table 1).

206

2.3 Observational data for ozone

208 The European and North American continental areas have each been divided into four sub-regions
(EU1 to EU4 and NA1 to NA4). Figure 1 displays the sub-regions for both continents, the locations
210 of the ozone receptors that have been used, and contours of “standard” anthropogenic NO_x
emissions averaged over the summer months of June-July-August (JJA) of 2006. Only rural
212 receptors below an altitude of 1000 m have been examined, with at least 75% annual data
availability. The choice of analysing only rural receptors is dictated by the need to provide
214 comparison with spatial scales consistent with the model resolution (see, e.g., Vautard et al., 2009).
Moreover, ozone measured by monitoring stations in urban areas is more sensitive to reactions with
216 NO_x, which may reduce ozone production.

218 The selection of the sub-regions is based on emissions, climate, and altitudinal aspects, as well as
practical constraints (data coverage, computational time). The four EU sub-regions are similar to
220 those in the analyses of meteorological forcing (Vautard et al., 2012) and particulate matter
(Solazzo et al., 2012) for AQMEII. Sub-region EU1, consisting of the British Isles, France, and
222 northern Spain, was selected for its mid-latitude, mixed maritime-continental climate and large
conurbations (London, Paris). Sub-region EU2, consisting of Central Europe, has a continental
224 climate with marked seasonality, many large cities, and areas with large emissions. Sub-region
EU3, consisting of the Po River Valley up to the Alpine area of Italy and south-eastern France has a
226 mixed climate, generally poor air quality, and is influenced by the Alpine barrier. The Southern
European domain covers the Mediterranean area (southern Italy, the east coast of Spain, and
228 Greece), with typical Mediterranean climate and large cities (e.g., Barcelona, Rome). The number
of rural receptors for the EU sub-regions is 201, 225, 77, and 140, respectively.

230
For NA, the number of rural receptors in the four sub-regions varies between 134 and 150. The NA
232 sub-regions are broadly derived from previous studies (e.g., Eder et al., 1993), and consider the
NO_x emissions intensity, with the additional constraint of a uniform number of receptors. Sub-
234 region NA1, consisting of the western portion of the United States and south-western Canada, has
high emissions along the coast of California, smaller emissions toward the interior of the continent,
236 a high amount of solar radiation, low relative humidity, and some large cities with poor air quality
(e.g., Los Angeles). Sub-region NA2 consists of the U.S. Plains states to the east of the Rocky
238 Mountains and is characterised by a continental climate in the north and a hot, humid climate in the
south, with a number of large cities with poor air quality (e.g., Houston, Dallas). Sub-region NA3,
240 consisting of northeastern NA including parts of south-central Canada, has a marked seasonal cycle,
most of the Great Lakes, some of the highest emissions areas in NA, and many large cities (e.g.,

242 New York City, Philadelphia, Toronto, Montreal). Finally, sub-region NA4, consisting of the
southeast United States, has high emissions and strong solar radiation.

244

Ozone data for EU were derived from hourly data collected by the AirBase and EMEP (European
246 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
248 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,
250 <http://java.epa.gov/castnet/>) networks in the United States and the NAPS (National Air Pollution
Surveillance, <http://www.ec.gc.ca/rnsps-naps/>) network in Canada. A total of 1445 stations are
252 available, more than half with a percentage of data validity higher than 80% (many U.S. ozone
stations only operate from May to October).

254

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

256 3.1. Operational SM and ensemble statistics for the continental-wide domains

van Loon et al. (2007) showed that the ensemble mean ozone daily cycle over EU, obtained by
258 averaging over all monitoring stations for the entire year of 2001, agrees almost perfectly with the
observations, and better than any individual member of the ensemble. This result provides
260 substantial evidence of the enhanced skill of MM predictions versus the individual SM predictions.
Such a result, though, while encouraging, poses some additional questions, such as what is the role
262 of repeated averaging (in time and space) in smoothing out peaks and reducing variability, and
whether any ensemble combination will show additional skill relative to SMs. For example,
264 Galmarini and Potempski (2004) showed that for the ETEX-1 case study the MM did not offer
significantly superior skill (and performed less well than for a long-term AQ case due to the
266 transient nature of the short-term ETEX tracer release). They thus concluded that in the absence of a
method for pre-selecting or discriminating between ensemble members, the MM improved
268 performance might be just coincidental and dependent on the ‘lucky shot’ of having the right
collection of models around the measured data.

270

Figure 2 presents annual frequency distributions of ozone mixing ratios averaged across the
272 receptor set that were estimated by the AQMEII SM and MM for EU (Fig. 2a) and NA (Fig. 2b). A
box-and-whisker representation has been used to show the frequency distribution, where the
274 rectangle represents the inter-quantile range (25th to 75th percentile), the small square identifies the
mean, the continuous horizontal line inside the rectangle identifies the median, the crosses identify

276 the 1st and 99th percentiles, and the whiskers extend between the minimum and maximum values.
The measured frequency distribution is also shown in each row. The top row displays the
278 distribution of hourly values (i.e., each bar is the distribution over 8760 receptor-averaged hourly
values), the middle row is the daily average distribution (over 365 receptor-averaged daily values),
280 and the bottom row is the mean diurnal range (each bar reflects the distribution over 24 receptor-
averaged hourly values, in which the same hours are averaged for each day of the year). Depending
282 on the averaging period, ozone mixing ratios are reduced by a factor of two for both continents,
which results in a dramatic reduction of the spread (e.g., min and max values are within the inter-
284 quantile ranges for the diurnal cycle) and a clustering of the diurnal time series, which in turn
results in improved statistical agreement. Thus, averaging over extended areas (continent) and
286 periods (year) has a dramatic effect in reducing the spread of the data. Note that in order to
maintain model anonymity each participating model has been assigned a random model number
288 (Mod 1 to 11 for EU and Mod 12 to 18 for NA) that do not correspond to the order of models
presented in Table 1.

290

The ability of the MM ensemble to sample measurement uncertainty for both continents is analysed
292 by means of the rank histograms presented in Fig. 3, which are a measure of the ensemble reliability
(Talagrand et al., 1998; Joliffe and Stephenson, 2003). The rank histogram is a widely adopted
294 diagnostic tool to evaluate the spread of the members of an ensemble. In a rank histogram, the
population of the k -th rank is the fraction of time that observations fall between the sorted members
296 $k-1$ and k , and the number of ranks or bins is one greater than the number of ensemble members.
Ideally, the frequency for each bin should be the same, meaning that the ozone estimate from each
298 ensemble member is as probable as from any other member, and that observations have an equal
probability of belonging to any bin (Hamill, 2000). In such a case the observations and the
300 ensemble members are selected from the same probability distribution, and the probability of an
observation falling into a particular bin is the same for all bins. In Fig. 3, spatially-averaged hourly
302 ozone data from the full year are used. For EU (Fig. 3a), the bin populations are rather uniform for
the first ten bins (between 6% and 11%), with bins 11 and 12 having a frequency of ~18% each.
304 This distribution indicates the ensemble mean has difficulty simulating high hourly mixing ratios,
which indicates a negative bias in the ensemble mean (i.e., underestimation). For NA the
306 intermediate bins of the rank histogram are more populated than the side bins (Fig. 3b), indicating
the possible presence of outlying members.

308

It is not clear whether the deviation from a uniform distribution for both EU and NA is due to
310 chance (for case in which ensemble members and observations are truly selected from the same
distribution) or if there is a compensating effect over such large domains and long time scales.
312 These aspects will be further examined in Section 3.3.

314 *3.2 Sub-regional SM and MM ensemble analyses*

Regional AQ models are often used on limited spatial and temporal scales (e.g., a few months or a
316 season over several hundred kilometres: Camalier et al., 2007; Bloomer et al., 2009; Boynard et al.,
2011; Hogrefe et al., 2011), for which mutual cancellation of model errors might not be as effective
318 as in the case of continental and yearly scales, as discussed for the results of Fig. 2. The analyses
presented in this section focus on the spatial variability of ozone mixing ratio statistics in four
320 distinct sub-regions of each of the continental domains of Fig. 1, examining the temporal variability
for the summer months JJA, when the ozone mixing ratios are typically the highest and are of most
322 concern for public health. Analysis and evaluation of SM performance over the whole year are
presented in companion papers in this Special Issue.

324
Sub-regional ozone diurnal cycles are shown in Fig. 4a (EU) and Fig. 4b (NA), including ensemble
326 mean and median (hourly data have been used for the analysis). Examining the observed
summertime diurnal cycles for the four EU sub-regions (Fig. 4a), it is evident that there is
328 considerable intra-continental variability of the daily ozone maximum, with the northern Italian and
Mediterranean sub-regions (EU3 and EU4) reaching 60 ppb or more whereas peak daily ozone
330 mixing ratios of ~45-50 ppb occur in the other two EU sub-regions. For sub-regions EU1, EU2, and
EU3 the daily maximum occurs at 1700 local time (LT), while the daily maximum occurs two hours
332 earlier in the EU4 sub-region due to the higher average insolation. Daily minimum ozone values
occur between 0700 LT and 0800 LT, and range between 20 and 30 ppb, with the Mediterranean
334 area (EU4) having the highest minimum due to the relative abundance of biogenic emissions (see,
e.g., Sartelet et al., 2012). The daily maximum ozone values for NA are of the same general
336 magnitude as EU, between 45 and 55 ppb for sub-regions NA1, NA2, and NA4, while only
reaching ~35 ppb for sub-region NA3 (the north-eastern NA region) due to the inclusion of some
338 remote monitoring stations from the Canadian NAPS network. Daily maximum values occur at
1600 LT for all four NA sub-regions. Daily minimum values typically occur at 0700 LT, and range
340 between 20 to 25 ppb for sub-regions NA1, NA2, and NA3 and less than 20 ppb for sub-region
NA4. This latter sub-region (south-eastern United States) exhibits a steep rise of ozone mixing
342 ratios in the late morning that is indicative of strong daytime photochemistry in this region.

344 The majority of individual models (indicated by the thin lines in Fig. 4) exhibit highly region-
346 dependant behaviour, although some common patterns are present. Models for EU have a
348 predominant tendency to underestimate (in some cases significantly) the peak daily mixing ratio
350 and/or displace the time of the peak mixing ratio, as well as to overestimate nighttime mixing ratios,
with the exception of sub-region EU2 (central Europe), which may be due to the strong daily
temperature gradient in this region. Nighttime overestimation is known to occur in some models
due to difficulties in dealing with stable conditions (e.g., Smyth et al., 2009; Herwehe et al., 201)

352 Model results for the NA sub-regions exhibit a lower spread throughout the diurnal cycle (Fig. 4b),
with the exception of one outlying model for sub-regions NA1, NA2, and NA3, which is
354 consistently biased low, especially at night. However, the majority of the models exhibited
356 nighttime overestimation to varying degrees, indicating that most of the AQ models have at least
some difficulty dealing with stable conditions despite the variety of vertical mixing schemes
implemented by the models. The case of the south-east U.S. sub-region (NA4), on the other hand,
358 with consistent model overestimation throughout the diurnal cycle, clearly requires a dedicated
investigation that is beyond the scope of this study.

360

Reasons for individual model biases are detailed in other studies of this special issue dedicated to
362 AQMEII and are not covered here. Collectively, though, the results of those studies have pointed to
a number of factors, such as: (a) the biogenic emissions adopted by each model in EU (Brandt et al.,
364 2012; Sartelet et al., 2012), confirmed by examining the performance of the CHIMERE model with
MEGAN biogenic emissions, which is the best performing SM for all EU sub-regions; (b) the
366 meteorological driver (Vautard et al., 2012), and the impact of overestimated wind speed on the
dispersion of primary pollutants (Solazzo et al., 2012), especially in EU; and (c) the lateral
368 boundary conditions used for ozone, especially for winter-time concentration in NA (Schere et al.,
2012; Appel et al., 2012; Nopmongcol et al., 2012).

370

The MM ensemble mean and median generally underestimate the amplitude of the ozone diurnal
372 cycle in EU despite one outlying model demonstrating a large positive bias. By contrast, the MM
mean and median accurately follow the measured ozone diurnal cycle for sub-regions NA1, NA2,
374 and NA3 (while largely overestimating for the NA4 sub-region) due to the mutual compensation of
a low-biased outlier and the tendency of the other ensemble members to overestimate ozone. It
376 should be noted that the mean and median curves overlapping is a consequence of the repeated data

378 averaging (both spatially and temporally), which has smoothed out the peaks of the distribution, as previously shown in Fig 2.

380 Figure 5 presents the error statistics for EU (Fig. 5a) and NA (Fig. 5b), in the form of a “soccer-goal” plot (Appel et al., 2011). NMSE versus NMB scores (see Appendix A for definition) are
382 reported for each individual model, together with scores for the ensemble mean and median, for each of the four sub-regions. Points falling within the dotted lines indicate model performance
384 within the criteria set by Russell and Dennis (2000) for ozone (bias within $\pm 15\%$ and error within $\pm 30\%$). For EU the majority of points lie in the left region of the soccer goal, indicating
386 underestimation, with the exception of Mod1, which substantially overestimates the ozone mixing ratio for all sub-regions. The ensemble mean and median scores for all sub-regions fall within the
388 15% box, and therefore comply with the performance criteria for ozone. For NA model results are well within the 15% box (mainly overestimated), the exception being the NA4 sub-region, where
390 three models show an overestimation between 15% and 20%. The ensemble mean and median for NA are approximately identical, as already noted for the diurnal cycle (see Fig. 4b).

392

3.3. Reliability of the Multi-model Ensemble

394 Biased rank histograms for all sub-regions of the two continents have a sloped shape (Fig. 6). Analysis is based on hourly data for the period JJA. The histograms for EU sub-regions 1 and 2
396 (Fig. 6a) have the most populated bins towards the end of the ranks, indicating model underestimation. The EU4 sub-region has empty or nearly empty initial and final bins, indicating an
398 excess of model variability. The histogram for the entire EU domain is fairly flat, a result of compensating biases between sub-regions EU1 and EU2 and sub-regions EU3 and EU4 (see Fig.
400 3a). As discussed at the beginning of section 3.2, when using long averaging periods and large spatial scales, seasonal and intra-continental variability can be hidden by the averaging and
402 compensating errors. Large biases are also present for the NA sub-regions (Fig 6b), with overestimation (left bins most populated) for all sub-regions, as seen in Fig. 5b. The spread also
404 suffers from deficiencies of the ensemble in all cases, with excess of spread for sub-region NA1 (middle bins more populated) or insufficient spread, such as in sub-regions NA2, NA3, and NA4
406 (side bins more populated). This latter case is typically due to not having captured all sources of error properly (Vautard et al., 2009), which may be due too many members of the ensemble using
408 the same meteorological drivers and/or emissions. Comparing the histograms in Fig. 6b for the entire NA domain for JJA and that of Fig. 3b for the entire NA domain for the full year highlights
410 that for the full year the bins were more uniform, with a tendency to form a “bell” shape, whereas

for JJA the distribution is drastically biased and bin populations are uneven. This is probably due to
412 the 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.

414

4. Multi-model analysis: selected vs. unselected model ensembles

416 4.1 Ensemble size

In this section we evaluate whether an ensemble built with a subset of individual models can
418 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 sub-regions of EU and NA
420 separately, using hourly ozone data for the period JJA.

422 Consider the distribution of some statistical measures (RMSE, PCC, MB, MGE, defined in
Appendix A) of the mean of all possible combinations of available ensemble members n (n is 11 for
424 EU and 7 for NA). The number of combinations of any k members is $\binom{n}{k}_{k=2,\dots,n-1}$. For example,

there are as many as 462 combinations of 5 models for EU, and 35 combinations of 3 models for
426 NA. The results of the statistical analysis are presented in Fig. 7. The continuous lines on each plot
represent the mean and median of the distribution of any k -model combinations. MM mean and
428 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
430 confirms the common practice to average over all available members to obtain enhanced
performance with respect to SM realisations. For MB, the mean trend is flat due to the quasi-
432 symmetric error fluctuations about the mean value for NA. Mean RMSE curves decrease steeply
from two to four models for all sub-regions except the sub-region NA4. A further striking feature is
434 that the best SM has similar (EU sub-regions EU1 and EU2; NA sub-regions NA1 and NA3) or
lower (EU sub-regions EU3 and EU4; NA sub-regions NA2 and NA4) RMSE than the ensemble
436 mean with all members. This is most probably due to having included members with large variances
in the ensemble (Potempski and Galmarini, 2009).

438

Analysis of mean RMSE for EU sub-regions (Fig. 7a), for which a large set of members is
440 available, shows a plateau is reached 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 mean RMSE is
442 negligible. Investigating the maximum RMSE (i.e., upper error bound), however, gives the result
 $\max_k RMSE(k) > \max_k RMSE(k+1)$. Thus, the mean of ensembles with a large number of members

444 has the property of reducing the maximum error. For example, sub-region EU3 has a large error
span for RMSE, between 2.5 ppb and 15 ppb for $k = 2$, which reduces to between 4 ppb and 7 ppb
446 for $k = 10$ (Fig. 7a). A similar trend is seen for PCC (all sub-regions), with a monotonic
improvement in the minimum PCC values with increasing k .

448
Values of minimum RMSE (lower bound) exhibit a more complex behaviour. A minimum, among
450 all combinations, systematically emerges for ensembles with a number of members $k < n$. Similarly,
a maximum of PCC is achieved by combinations of a subset of members. This result suggests that
452 ensembles of a few members systematically outperform the ensemble of all members. In addition,
adding new members to such an optimal ensemble (thus moving towards a higher value of k)
454 deteriorates the quality of the ensemble, as the minimum RMSE increases and the maximum PCC
decreases (Fig. 7).

456

4.2 Ensemble combinations of minimum RMSE

458 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-
460 best SM in terms of RMSE). The SM RMSE ranking is defined by domain, and individual models
may not have the same SM RMSE ranking over the different sub-regions.

462

An important point worth noting is that the RMSE ranking shows that the optimal ensemble is in
464 some cases achieved by the MM ensemble containing low-ranking members, which suggests that all
members should be considered to build a skilful ensemble. Therefore, an ensemble of top-ranking
466 models can be worse than an ensemble of top-ranking and low-ranking models: that is, outliers may
need to be included in the ensemble to obtain the best performance.

468

It can be argued that large ensembles are needed to capture extreme events (e.g., high mixing
470 ratios). Figure 8 presents a scatter plot of 1-hour daily maximum ozone mixing ratios for the EU
sub-regions (analysis for NA sub-regions with fewer individual model members produced similar
472 results and is not shown). The x -axis represents the 1-hour maximum of the ensemble of all
available members, while the y -axis represents the 1-hour maximum of the ensemble of the selected
474 members with minimum RMSE (bold-face combinations of Table 2). The data distribution along
the diagonal line for each region shows that ensembles of selected models and full ensembles have
476 the same probability to capture the extreme events. In particular, for the EU1 and EU3 sub-regions,
the maximum predicted mixing ratio is higher with the small ensemble. This is because a poorly

478 performing SM added to an ensemble can improve RMSE and compensating biases can reduce
overall bias.

480

As an example, consider the case presented in Fig. 9, in which ozone mixing ratios of observations
482 (Fig. 9a), the ensemble of ranked models 1 and 5, (Fig. 9b), ranked model 2 (Fig. 9c), and ranked
model 11 (Fig. 9d) are displayed at the receptor sites. Note that the ranked combination 1-5-11
484 represents a minimum RMSE for the EU1 sub-region (Table 2). An interesting question to pose is
why the lowest-ranked model (11) improves the ensemble more than a highly ranked model.
486 Examining the receptor sites in the British Isles and France (Domain 1 of Fig. 1a), the MM mean of
Fig. 9b clearly underestimates the observations in the south of France. When the 11th-ranked model
488 (Fig. 9d) is added to the ensemble in Fig. 9b, compensating errors result in lower RMSE than the
combination with the 2nd-best model in Fig. 9c. This is because the 2nd-best model has a
490 performance very similar to the best performing model (which is already included in the ensemble),
and thus brings no new information to the existing ensemble, whereas the 11th-ranked model, while
492 performing poorly across the entire domain, matched the high mixing ratios in southern France (i.e.,
the only place where the higher-ranked models performed worse). Since RMSE weights large errors
494 more heavily, including the 11th-ranked model results in less error at a greater number of receptor
sites than when the 2nd-ranked model is included instead.

496

Statistical results and box-and-whisker plots for the full ensemble and for the selected member
498 ensemble for each sub-region are presented in Table 3 and Fig. 10, respectively. RMSE is, as
expected, lower for the selected-member ensemble for all sub-regions. PCC varies only slightly,
500 indicating that the association between observations and MM ensemble is not strictly related to
model error. The minimum RMSE combinations also improve the estimation of the modelled
502 spread (the standard deviation of the MM ensemble, σ) compared to measured spread for almost all
sub-regions (Table 3), and especially for the EU sub-regions. Therefore, reducing the number of
504 members does not degrade the ensemble variability, but instead actually compares better to the
spread of the observations. This is most likely due to the reduced variability induced by members
506 sharing similar emissions and boundary conditions. Figure 10 presents a graphical depiction of how
the selected-member ensemble compares against the full-member ensemble in terms of spread,
508 maximum and minimum, and percentile distribution. The improvement to spread of the selected-
member ensemble mean is most visible.

510

512

5. Reduction of data complexity: a clustering approach

514 Results discussed in the previous section have shown that a skilful ensemble is built with an optimal
516 number of members and often includes low-ranking skill-score members as well. In order to
discern which members should be included in the ensemble, a method for clustering highly
associated models and then discarding redundant information was developed using the PCC as the
518 determining metric (we note that PCC is independent of model bias; therefore, the analysis would
be the same for unbiased models). The most representative models of each cluster, chosen based on
520 a distance metric, are then used to generate a reduced or selected-member ensemble. In this way,
the information that each member provides to the ensemble is “unique” to the greatest possible
522 degree.

524 The Euclidean distance metric has been used to calculate the distance between the PCC of any two
models and between clusters. The points that are farthest apart are identified and used as the initial
526 cluster centres. Then, the other models are allocated to the closest centre by the Euclidean distance
from each centre. Results for this procedure are presented in Fig. 11 (EU) and Fig. 12 (NA) as
528 hierarchical diagrams called dendrograms. The “height” of each inverted U-shaped line on the x -
axis represents the distance between the two clusters being connected. Independent clusters are
530 identified by different colours. Sensitivity analysis of other distance metrics (not shown) has found
that the clustering of models is independent from the metric used to calculate the distance, thus
532 leaving the group associations unaltered. However, while the distance itself may change, it does not
affect the results of this study. The y -axis of Figs. 11 and 12 identifies the models by their number
534 and RMSE ranking (discussed in Section 4.2). The ranking information allows tracking of each
model’s position and whether aggregation results from differences between the models themselves
536 (e.g., AQ model, meteorological drivers, emissions) or if the model’s performance itself (e.g.,
RMSE) has an influence.

538

For EU (Fig. 11), the maximum PCC distance (degree of model disassociation) varies between 0.12
540 (sub-region EU4) to 0.28 (sub-region EU2). By contrast, analysis of NA sub-regions (Fig. 12)
shows the maximum distance is 0.08 for all sub-regions, with the exception of sub-region NA2
542 (~ 0.03). Association between models is thus stronger for NA, indicating a lower degree of
independence. This is likely due to four out of the seven models using the same meteorological
544 driver for NA, and six models using the same emissions. For EU it is possible to isolate two
repeating groups of models whose PCC distances are very small: Mod6 and Mod7, and Mod11 and
546 Mod3. Models of the former group are essentially the same, as they share both AQ and

meteorological models, and used the same emissions and boundary conditions. They also have
548 similar RMSE rankings. Mod11 and Mod3 differ in the AQ model used, but used the same
meteorological model (MM5) and anthropogenic emissions. The NA cluster analysis, with fewer
550 members, shows repeated association of several pairs of models: Mod15 and16 (same
meteorological driver, anthropogenic emissions, and boundary conditions); Mod13 and Mod17
552 (same AQ model, only different boundary conditions), and Mod14 and Mod18 (same
meteorological driver). Mod12 is associated with Mod14 and Mod18, with the exception of the
554 NA3 sub-region.

556 In order to find an optimal set of clusters, a threshold at which models are said to be independent
(imagine cutting the dendrograms vertically) is defined. The selection of the cutting height is in part
558 arbitrary. The common practice suggests cutting the dendrogram at the height where the distance
from the next clustered groups is relatively large, and the retained number of clusters is small
560 compared to the original number of models (Riccio et al., 2012). Members of the ensemble
generated with a higher threshold are more distant and therefore more independent. The cluster
562 representatives and selected-member ensembles are summarised in Table 4 for both continents and
for different PCC distances. For clusters composed by only two members and with symmetric
564 structures (same mutual distance among all members, such as the third cluster of the EU2 sub-
region in Fig. 11b), it was not possible to identify a model whose distance from the centre of the
566 cluster was a minimum in terms of RMSE. In these cases, more than one model is selected to
represent the cluster.

568
The number of independent members varies between 3 and 6 for EU and between 2 and 4 for NA
570 (this difference is probably due to the smaller number of models for NA). It is interesting to note
that the number of independent clusters matches the number of models needed to generate the MM
572 ensembles with minimum RMSE in Fig. 7 for both continents. The two methods are in fact
independent, as the clustering analysis makes no use of observational data (only model-to-model
574 PCC is in fact used in the cluster analysis). Looking at the minimum RSME combination in Table 2,
it can be seen that the ensembles of minimum RMSE have two or more members belonging to the
576 same cluster, and that for the NA4 sub-region all members are from the same cluster. This is a result
of too few independent members due to models sharing of boundary conditions, meteorology, and
578 emissions.

580 The RMSE of MM ensembles in Table 4 were compared to the RMSE curves discussed in Fig. 7 by
connecting, for any number of models, the minimum (thick lines) and maximum (dotted lines)
582 RMSE values. The results are presented in Fig. 13. The short lines in Fig. 13 represent the RMSE of
combinations from Table 4 (obtained with the clustering technique) and are reported along with the
584 ranked combination. In the case of clusters with only two members (symmetric clusters), it was not
possible to identify the representative member, and therefore both members have been retained for
586 the analysis. Comparing the position of the cluster's combination against the RMSE of the full
member ensemble in Fig. 13 allows one to infer whether the new methodology is able to produce
588 reduced ensembles that are more skilful than the full ensemble mean. Note that combinations of
independent models have, in most cases, lower RMSE than the full ensemble, and that for all sub-
590 regions there are ensembles that clearly outperform the full ensemble. 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 sub-regions EU1, EU2, EU3, and
592 EU4, respectively, have a lower RMSE than the mean of all ensemble members and are close to the
minimum curve. Conversely, there are situations in which the ambiguous definition of
594 representative cluster leads to high-RMSE MM combinations, as for the four-member combination
of the EU4 sub-region (1-2-4/5-11) and NA1 sub-region (2-4-5). Further work is needed to remove
596 such ambiguity.

598 **6. Conclusions**

This study collectively evaluates and analyses the performance of eleven regional AQ models and
their ensembles in the context of the AQMEII inter-comparison exercise. The scale of the exercise
600 is unprecedented, with two continent-wide domains being modelled for a full year. The focus of
this study was on the *collective* analysis of surface ozone mixing ratios, rather than on inter-
602 comparing metrics for each individual model. The study began with an analysis of ozone time series
for sub-regions of EU and NA, followed by an interpretation of the uncertainties of the individual
604 models and ensemble. Analysis of model error in each sub-region demonstrates that most of the
error in the models is introduced by bias from emissions, boundary conditions, and meteorological
606 drivers.

608 While MM ensembles demonstrate improved performance over the individual model realizations,
the most skilful ensemble is not necessarily generated by including all available model results, but
610 instead by selecting models that result in a minimization in ensemble error. In addition, an ensemble
of top-ranking model results can be worse than an ensemble of top-ranking and low-ranking model
612 results. Until now, the prevailing assumption has been that as long as a large set of results was

614 treated statistically in one ensemble, the ensemble would perform better than any individual
ensemble member. Furthermore, it was assumed that the better the model results the better the
616 ensemble. However, the analysis presented here suggests that this is not necessarily the case, as
outliers also need to be included in the ensemble to enhance performance. Furthermore, the skill
618 score does not necessarily improve by 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
620 overall worsening of performance. Despite the remarkable progress of ensemble AQ modelling over
the past decade and the effort spent to build a theoretical foundation, there still are many
622 outstanding questions regarding this technique. Among them, what is the best and most beneficial
way to build an ensemble of members? And how to determine the optimum size of the ensemble in
624 order to capture data variability while minimizing the error?

626 To try address these questions we apply a method for reducing data complexity known as a
clustering technique, which has the advantage of simplifying information provided by the large
628 amounts of data (such as AQ model outputs) by classifying, or clustering, the data into groups
based on a selected metric, where there is no prior knowledge of grouping. Results show that, while
630 the clustering technique needs further refinement, by selecting the appropriate cluster distance and
association criteria, one can generate an ensemble of selected members whose error is significantly
632 lower than that of the full-member ensemble mean. While the results of the clustering analysis are
directly relevant for ensemble model evaluation applications, it is also applicable to other ensemble
634 communities, for example AQ forecasting, climate analysis, and oceanography.

636 **Aknowlegments**

The authors acknowledge the Center for Energy, Environment and Health (CEEH), financed by The
638 Danish Strategic Research Program on Sustainable Energy under contract no 2104-06-0027.
Homepage: www.cee.dk.

640

Appendix A: Statistical Measures

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

644 Mean bias:

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

646 Root mean square error:

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

648 Mean Gross Error:

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

650 Normalised mean square error

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

652 Fractional Bias

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

654 Normalised Mean Bias:

$$NMB = \frac{\sum_i (y_i - obs_i)}{n \overline{y \cdot obs}} \quad (A6)$$

656

Pearson correlation coefficient:

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

660 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](https://doi.org/10.1016/j.atmosenv.2011.11.016)

662 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

664 References

meteorological and air quality models. Environmental Modelling & Software 26, 434-443.

666 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

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

670 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.

674 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.

676 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.

678 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

680

- 682 Brandt, J., J. H. Christensen, L. M. Frohn, C. Geels, K. M. Hansen, G. B. Hedegaard, M. Hvidberg and C. A.
 684 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
- 686 Byun, D.W., Schere, K.L., 2006. Review of the governing equations, computational algorithms, and other
 688 components of the Models-3 Community Multiscale Air Quality (CMAQ) modeling system, Appl.
 Mech. Rev. 55, 51–77.
- 690 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.
- 692 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.,
 694 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.,
 696 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.
- 698 Ciaramella, A., Giunta, G., Riccio, A., Galmarini, S., 2009. Independent model selection for ensemble
 dispersion forecasting, Applications of Supervised and Unsupervised Ensemble Methods, Oleg Okun
 700 and Giorgio Valentini (eds), Springer, pp 213-231.
- 702 Delle Monache, L., Stull, R., 2003. An ensemble air quality forecast over western Europe during an ozone
 episode. Atmospheric Environment 37, 3469-3474.
- 704 Dennis et al., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems.
 Environ Fluid Mech DOI 10.1007/s10652-009-9163-2
- 706 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.
- 708 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.
- 710 ENVIRON, 2010. User's guide to the Comprehensive Air Quality model with extensions (CAMx) version
 5.20 (March, 2010), <http://www.camx.com>.
- 712 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
- 714 Galmarini, S., Bianconi, R., Bellasio, R., Graziani, G., 2001. Forecasting the consequences of accidental
 releases of radionuclides in the atmosphere from ensemble dispersion modelling. Journal of
 716 Environmental Radioactivity 57, 203-219.
- 718 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.
- 720 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
- 722 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
 724 air quality model (AURAMS). Atmos. Res. 82, 248-275.
- 726 Guenther, A., Zimmerman, P., Wildermuth, M., 1994. Natural volatile organic compound emission rate
 estimates for US woodland landscapes. Atmos. Environ., 28, 1197–1210.
- 728 Hamill, T.H., 2000. Interpretation of Rank Histograms for verifying ensemble forecasts. Monthly Weather
 Review 129, 550-560
- 730 Hauglustaine, D.A., Hourdin, F., Walters, S., Jourdain, L., Filiberti, M.-A., Lamarque, 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,
 732 doi:10.1029/3JD003957.
- 734 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.

- 736 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.
- 738 Hogrefe, C., Hao, W., Zalewsky, E.E., Ku, J.-Y., Lynn, B., Rosenzweig, C., Schultz, M.G., Rast, S.,
740 Newchurch, M.J., Wang, L., Kinney, P.L., Sistla, G., 2011. An analysis of long-term regional scale
ozone simulations over the north-eastern United States: Variability and trends. *Atmos. Chem. Phys.*,
11, 567–582.
- 742 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.
- 744 Jacob, D.J., Winner, D.A., 2009. Effect of climate change on air quality. *Atmospheric Environment* 43, 51-
63.
- 746 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.*
748 *Phys.*, 10, 341-364. DOI:10.5194/acp-10-341-2010
- Jolliffe, I. T., and D. B. Stephenson (Eds.) (2003), *Forecast Verification: A Practitioner's Guide in*
750 *Atmospheric Science*, 240 pp., John Wiley, Hoboken, N. J.
- 752 Kumar, A., Barnston, A.G., Hoerling, M.P., 2001. Seasonal prediction, probability verifications, and
ensemble size. *J. Climate* 14, 1671-1676.
- 754 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.
- 756 Mallet, V., Sportisse, B., 2006. Ensemble-based air quality forecasts: A multimodel approach applied to
ozone. *Journal of Geophysical Research* 111, D18302.
- 758 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.,
760 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
762 summer of 2006 Second Texas Air Quality Study field study. *Journal of Geophysical Research* 114,
D00F11, doi:10.1029/2008JD011697.
- 764 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,
766 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.
- 768 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 PM_{2.5}
770 forecast models using data collected during the ICARTT/NEAQS 2004 field study. *Journal of*
Geophysical Research 112, D10S20, doi:10.1029/2006JD007608.
- 772 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*
774 [10.1016/j.atmosenv.2011.11.023](https://doi.org/10.1016/j.atmosenv.2011.11.023)
- Potempski, S., Galmarini, S., 2009. Est Modus in Rebus: analytical properties of multi-model ensembles.
776 *Atmospheric Chemistry and Physics* 9, 9471-9489.
- Pouliot, G., Pierce, T., Denier van der Gon, H. ., Schaap, M., Moran, M., Nopmongcol, U., Comparing
778 Emissions Inventories and Model-Ready Emissions Datasets between Europe and North America for
the AQMEII Project. *Submitted for publication to Atmospheric Environment (this issue)*
- 780 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
- 782 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–
784 1912.
- Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., Potempski, S., 2012. On the systematic
786 reduction of data complexity in multi-model ensemble atmospheric dispersion modelling. *Journal*
Geophysical Research, *in press*.
- 788 Russell, A., Dennis, R., 2000. NARSTO critical review of photochemical models and modelling.
Atmospheric Environment 34, 2283 – 2324.
- 790 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

- 792 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.
- 794 Schere, K., Flemming, J., Vautard, R., Chemel, C., et al. Trace Gas/Aerosol concentrations and their impacts
on continental-scale AQMEII modelling sub-regions. *Atmospheric Environment*
796 doi:10.1016/j.atmosenv.2011.09.043
- Schmidt, H., Derognat, C., Vautard, R., and Beekmann, M., 2001. A comparison of simulated and observed
798 ozone mixing ratios for the summer of 1998 in Western Europe. *Atmospheric Environment*, 36, 6277-
6297.
- 800 Simpson, D., A. Guenther, C. N. Hewitt, R. Steinbrecher, 1995. Biogenic emissions in Europe. 1. Estimates
and uncertainties. *J. Geophys. Res.*, 100D, 22875–22890.
- 802 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
804 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,
806 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.
- 808 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.
810 *Atmos. Environ.*, 43, 1059-1070.
- Sofiev, M., Siljamo, P., Valkama, I., Ilvonen, M., Kukkonen, J., 2006. A dispersion modeling system
812 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
814 evaluation for particulate matter in Europe and North America in the context of the AQMEII project.
Atmospheric Environment
- 816 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.
- 818 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.
820 *Atmos. Environ.* 41, 2083-2097.
- Vautard, R., M. Schaap, R Bergström, B. Bessagnet, J. Brandt, P.J.H. Builtjes, J.H. Christensen, C. Cuvelier,
822 V. Foltescu, A. Graf, A. Kerschbaumer, M. Krol, P. Roberts, L. Rouil, R. Stern, L. Tarrason, P.
Thunis, E. Vignati, P. Wind, 2009, Skill and uncertainty of a regional air quality model ensemble.
824 *Atmos. Environ.*, 43, 4822-4832.
- Vautard, R., Moran, M.D., Solazzo, E., Gilliam, R.C., Matthias, V., et al. Evaluation of the meteorological
826 forcing used for AQMEII air quality simulations. *Atmospheric Environment*
10.1016/j.atmosenv.2011.10.065
- 828 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.
- 830 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,
832 doi:10.1029/2006GL027610.
- 834 Wolke, R., Knuth, 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.
- 836

Captions

838

Table 1. Participating models and important characteristics.

840

Table 2. RMSE-ranked combinations of models that give minimum RMSE for each sub-region. The minimum of all combinations is listed in bold.

842

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

844

Table 4. Ranking of cluster representatives for EU and NA sub-regions for varying PCC distance.

848

Figure 1 – Continental maps of (a) Europe and (b) North America with locations of sub-regions marked. The dots and other symbols denote the positions of the rural ozone receptors used in the evaluation analysis. The contours indicate the summertime anthropogenic NO_x emissions (in kg km^{-2}) from the standard inventories.

854

Figure 2. Box-plots of ozone frequency distribution at receptors, averaged in space over (a) EU domain and (b) NA domain and in time for the whole 2006 year, for observations (MEAS), individual SMs, and two MM ensembles (Mean, Median).

856

Figure 3. Rank histogram for the whole domain of (a) EU and (b) NA, full-model ensemble, hourly data for the whole 2006 year.

860

Figure 4. Time series (JJA) of diurnal ozone cycle for (a) EU and (b) NA sub-regions

862

Figure 5. Normalized Mean Bias vs. Normalised Mean Square Error for (a) EU and (b) NA. Sub-regions 1 to 4 are represented by number and coloured by model or ensemble. Mean and median for each sub-region are highlighted by boxes.

864

Figure 6. Rank histogram for (a) EU and (b) NA by sub-regions, full-model ensemble, hourly data for the period JJA.

868

Figure 7. RMSE, MGE or MB, and PCC of the ensemble mean of any possible combination of members for (a) EU, and (b) NA. Continuous lines denote the mean and the median of the distributions

870

Figure 8. Daily maximum concentrations for EU sub-regions, for the period JJA. Horizontal axis: ensemble maximum of all available members. Vertical axis: ensemble maximum of model combinations with minimum RMSE.

872

Figure 9. Hourly ozone concentrations ($\mu\text{g m}^{-3}$) for the period JJA at receptor positions: (a) observations; (b) ensemble of ranked models 1 and 5; models ranked (c) 2nd; and (d) 11th.

878

Figure 10. Box-plots of observed ozone concentration, full-model ensemble and selected-model (combinations with minimum RMSE) ensemble. Top row: EU sub-regions; bottom row: NA sub-regions.

880

Figure 11. Dendrograms of model clustering as function of mutual PCC distance for EU sub-regions

884

Figure 12. Dendrograms of model clustering as function of mutual PCC distance for NA sub-regions

886

Figure 13. 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 rankings of MM members. Different colours correspond to different sub-regions for (a) EU and (b) NA.

888

890 **Tables**
Table 1

	Model		Res (km)	No. Vertical layers	Emissions	Chemical BC
	Met	AQ				
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 [†]
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

892 [§] 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).

894 *Standard inventory for NA includes biogenic emissions (see text).

896 [†]Standard anthropogenic inventory but independent emissions processing, exclusion of wildfires, and different version of BEIS (v3.09) used.

898 **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	

900

902

904

906

908

910

Table 3.

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

912

914

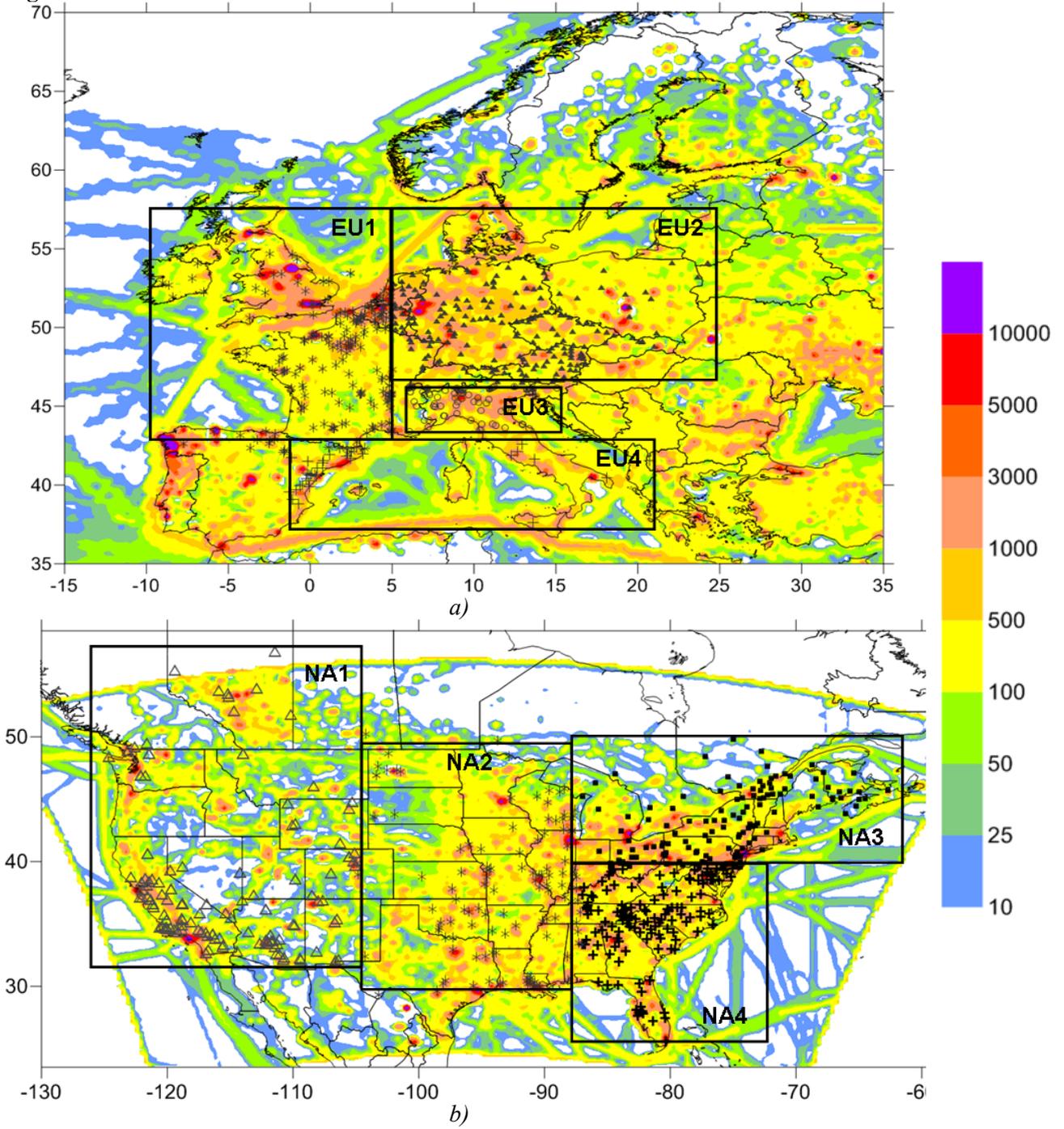
Table 4

	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

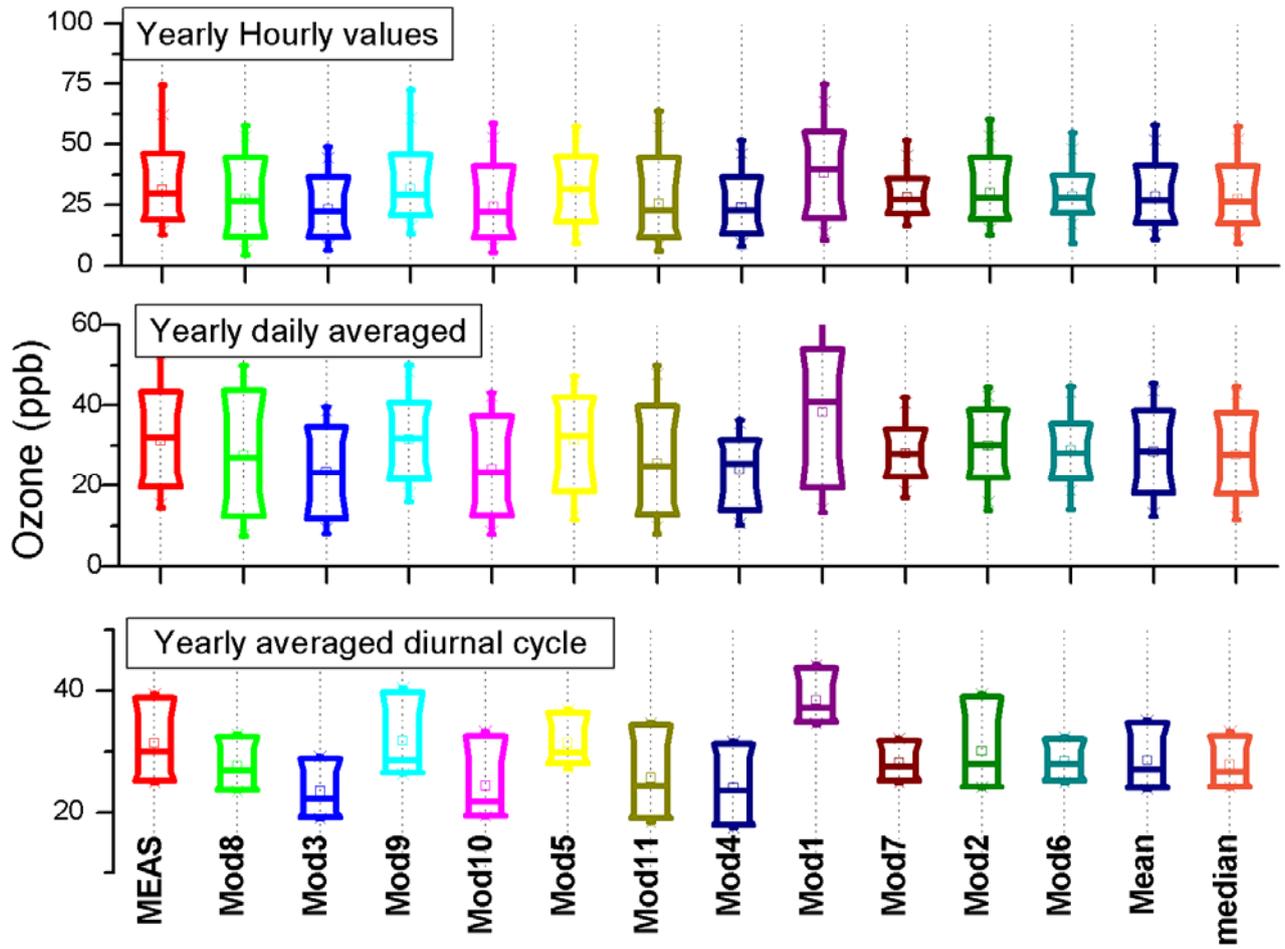
916

918

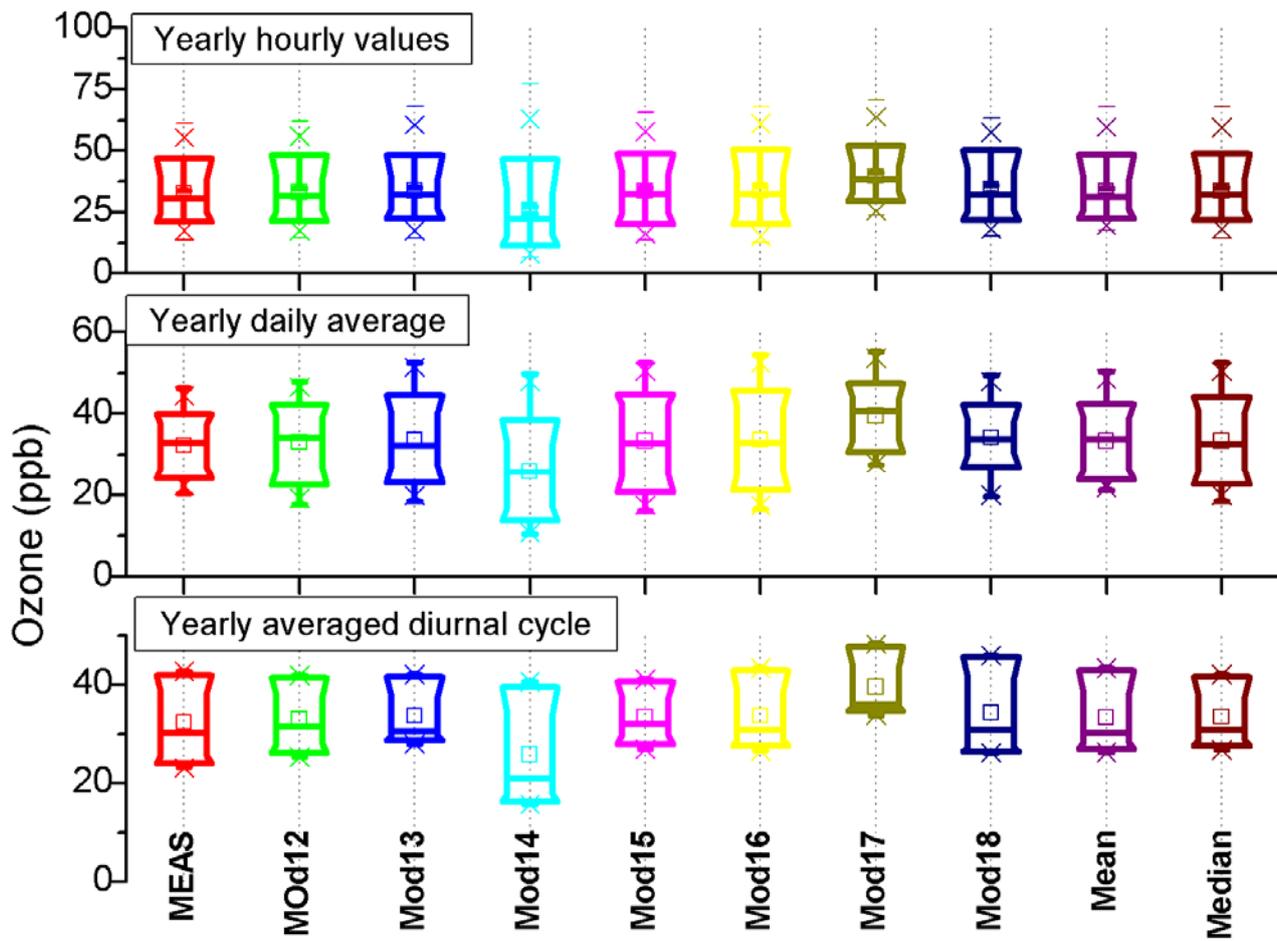
Figures



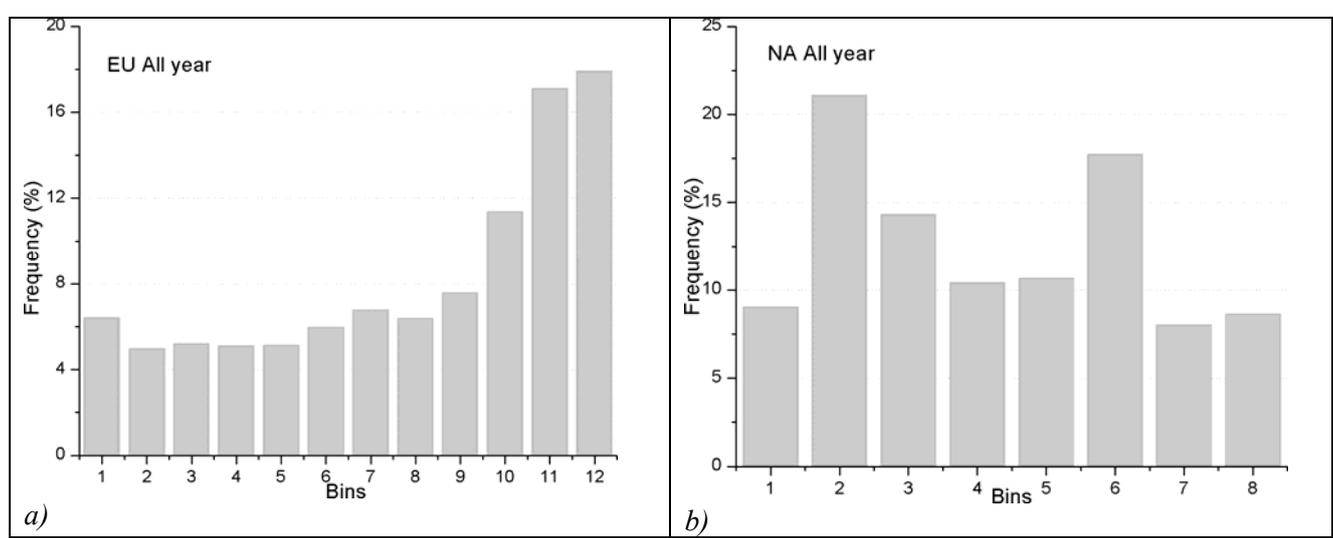
922 Figure 1



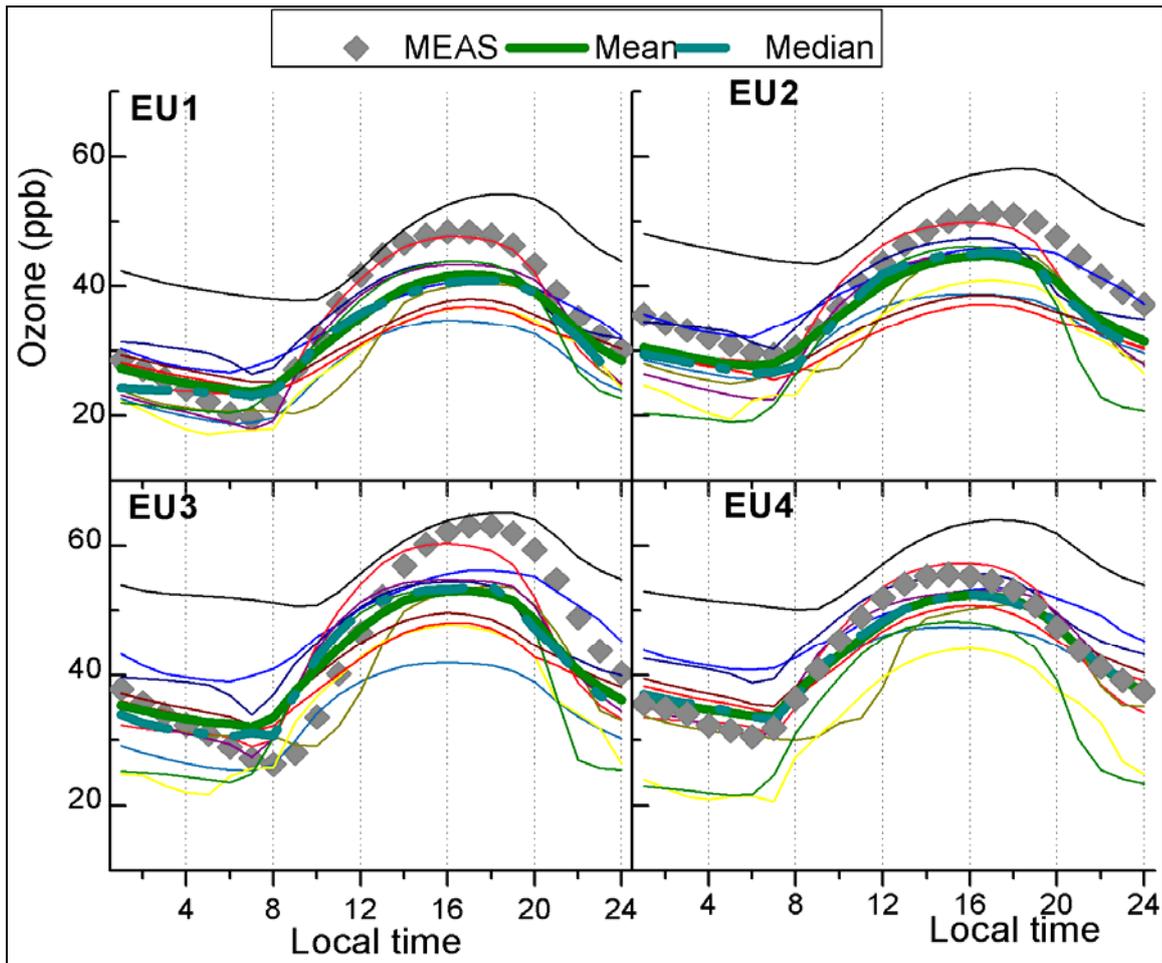
a)



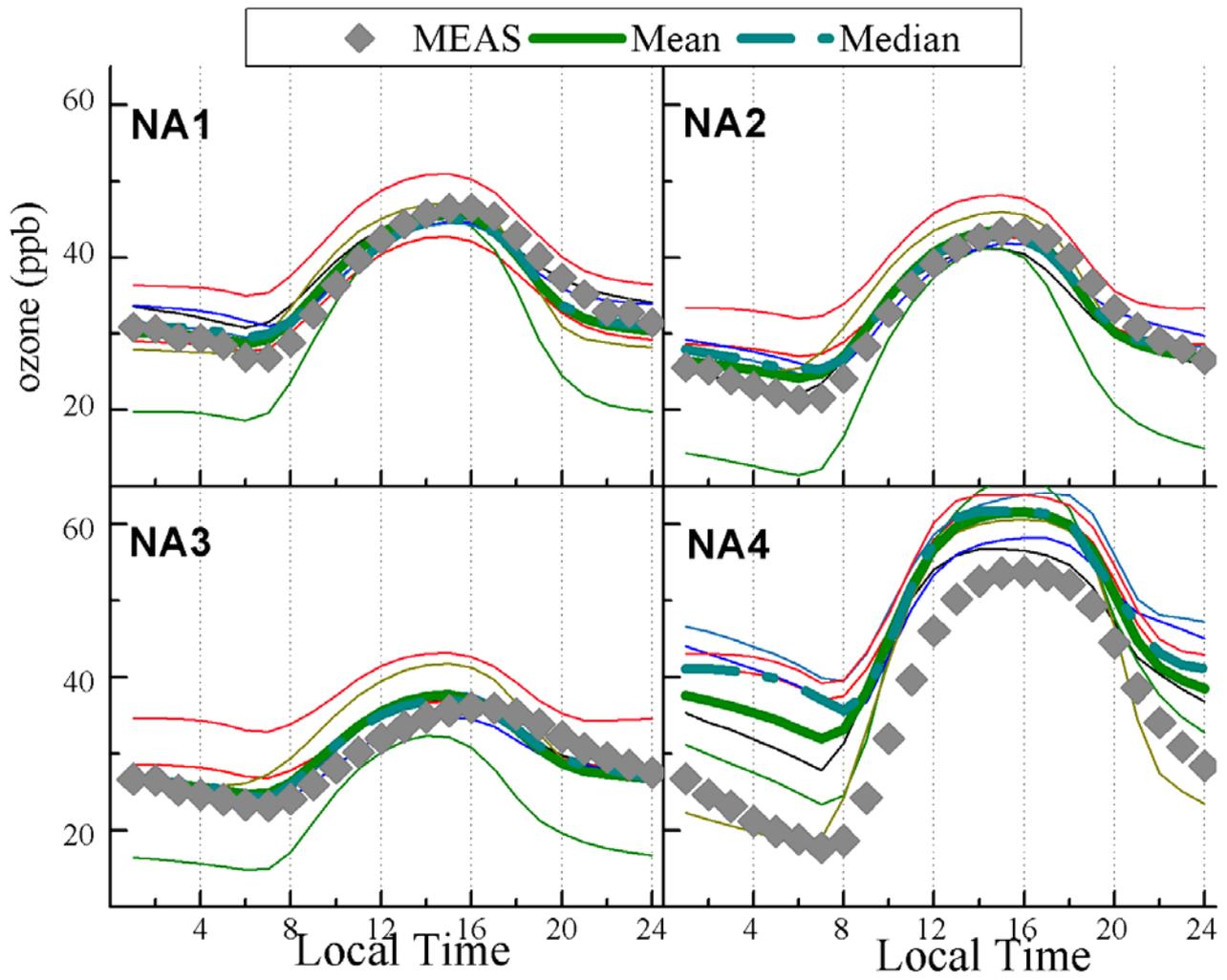
b) Figure 2



a) Figure 3



a)

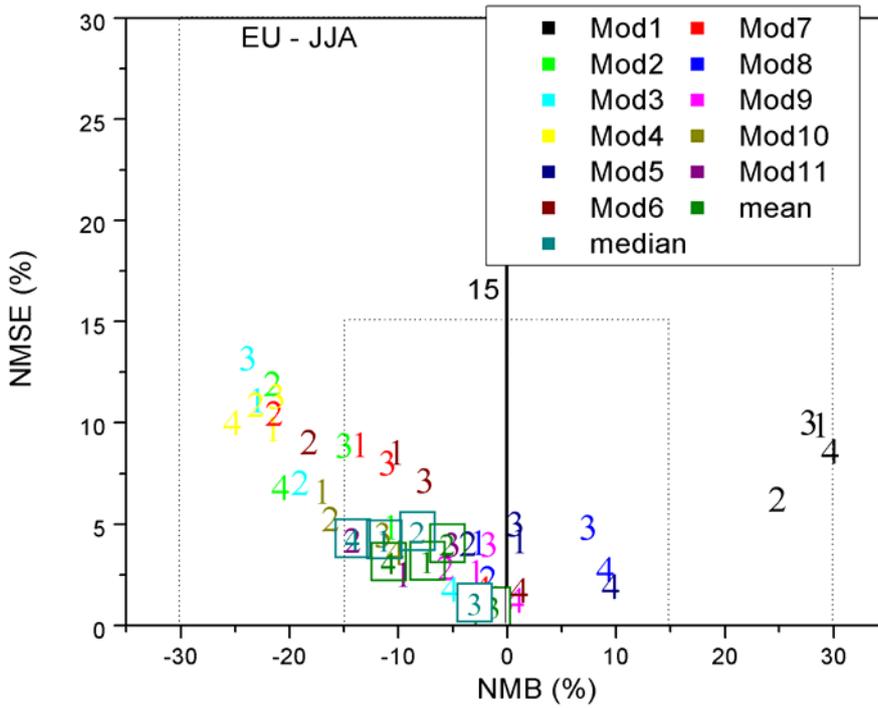


b)

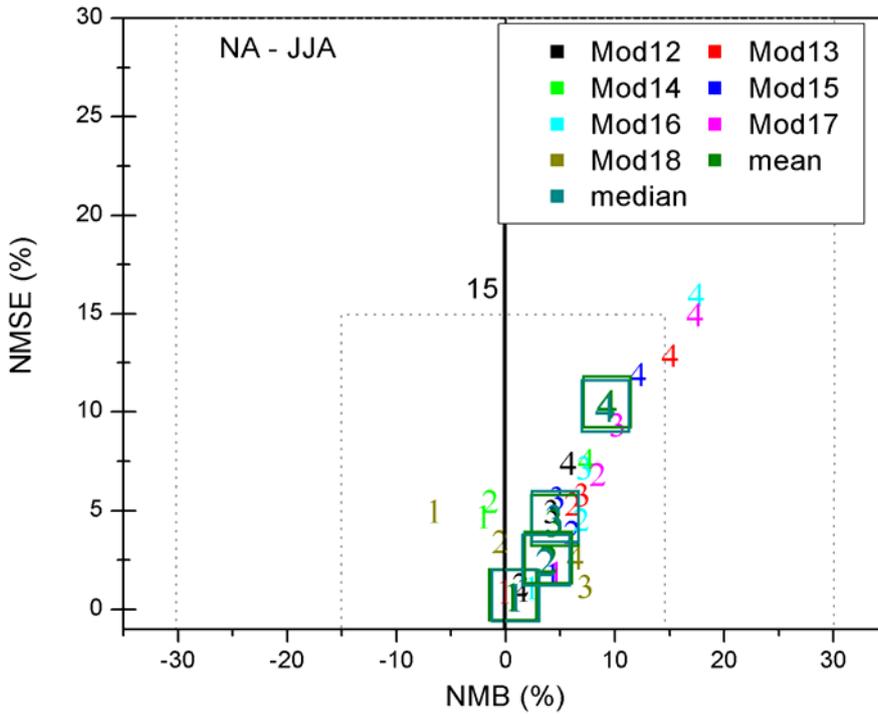
928 **Figure 4.**

930

932



a)



b)

Figure 5.

934

936

938

940

942

944

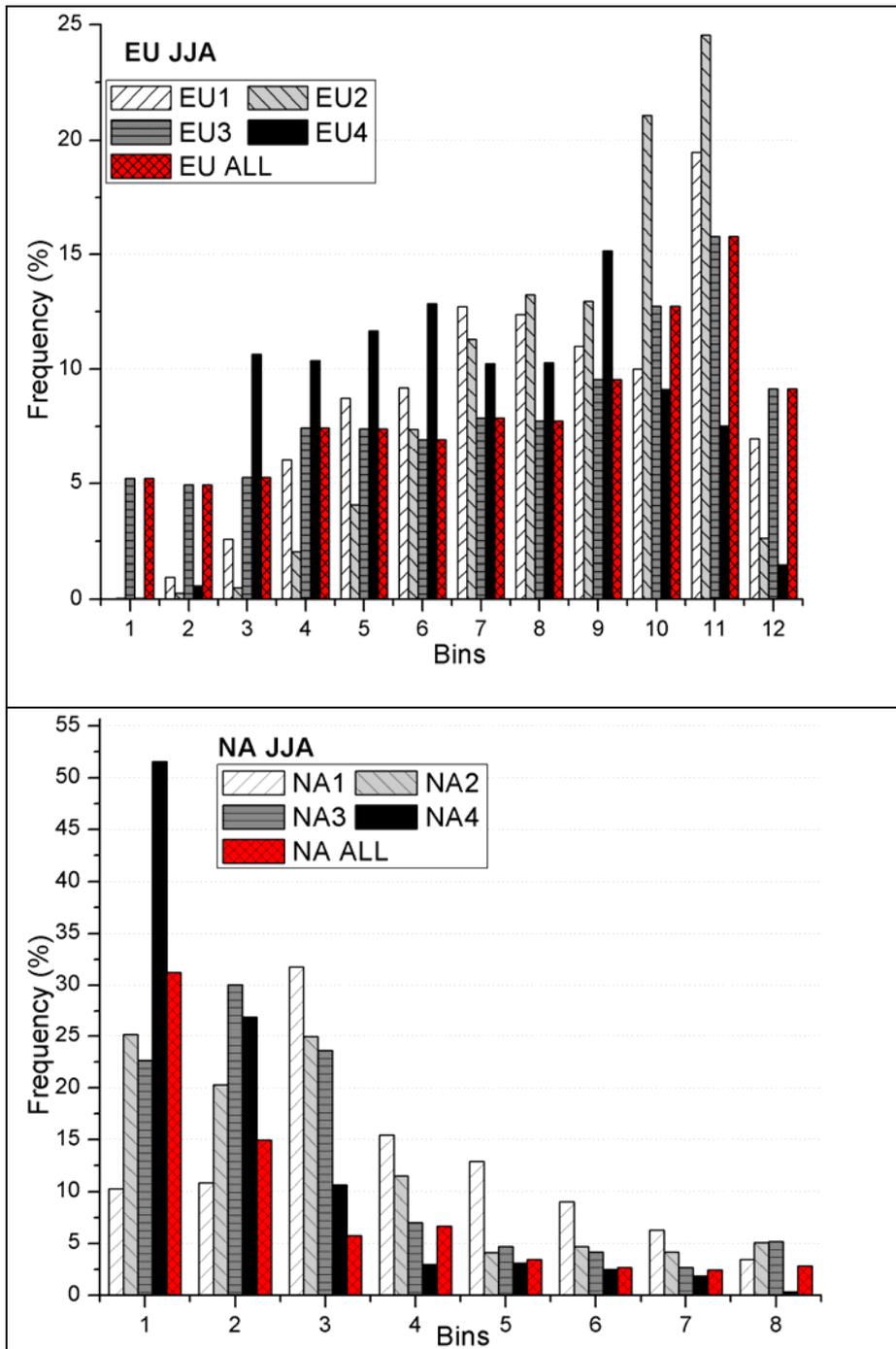
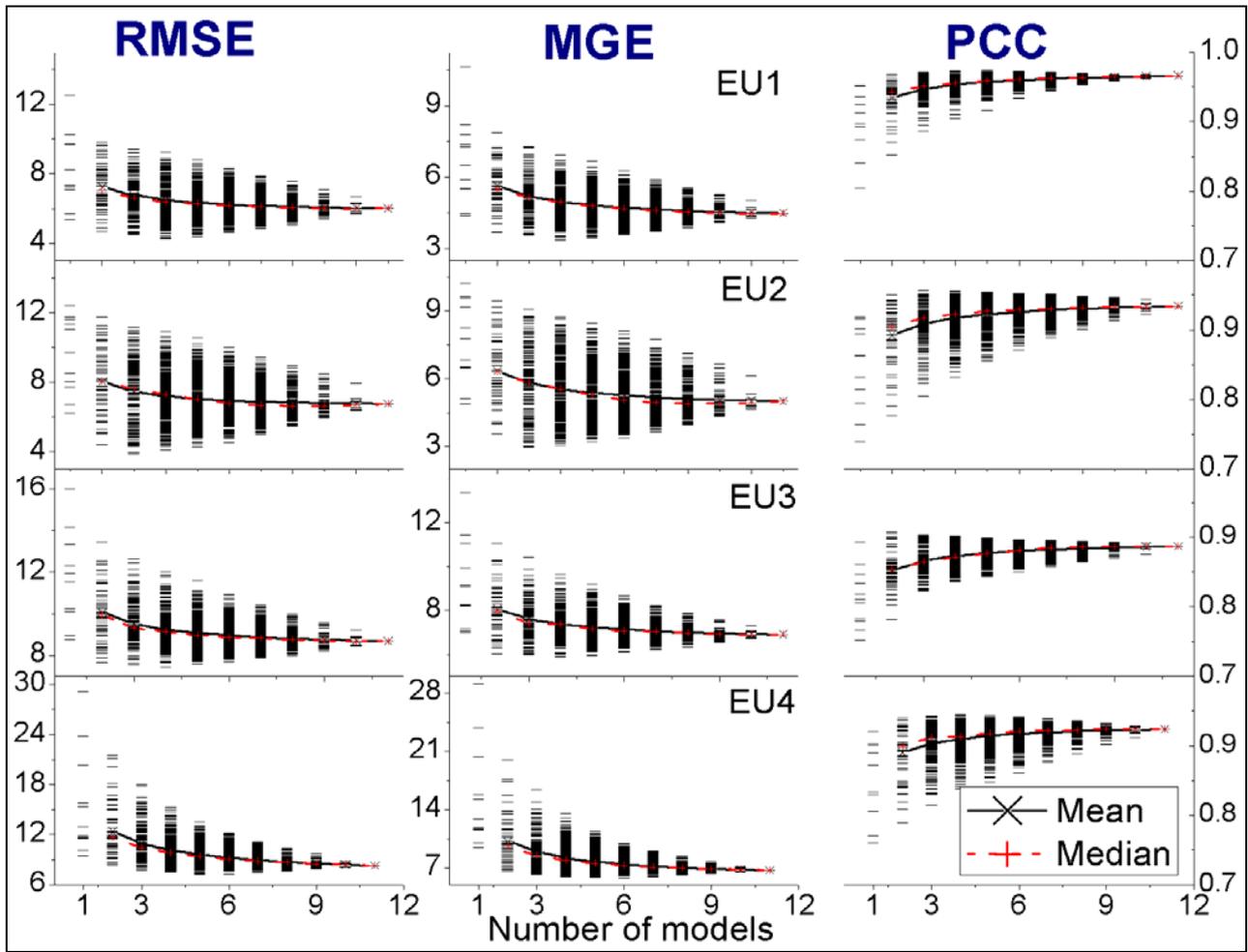
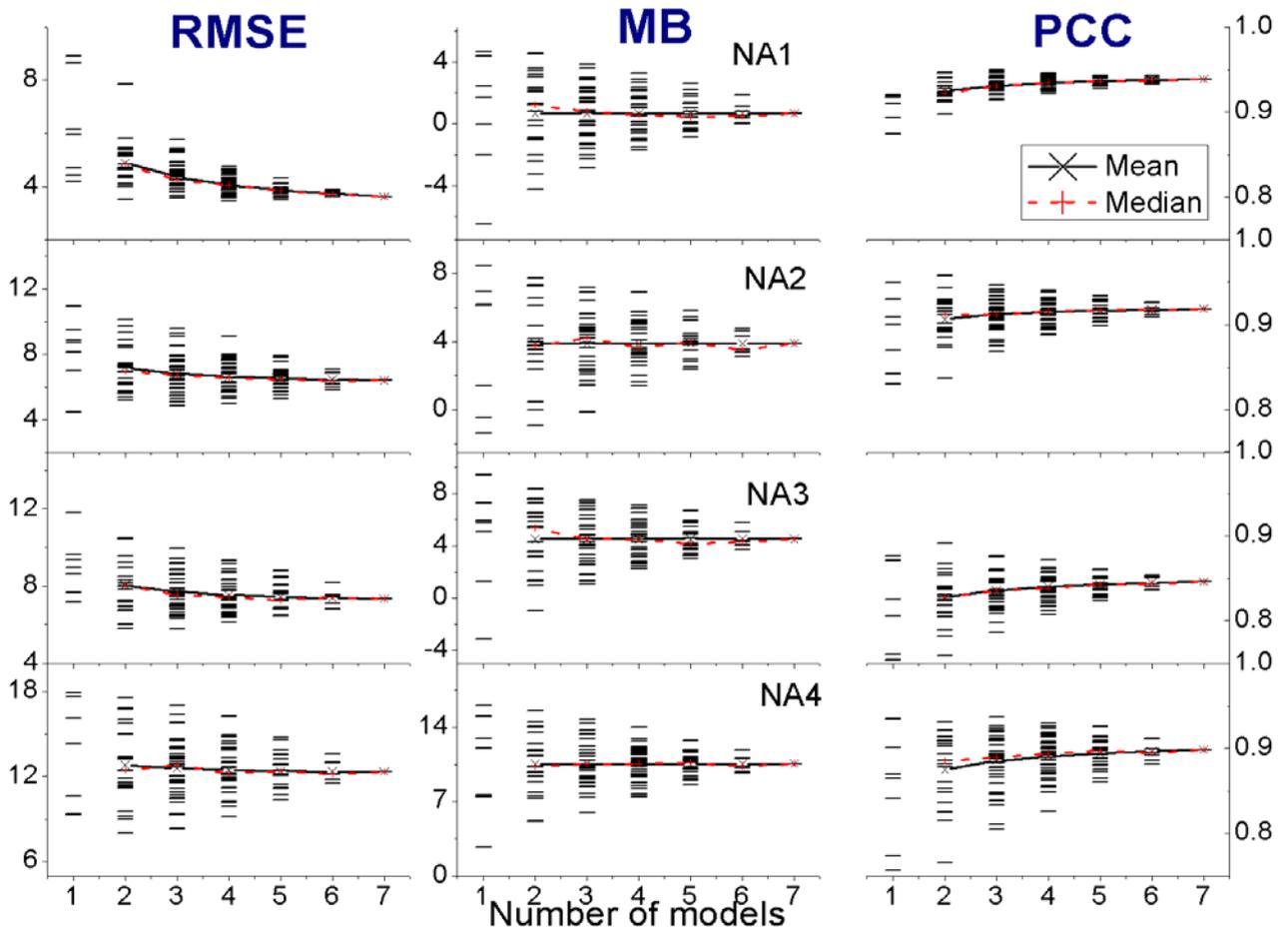


Figure 6



948 **Figure 7a**



950 **Figure 7b**

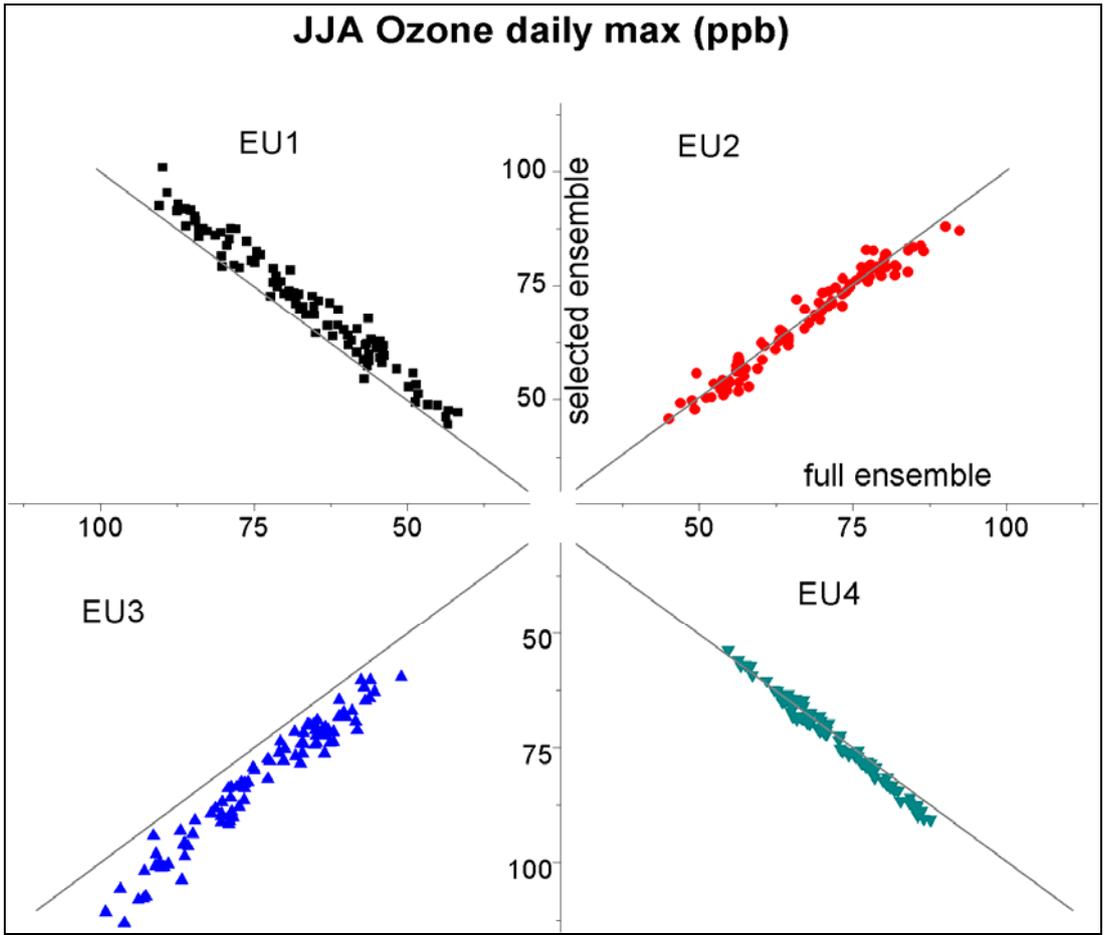


Figure 8

952
 954
 956
 958
 960
 962
 964
 966
 968
 970
 972
 974

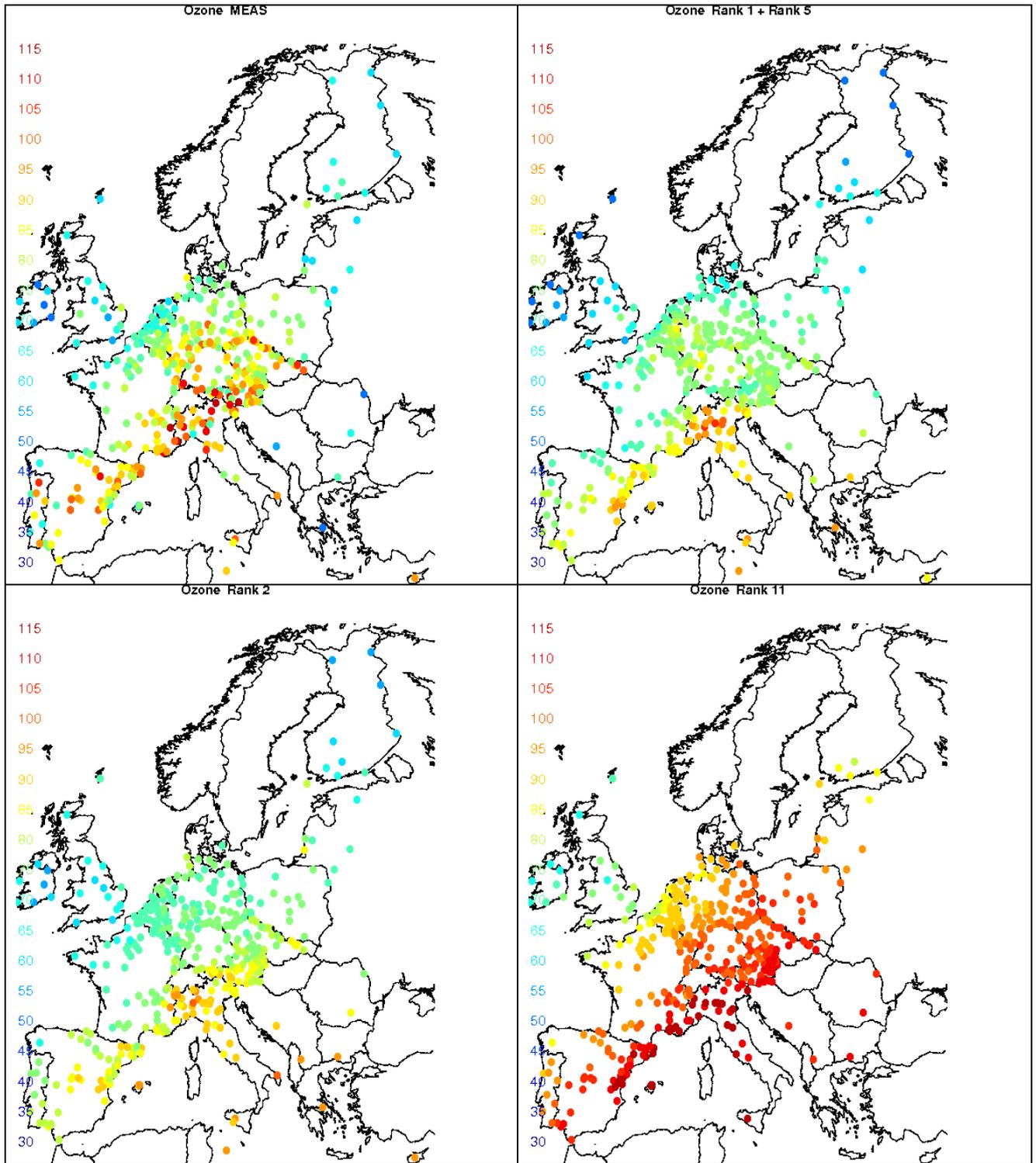


Figure 9

976

978

980

982

984

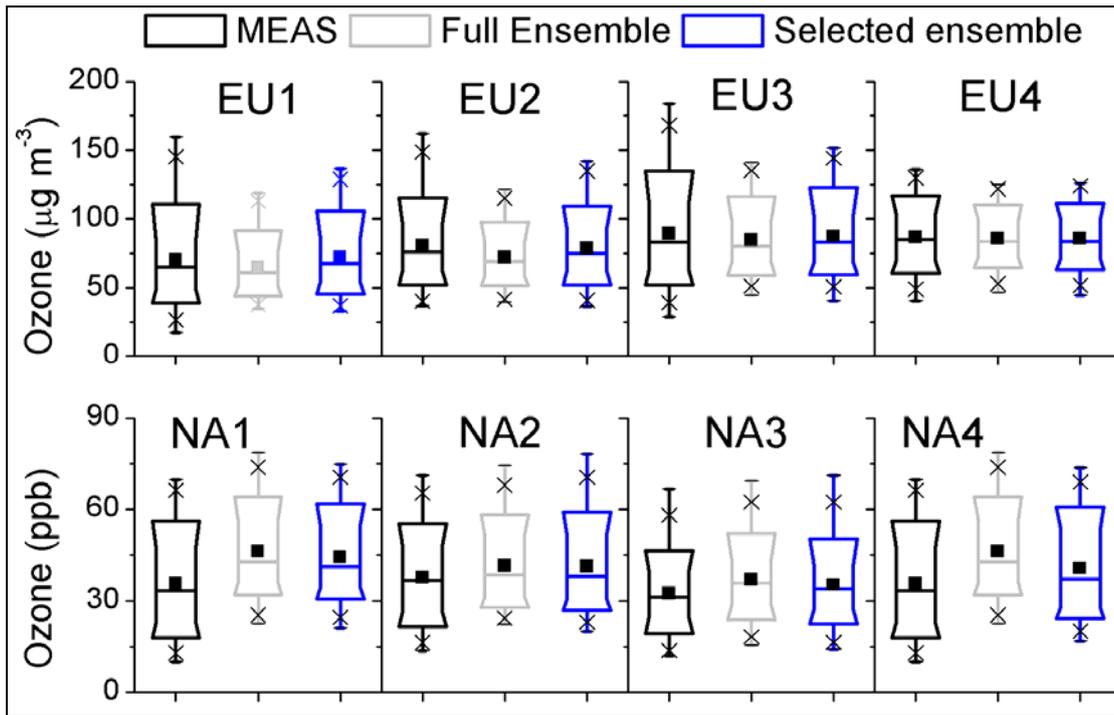
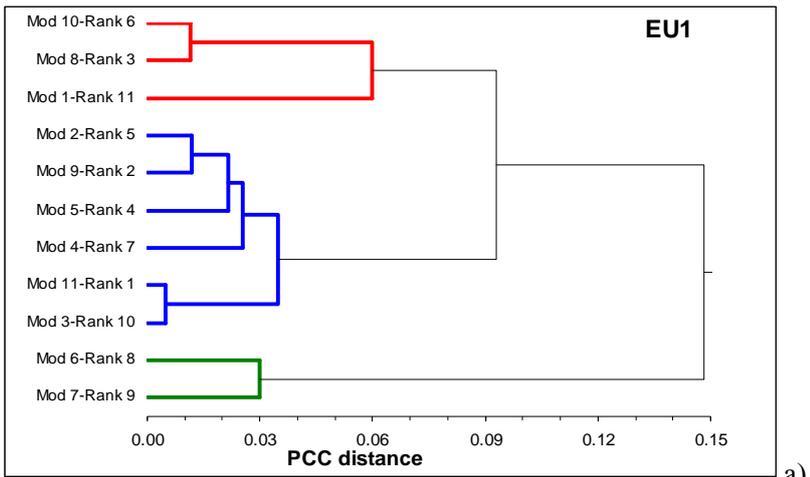
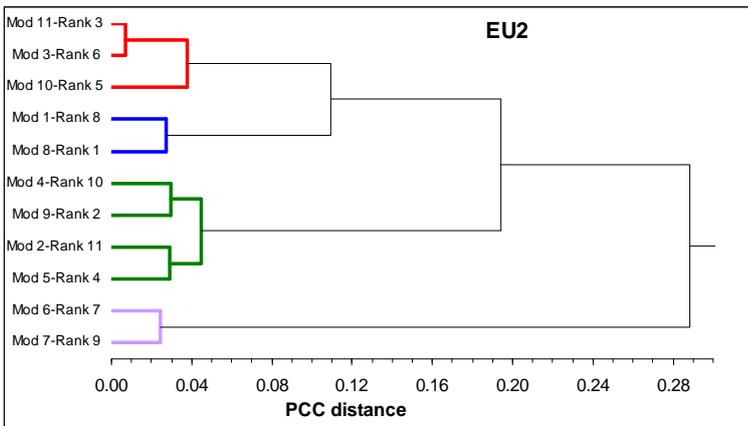


Figure 10

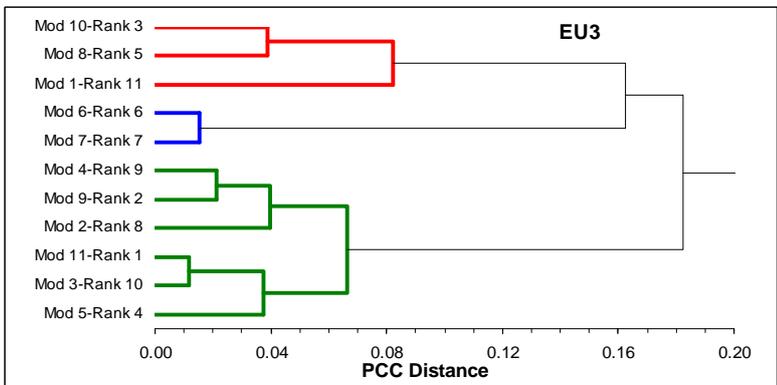
986
 988
 990
 992
 994
 996
 998
 1000
 1002
 1004
 1006
 1008
 1010
 1012
 1014
 1016
 1018



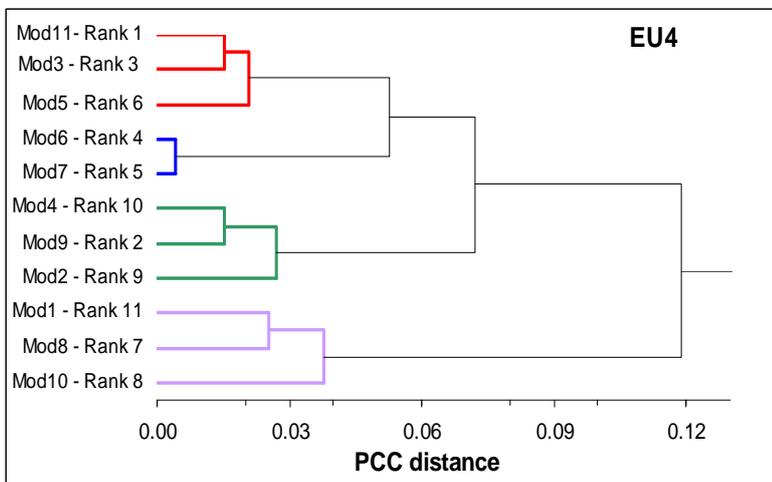
a)



b)



c)



d)

Figure 11

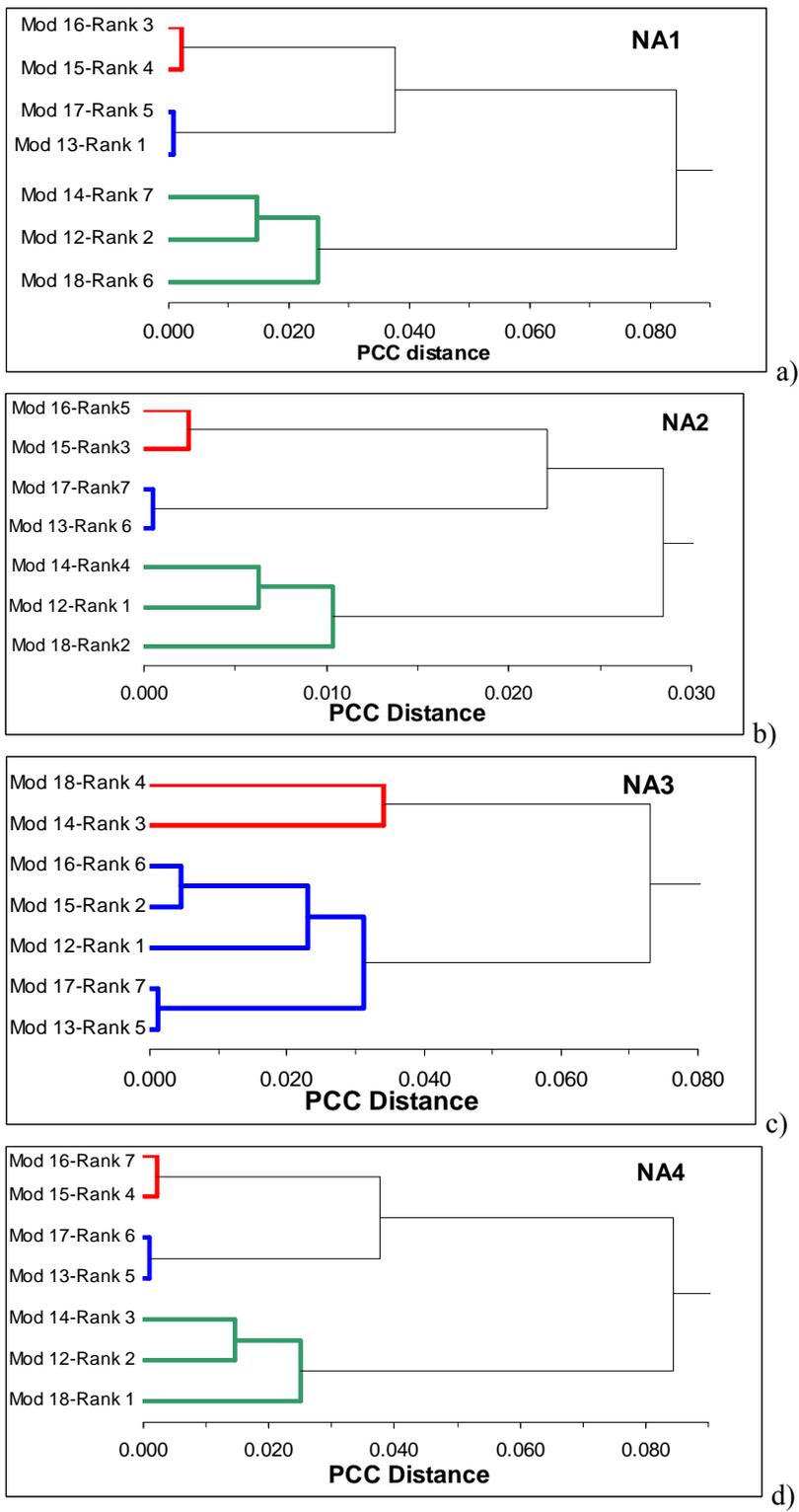
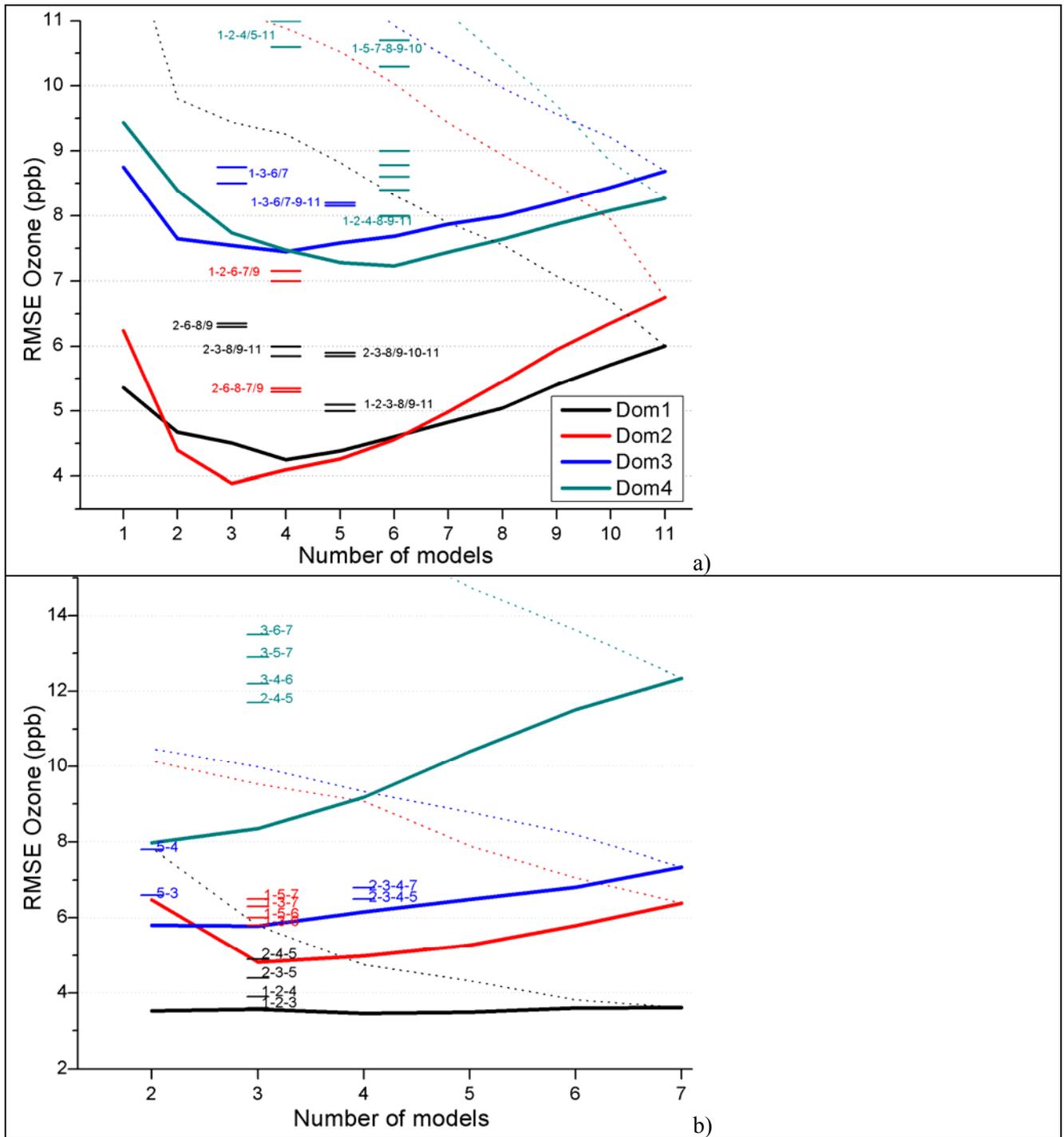


Figure 12

1022

1024



1026 **Figure 13**