

Dear Prof. Carmichael,

we have uploaded the revised version of the paper entitled “Insights into the deterministic skill of air quality ensembles from the analysis of AQMEII data (acp-2016-513)”. We would like to thank you for all the detailed comments concerning the mathematical/physical rigor and clarity of the paper, and we tried our best to comply with the reviewers’ comments. Accordingly, we believe we have been able to significantly enhance both the content of the paper and quality of English.

We have re-written and re-organized the majority of the presented material in order (a) to emphasize the originality of the work and (b) to remove any comparison between AQMEII phase 1 and 2. Specifically, we have:

- Added a more focused title;
- Introduced a new section 3 with the experimental setup;
- Merged the old sections 3 and 4 into one (Results), with three subsections; one for the single models, one for the unconditional ensemble mean and its pitfalls and one with the conditional ensemble estimators;
- Replaced all tables with others containing more information;
- Expanded the conclusions with the ‘take-home’ messages.

We believe, thanks to the reviewers, we have been able to accomplish a worthwhile improvement of the quality of the paper and have, we hope, clarified the points raised by the reviewers.

Sincerely yours,

Ioannis Kioutsioukis and Stefano Galmarini

Referee #1: Matthieu. Plu

We thank Dr Matthieu Plu for the positive and helpful comments that have improved the manuscript. They have all been taken on board and addressed in the revised version of our manuscript.

General Comments:

The manuscript entitled "Improving the deterministic skill of air quality ensembles" reports about the properties and the scores of different ensembles applied on the model outputs of the two phases of AQMEII (2006 and 2010) over Europe. The presentation of the manuscript and of figures is good. To my view, the study has several merits and the results that are presented are original enough to be published. However, the manuscript should be improved along the following recommendations in order to go over the Discussion step.

Specific Comments:

1. There is a lack of focus of the manuscript on the main relevant original ideas that are demonstrated. Many interesting results are presented, and the manuscript needs to be focussed on one or two main new scientific questions that the manuscript addresses. These lines should be followed from the abstract to the conclusion. To make my argument more understandable, I would like to point out the following:

- the results that are reported in the first paragraph of the abstract - from lines 6 to 10 - are not new as they were demonstrated in past research articles. These lines mislead the reader on the purpose of the article;

- many times (page2-line12, page4-line4, page7-line12), "two ensembles" appear in the text, but it is not clear whether it refers to two different ensemble methods (and actually, the article compares more than two methods) or to the two AQMEII phases.

A suggestion would be to present the article as a comparison of different ensemble methods (mmW, mmS, KZ, ...), applied on two different datasets (both phases of AQMEII). Actually, the manuscript does not present new methods, but it compares the performance of existing ensemble methods on different pollutants and on different periods. What I consider also to be original are the diagnostics (such as Figure 5) that have been developed and that are used to analyse the ensemble properties. The objectives written page 4 (lines 20-23) may also be the relevant lines to follow, which is not fully obvious in the present manuscript.

Response: Thank you for the valuable suggestion. The manuscript has been rewritten as an analysis of the performance of different ensemble techniques rather than a comparison of the results from the two phases of the AQMEII activity, focusing on the originality of the study that includes: (a) the comparison of several ensemble methods on pollutants of different skill using different datasets, (b) the introduction of an approach based on high-dimension spectral optimization, (c) the introduction of innovative charts for the interpretation of the error of the unconditional ensemble mean with respect to indicators

reflecting the skill difference and error dependence of the models as well as the effective number of models.

2. In the same line of thoughts, the title of the manuscript is too general and it should be more specific. A general title such as the one that appears now could apply to many papers that have already been published.

Response: Done as suggested. The title has been changed to "Insights in the deterministic skill of air quality ensembles from the analysis of AQMEII data"

3. Comparing the ensemble performance between the two AQMEII phases does not bring much to the study and can be misleading, since:

- the observation dataset changes (no PM10 observations available from UK nor France in Phase II, page8-lines7-8),
- the period (meteorological regimes, types of pollution, etc) change,
- the individual models change in depth.

The differences in ensemble performances between phases I and II (page9 for instance) are subject to all these differences. The attribution of differences of ensemble performances between the two phases should be done cautiously, making only one variable change at each time for any interpretation. If it is not possible, I suggest then to remove the discussions about the differences between Phase I and Phase II of AQMEII.

Response: Done. The new presentation is after an analysis of the performance of different ensemble techniques rather than a comparison of the results from the two phases of the AQMEII activity.

4. There is a lack of description of the experimental setup of the two phases of AQMEII, that would help the reader to understand some of the conclusions that are drawn, such as the arguments at page9-line19, page16-line, among others. The manuscript as it is written now is not self-consistent. To improve this, I would suggest to add in section 3.2 the key facts of both AQMEII phases: general experimental setup (domain, periods, common input data and setup for all models) and the different models that participate (name, chemical and aerosol schemes, resolution, meteorological model, etc). At least the key facts that are needed to understand the discussions should appear in the manuscript. For the rest, the manuscript should cite some AQMEII reference articles.

Response: Done as suggested. A new section 3 with the experimental setup has been added.

Minor comments:

- page 3, if the "Recent results" (line 7) refer to the citation (Eskes, 2002) (line 9), then the word "Recent" does not apply; if they refer to an actual recent other work, please cite it,

Response: Done as suggested. The word 'recent' has been removed.

- the manuscript would gain in clarity if the KZ methods (page6) were described more in depth; for instance the page12-lines(7-10) sentence is somehow enigmatic.

Response: Done as suggested. Two paragraphs have been introduced in section 2. The first describes the rationale behind spectral optimization (spectral decomposition equations are provided in the Appendix). The second presents the examined ensemble estimators with reference to their theoretical basis.

- Is the quarter (September-October-November) chosen for NO₂ the most relevant one? Do not the December-January-February quarters show higher NO₂ concentration levels?

Response: We choose a continuous seasonal time series for each pollutant.

- The sentence page7-line22 would better fit in section 3.1.

Response: Done as suggested.

- page8-lines5-8: the sentence "the decline ... due to .. sampling stations." should be proven by some diagnoses or adequate citation.

Response: The sentence has been removed from the text.

- page10-line 2: the sentence "the benefits of ensemble ... members)." is not fully clear and maybe not true: what happens if we take the 6 "worse performing" models?

Response: We have rephrased the sentence to emphasize that we do not mean particular models.

- page10-line17: reference to the relevant figure is needed.

Response: The sentence has been removed from the text.

- page17-line28: remove '**

Response: Done as suggested.

Referee #2: Anonymous

We thank the anonymous reviewer for the many helpful suggestions that have improved the manuscript. They have all been taken into consideration and addressed in the revised version of our manuscript.

General/Specific Comments:

The paper analyses the two phases of the AQMEII initiatives to test different techniques for improving deterministic estimates from multi-model ensembles. Even though the paper is generally well written, my opinion is that the scientific novelty is scarce, and most of the conclusions are not solid. Here are my motivations:

1) As stated in the Abstract: Line 5-7 “we demonstrate. . . is far from optimum,”. This has been already proved several times in previous publications. (see Solazzo et al. 2013, Riccio et al. 2007, Galmarini et al. 2013, among others). In these papers, the same concepts and techniques of reducing the dimensionality of multi-model ensembles and optimal combination have been widely and repeatedly presented.

Response: We have re-written many parts of the manuscript to make more clear the focus and originality of the study. The scientific novelty of the study includes: (a) the comparison of several ensemble methods on pollutants of different skill using different datasets, (b) the introduction of an approach based on high-dimension spectral optimization, (c) the introduction of innovative charts for the interpretation of the error of the unconditional ensemble mean with respect to indicators reflecting the skill difference and error dependence of the models as well as the effective number of models. The manuscript has been rewritten to better reflect its originality.

2) Pag 4 line 4-13. The differences between the two experiments are described. The differences in the meteorology (two different years) and stations (amount of observations and their locations) are those that undermine more the statistical significance of the results. Most of them are presented (see Table 2 and Table 4) without bootstrap confidence intervals or other techniques to assess if the differences between the two phases are statistically significant. The numbers of Phase I and II are often very close, and despite that, the authors build many conclusions on the top of these small differences. Also, most of the differences (if any) could be explained by the meteorology or the underlying changes in the station network. The authors should, at least, have made an attempt to make the two experiments more homogenous, i.e. by keeping a similar kind of stations over the two phases (same amount of urban, background stations).

Response: We have tested the statistical hypotheses on the differences of the distributions and their means through the Kolmogorov-Smirnov test and the t-test respectively. Although many differences were generally significant at the 1% level, we have decided to remove the comparisons of the two phases. The two experiments were independently designed and executed and have many differences. The harmonization of the validation set would remove an uncertain factor. Even then, the attribution of the differences between the two datasets to the uncertain factors (meteorology, models, coupling, etc.) in a statistical framework

would still include a considerable amount of uncertainty. Moreover, such quantitative decomposition is beyond the objectives stated in this study. Therefore, the manuscript has been rewritten as an analysis of the performance of different ensemble techniques rather than as a comparison of the results from the two phases of the AQMEII activity.

3) Section 4.1 Forecasting performances. The authors want to prove that the weighting scheme might be used in forecasting mode. There are two issues here that undermine the conclusions of this section. My understanding is that some of the models participating at the inter-comparison are not running in forecasting mode (they use meteorological reanalysis as boundary conditions). While they should run as an operational real-time forecasting model to be considered as realistic forecasts. Running these model in forecasting mode would change the model behaviors and error structures. Hence the conclusions achieved might change as well. How the bias of the models is removed in this test? Using the bias computed over the entire period (as previously mentioned) to correct forecast issued over the same test period would not be possible in real-time forecasting. This simple bias removal technique might not be so effective especially in forecasting mode when data from the future cannot be used.

Response: With respect to the first issue, the term ‘forecast’ has been changed to ‘simulation’ throughout the text. Concerning the second issue, bias removal is beneficial to the ensemble mean according to the bias-variance-covariance decomposition. It is not necessary for the approaches relying on reduced-dimensionality ensembles but the formulas for the analytically optimized weights have been derived with the assumption of bias-free members. As for the implementation, the mean bias over the training period is removed from the time-series of the test dataset. An explanation has been added in the text to clarify that the bias calculated in the train dataset (for the examined training periods of 5-60 days) is subtracted from the test dataset.

Our results indicate that after 30-60 days, the variable biases and weights have no effect in the skill of the weighted ensemble mean. Besides that, the seasonal bias reflects the systematic errors of the single models and it is considered a known quantity for validated models. Those considerations support the possible application of the approaches in real-time forecasting.

Minor comments:

The sentence: “In addition, mathematical tools such as ensemble forecasting provide an extra channel for uncertainty quantification and eventually reduction. Such method seems similar to the Monte Carlo approach; in practice, the similarity is only phenomenological since the probability density function of the uncertainty is not sampled in any statistical context like random, latin-hypercube, etc.” is not clear at all. Ensemble forecasting cannot be considered as a mathematical tool in general. What does it mean:” Similarity in only phenomenological. . .”?

Response: The sentence has been removed from the text.

“benefits from ensemble forecasting arise from the averaging out of the unpredictable components (Kalnay, 2003).” It would be correct to say that benefits arise from averaging estimates with uncorrelated errors.

Response: Done. The sentence has changed accordingly.

Pag 3 line 25 “One of the challenges in ensemble forecasting is the processing of the deterministic models”. This is true only if you are talking about a multi-model ensemble.

Response: Done. The sentence has changed accordingly.

Eq1 bias, var, cov? Should be presented with a more detailed notation

Response: Done.

Eq 2 E is the mean over what?

Response: Eq 1 and Eq 2 are related through their expectations over multiple stations.

Line 6 page 8 keep the same stations over the two phases

Response: We do not compare differences between the two phases in the revised manuscript.

Line 24 page 8 indirect feedback of what? Some details should be added

Response: The sentence has been removed from the text.

Line 19 page 5 I'd say the minimum (what does it mean ideal?)

Response: Done. The sentence has changed accordingly.

Section 2.1 86 % is a general value or something related to this paper

Response: It is general, the first N_{EFF} members account for 86% of the variability.

The same bunch of authors (or most of them) appears in previous publications regarding AQMEII phase I and II. I have some doubts (but I might be wrong) that they all give an active contribution to this paper or at least original compared to what already provided in the previous publications regarding these experiments. It would be fair to include in detail a description of the contribution of each author to this paper.

Response: The authors present in many AQMEII publications, as well as this one, are from the modeling groups that performed the simulations. Without the simulations, none of the published analyses would be possible.

1 Insights into the deterministic skill of air quality ensembles from the 2 analysis of AQMEII data

3
4 Ioannis Kioutsioukis^{ab}, Ulas Im^c, Efisio Solazzo^b, Roberto Bianconi^d, Alba Badia^e, Alessandra
5 Balzarini^f, Rocío Baró^l, Roberto Bellasio^d, Dominik Brunner^g, Charles Chemel^h, Gabriele
6 Curci^{ij}, Hugo Denier van der Gon^k, Johannes Flemming^m, Renate Forkelⁿ, Lea Giordano^g,
7 Pedro Jiménez-Guerrero^l, Marcus Hirtl^o, Oriol Jorba^e, Astrid Manders-Groot^k, Lucy Neal^p,
8 Juan L. Pérez^q, Guidio Pirovano^f, Roberto San Jose^q, Nicholas Savage^p, Wolfram Schroder^r,
9 Ranjeet S Sokhi^h, Dimiter Syrakov^s, Paolo Tuccella^{ij}, Johannes Werhahnⁿ, Ralf Wolke^r,
10 Christian Hogrefe^t, Stefano Galmarini^b

- 11
12 a. University of Patras, Department of Physics, University Campus 26504 Rio, Greece.
13 b. European Commission, Joint Research Centre, Directorate for Energy, Transport and Climate, Air
14 and Climate Unit, Ispra (VA), Italy.
15 c. Aarhus University, Department of Environmental Science, Roskilde, Denmark
16 d. Enviroware srl, Concorezzo (MB), Italy.
17 e. Earth Sciences Department, Barcelona Supercomputing Center (BSC-CNS), Barcelona, Spain.
18 f. Ricerca sul Sistema Energetico (RSE) SpA, Milan, Italy
19 g. Laboratory for Air Pollution and Environmental Technology, Empa, Dübendorf, Switzerland.
20 h. Centre for Atmospheric & Instrumentation Research, University of Hertfordshire, College Lane,
21 Hatfield, AL10 9AB, UK.
22 i. Department of Physical and Chemical Sciences, University of L'Aquila, L'Aquila, Italy.
23 j. Center of Excellence for the forecast of Severe Weather (CETEMPS), University of L'Aquila,
24 L'Aquila, Italy.
25 k. Netherlands Organization for Applied Scientific Research (TNO), Utrecht, The Netherlands.
26 l. University of Murcia, Department of Physics, Physics of the Earth. Campus de Espinardo, Ed.
27 CIOyN, 30100 Murcia, Spain.
28 m. ECMWF, Shinfield Park, RG2 9AX Reading, United Kingdom.
29 n. Karlsruher Institut für Technologie (KIT), IMK-IFU, Kreuzteckbahnstr. 19, 82467 Garmisch-
30 Partenkirchen, Germany.
31 o. Zentralanstalt für Meteorologie und Geodynamik, ZAMG, 1190 Wien, Austria.
32 p. Met Office, FitzRoy Road, Exeter, EX1 3PB, United Kingdom.
33 q. Environmental Software and Modelling Group, Computer Science School - Technical University
34 of Madrid, Campus de Montegancedo - Boadilla del Monte-28660, Madrid, Spain.
35 r. Leibniz Institute for Tropospheric Research, Permoserstr. 15, D-04318 Leipzig, Germany.
36 s. National Institute of Meteorology and Hydrology, Bulgarian Academy of Sciences, 66
37 Tzarigradsko shaussee Blvd., Sofia 1784, Bulgaria.
38 t. Atmospheric Modelling and Analysis Division, Environmental Protection Agency, Research
39 Triangle Park, USA.

40
41

1 **Abstract**

2 Simulations from chemical weather models are subject to uncertainties in the input data (e.g.
3 emission inventory, initial and boundary conditions) as well as those intrinsic to the model
4 (e.g. physical parameterization, chemical mechanism). Multi-model ensembles can improve
5 the forecast skill provided that certain mathematical conditions are fulfilled. In this work, four
6 ensemble methods were applied to two different datasets and their performance was compared
7 for ozone (O₃), nitrogen dioxide (NO₂) and particulate matter (PM₁₀). Apart from the
8 unconditional ensemble average, the approach behind the other three methods relies on
9 adding optimum weights to members or constraining the ensemble to those members that
10 meet certain conditions in time or frequency domain. The two different datasets were created
11 for the first and second phase of the Air Quality Model Evaluation International Initiative
12 (AQMEII). The methods are evaluated against ground level observations collected from the
13 EMEP and Airbase databases. The goal of the study is to quantify to what extent we can
14 extract predictable signals from an ensemble with superior skill over the single models and
15 the ensemble mean. Verification statistics shows that the deterministic models simulate better
16 O₃ than NO₂ and PM₁₀, linked to different levels of complexity in the represented processes.
17 The unconditional ensemble mean achieves higher skill compared to each station's best
18 deterministic model at no more than 60% of the sites, indicating for the rest a combination of
19 members with unbalanced skill difference and error dependence. The promotion of the right
20 amount of accuracy and diversity within the ensemble results in an average additional skill up
21 to 31% compared to using the full ensemble in an unconditional way. The skill improvements
22 were higher for O₃ and lower for PM₁₀, associated to the extent of potential changes in the
23 joint distribution of accuracy and diversity in the respective ensembles. The skill
24 enhancement was superior using the weighting scheme but the training period required to
25 acquire representative weights was longer compared to the sub-selecting schemes. Further
26 development of the method is discussed in the conclusion.

27

28 Keywords: AQMEII, multi-model ensembles, air quality model, error decomposition,
29 verification.

30

1 1 Introduction

2 Uncertainties in atmospheric models such as the chemical weather models, whether due to the
3 input data or the model itself, limit the predictive skill. The incorporation of data assimilation
4 techniques and the continued effort in understanding the physical, chemical and dynamical
5 processes, result in better forecasts (Zhang et al., 2012). In addition, ensemble methods
6 provide an extra channel for forecast improvement and uncertainty quantification. The
7 benefits from ensemble averaging arise from filtering out the components of the forecast with
8 uncorrelated errors (Kalnay, 2003).

9 The European Centre for Medium-Range Weather Forecast (ECMWF) reports an increase in
10 forecast skill of 1 day per decade for meteorological variables, evaluated on the geopotential
11 height anomaly (Simmons, 2011). The air quality modelling and monitoring has a shorter
12 history that does not allow a similar adequate estimation of such trend for the numerous
13 species being modelled. Moreover, the skill changes dramatically from species to species
14 strongly connected to the availability of accurate emission data. Results for ozone suggest that
15 medium-range forecasts can be performed with a quality similar to the geopotential height
16 anomaly forecasts (Eskes et al., 2002). Besides the continuous increase in skill due to the
17 improved scientific understanding, harmonized emission inventories, more accurate and
18 denser observations as well as ensemble averaging, an extra gain of similar magnitude can be
19 achieved for ensemble-based deterministic modelling using conditional averaging (e.g.,
20 Galmarini et al., 2013; Mallet et al., 2009; Solazzo et al., 2013).

21 Ideally, for continuous and unbiased variables, the multi-model ensemble mean outcores the
22 skill of the deterministic models provided that the members have similar skill and
23 independent errors (Potempski and Galmarini, 2009; Weigel et al., 2010). Practically, the
24 multi-model ensemble mean usually outcores the skill of the deterministic models if the
25 evaluation is performed over multiple observation sites and times. This occurs because over a
26 network of stations, there are some where the essential conditions (e.g. the skill difference
27 between the models is not too large) for the ensemble members are fulfilled, favouring the
28 ensemble mean; for the remaining stations, where the conditions are not fulfilled, local
29 verification identifies the best model but generally no single model is the best at all sites.
30 Hence, although the skill of the numerical models varies in space (latitude, longitude, altitude)
31 and time (e.g., hour of the day, month, season), the ensemble mean is usually the most
32 accurate spatio-temporal representation.

1 One of the challenges in multi-model ensemble forecasting is the processing of the
2 deterministic models datasets prior to averaging in order to construct another dataset for
3 which its members ideally constitute an *independent and identically distributed* (i.i.d.) sample
4 (Kioutsioukis and Galmarini, 2014; Bishop and Abramowitz, 2013). This statistical process
5 favours the ensemble mean at each observation site. Two basic pathways exist to achieve this
6 goal: model weighting or model sub-selecting. There are several methods to assign weights to
7 ensemble members such as the singular value decomposition (Pagowski et al., 2005),
8 dynamic linear regression (Pagowski et al., 2006; Djalalova et al., 2010), Kalman filtering
9 (Delle Monache et al., 2011), Bayesian model averaging (Riccio et al., 2007; Monteiro et al.,
10 2013) and analytical optimization (Potempski and Galmarini, 2009) while model selection
11 usually relies on the quadratic error or its proxies, in time (e.g. Solazzo et al., 2013;
12 Kioutsioukis and Galmarini., 2014) or frequency space (Galmarini et al., 2013). The majority
13 of those ensemble studies focuses on O₃ and only recently the studies also involve particulate
14 matter (Djalalova et al., 2010; Monteiro et al., 2013).

15 In this work, we apply and intercompare both approaches (weighting and sub-selecting) using
16 the Air Quality Model Evaluation International Initiative (AQMEII) datasets from phase I and
17 phase II. The ensemble approaches are evaluated against ground level observations from the
18 EMEP and Airbase databases, focusing on the pollutants O₃, NO₂ and PM₁₀ that exhibit
19 different levels of forecast skill. The differences between the multi-model ensembles of phase
20 I (hereafter AQMEII-I) and phase II (hereafter AQMEII-II) originate from many sources,
21 related to both the input data and the models: (a) the simulated years are different (2006 vs.
22 2010), therefore the meteorological conditions are different; (b) emission methodologies have
23 changed; (c) boundary conditions are very different; (d) the composition of the ensembles is
24 different; (e) the models in AQMEII-II use on-line coupling between meteorology and
25 chemistry; (f) the models may have been updated with new science processes apart from
26 feedback processes. The uncertainties arising from observational errors are not taken into
27 consideration.

28 In spite of these differences we consider the analysis of the two sets of ensembles revealing.
29 In detail, the objectives of the paper are (a) to interpret the skill of the unconditional multi-
30 model mean within AQMEII-I and AQMEII-II (b) to calculate the maximum expectations in
31 the skill of alternative ensemble estimators and (c) to evaluate the operational implementation
32 of the approaches using cross-validation. The originality of the study includes: (a) the

1 comparison of several ensemble methods on pollutants of different skill using different
 2 datasets, (b) the introduction of an approach based on high-dimension spectral optimization,
 3 (c) the introduction of innovative charts for the interpretation of the error of the unconditional
 4 ensemble mean with respect to indicators reflecting the skill difference and error dependence
 5 of the models as well as the effective number of models. Therefore we carry out an analysis of
 6 the performance of different ensemble techniques rather than a comparison of the results from
 7 the two phases of the AQMEII activity.

8 The paper is structured as follows: section 2 provides a brief description of the ensemble's
 9 basic properties through a series of conditions expressed by mathematical equations. In
 10 section 3, the experimental setup is described. Results are presented in section 4, where the
 11 skill of the deterministic models, the unconditional ensemble mean and the conditional
 12 ensemble estimators are analysed and intercompared. Conclusions are drawn in Section 5.

13 2 Minimization of the ensemble error

14 The notation conventions used in this section are briefly presented in the following. Assuming
 15 an ensemble composed of M members (i.e. output of modelling systems) denoted as f_i ,
 16 $i=1,2,\dots,M$, the multi-model ensemble mean can be evaluated from $\bar{f} = \sum_{i=1}^M w_i f_i$, $\sum w_i = 1$. The
 17 weights (w_i) sum up to one and can be either equal (uniform ensemble) or unequal
 18 (nonuniform ensemble). The desired value (measurement) is μ .

19 Assuming a uniform ensemble, the squared error (MSE) of the multi-model ensemble mean
 20 can be broken down into three components, namely, the average bias (1st term), the average
 21 error variance (2nd term) and the average error covariance (3rd term) of the ensemble members
 22 (Ueda and Nakano, 1996):

$$\begin{aligned}
 \text{MSE}(\bar{f}) = & \left(\frac{1}{M} \sum_{i=1}^M (f_i - \mu) \right)^2 + \frac{1}{M} \left(\frac{1}{M} \sum_{i=1}^M (f_i - \mu)^2 \right) \\
 & + \left(1 - \frac{1}{M} \right) \left(\frac{1}{M(M-1)} \sum_{i=1}^M \sum_{i \neq j} (f_i - \mu)(f_j - \mu) \right)
 \end{aligned}
 \tag{Eq.1}$$

23 The decomposition provides the reasoning behind ensemble averaging: as we include more
 24 ensemble members, the variance factor is monotonically decreasing and the MSE converges
 25 towards the covariance factor. Covariance, unlike the other two positive definite terms, can be

1 either positive or negative; its minimization requires an ensemble composed by independent
 2 or even better, negatively correlated members. In addition, bias correction should be a
 3 necessary step prior to any ensemble manipulation. More details regarding this decomposition
 4 within the air quality ensembles context can be found in Kioutsioukis and Galmarini, 2014.

5 In a similar fashion, the squared error of the multi-model ensemble mean can be decomposed
 6 into the difference of two positive-definite components, with their expectations characterized
 7 as accuracy and diversity (Krogh and Vedelsby, 1995):

$$MSE(\bar{f}) = \frac{1}{M} \sum_{i=1}^M (f_i - \mu)^2 - \frac{1}{M} \sum_{i=1}^M (f_i - \bar{f})^2 \quad \text{Eq.2}$$

8 This decomposition proves that the error of the ensemble mean is guaranteed to be less than
 9 or equal to the average quadratic error of the component models. The minimum ensemble
 10 error depends on the right trade-off between accuracy (1st term on the r.h.s. of Eq. 2) and
 11 diversity (2nd term on the r.h.s. of Eq. 2). If the evaluation is applied on multiple sites, then the
 12 equations 1 and 2 should be replaced with their expectations over the stations.

13 An error decomposition approach can also be applied on the spectral components (SC) of the
 14 observed and modelled time-series. The data can be spectrally decomposed with the
 15 Kolmogorov-Zurbenko (*kz*) filter (Zurbenko, 1986) while the original time-series can be
 16 obtained with the linear combination of the spectral components. Assuming the pollution data
 17 at the frequency domain yields *N* principal spectral bands, the squared error of the multi-
 18 model ensemble mean can be broken down into *N*² components (Galmarini et al., 2013;
 19 Solazzo and Galmarini, 2016):

$$MSE(\bar{f}) = \sum_{i=1}^N MSE(SC_{\bar{f}_i}) + \sum_{i \neq j} Cov(SC_{\bar{f}_i}, SC_{\bar{f}_j}) \quad \text{Eq.3}$$

20 This decomposition shows that the error of the ensemble mean could be split into the sum of
 21 *N* errors associated with different parts of the spectrum (1st term), provided the spectral
 22 components are independent (the covariance term is zero). The minimization of the error at
 23 each spectral band can be achieved with another approach such as the decompositions
 24 presented in Eq.1 and Eq.2.

1 The three decompositions presented assume uniform ensembles, i.e. all members receive
 2 equal weight. For the case of a non-uniform ensemble, the MSE of the multi-model ensemble
 3 mean can be analytically minimized to yield the optimal weights, provided that the
 4 participating models are bias-corrected (Potempski and Galmarini, 2009):

$$\bar{\mathbf{w}} = \frac{\mathbf{K}^{-1}\mathbf{l}}{(\mathbf{K}^{-1}\mathbf{l}, \mathbf{l})} \quad \text{Eq.4}$$

5 where, \mathbf{w} is the vector of optimal weights, \mathbf{K} is the error covariance matrix and \mathbf{l} the unitary
 6 vector. In its simplest form, the equation assigns one weight for each model at each
 7 measurement site; more complicated versions like multidimensional optimisation for many
 8 variables (e.g. chemical compounds) at many sites simultaneously are not discussed here.

9 Unlike the straightforward calculation of the optimal weights, the sub-selecting schemes make
 10 use of a reduced-dimensionality ensemble. An estimate of the effective number of models
 11 (N_{EFF}) sufficient to reproduce the variability of the full ensemble is calculated as (Bretherton
 12 et al., 1999):

$$N_{EFF} = \frac{(\sum_{i=1}^M s_i)^2}{\sum_{i=1}^M s_i^2} \quad \text{Eq.5}$$

13 where s_i is eigenvalue of the error covariance matrix. Theoretical evidence shows that the
 14 fraction of the overall variance expressed by the first N_{EFF} eigenvalues is 86%, provided that
 15 the modelled and observed fields are normally distributed (Bretherton et al., 1999). The
 16 highest eigenvalue is denoted as s_m .

17 It is apparent from the above considerations that the skill of the unconditional ensemble mean
 18 has the potential for certain advantages over the single members, provided some properties
 19 are satisfied. As those properties are not systematically met in practice, superior ensemble
 20 skill can be achieved through sub-selecting or weighting schemes presented in this section.

21 An inter-comparison of the following approaches in ensemble averaging is investigated in this
 22 work using observed and simulated air quality time-series:

- 23 • Unconditional ensemble mean (*mme*)
- 24 • Conditional (on selected members) ensemble mean in time domain (*mme<*): the
 25 optimal trade-off between accuracy and diversity (equation 2) is identified across all
 26 possible combinations of the available M models (Kioutsioukis and Galmarini, 2014).

- 1 The number of members in the ensemble combination that gives the minimum error
2 will be used as the effective number of models (N_{EFF}) rather than its estimate based on
3 the independent components of the ensemble (eq. 5).
- 4 • Conditional (on selected members) ensemble mean in frequency domain (*kzFO*):
5 following equation 3, an ensemble estimator is synthesized from the best member at
6 each spectral band (Galmarini et al., 2013). The original time-series are decomposed
7 into four spectral components (see Appendix I), namely the intra-diurnal, diurnal,
8 synoptic and long-term component, using the Kolmogorov-Zurbenko filter (Zurbenko,
9 1986).
 - 10 • Conditional (on selected members) ensemble mean in frequency domain (*kzHO*): it is
11 an extension of the *kzFO*, where the spectral components of the ensemble estimator
12 are averaged from N_{EFF} members at each spectral band (rather than the best).
 - 13 • Conditional (optimally weighted) ensemble mean (*mmW*): according to equation 4
14 (Potemski and Galmarini, 2009).

15 The skill of the models and the examined ensemble averages have been scored with the
16 following statistical parameters: (1) normalised mean square error (NMSE), i.e. the mean
17 square error (MSE) divided by $\bar{O}\bar{M}$, where \bar{O} and \bar{M} are the mean value of the observation and
18 the model respectively, (2) probability of detection (POD) and false alarm rate (FAR), i.e. the
19 proportion of occurrences (e.g. events exceeding threshold value) that were correctly
20 identified and the proportion of non-occurrences that were incorrectly identified respectively
21 (3) Taylor plots (Taylor, 2001), which summarize standard deviation, root mean square error
22 (RMSE) and Pearson product-moment correlation coefficient in a single point on a two-
23 dimensional plot.

24 **3 Setup: experiments, models and observations**

25 The two AQMEII ensemble datasets have simulated the air quality for Europe [(-10,39)W;
26 (30,65)N] and North America [(-125,-55)W; (26,51)N]. Despite the common domains, the
27 modelling systems across the two phases have profound differences. The simulation year was
28 2006 for AQMEII-I and 2010 for AQMEII-II, therefore the two sets are dissimilar with
29 respect to the input data (emissions, chemical boundary conditions, meteorology). Boundary
30 conditions are obtained from GEMS (Global and Regional Earth-System Monitoring using
31 Satellite and in-situ data) in AQMEII-I and MACC (Monitoring Atmospheric Composition &

1 Climate) in AQMEII-II. The air quality models of the second phase are coupled with their
2 meteorological driver (chemistry feedbacks on meteorology), while those of the first phase
3 are not. The participating models are also different. Detailed analysis of the emissions,
4 boundary conditions and meteorology for the modelled year 2006 (AQMEII-I) is presented in
5 Pouliot et al. (2012), Schere et al. (2012) and Vautard et al. (2012). For 2010 (AQMEII-II),
6 similar information is presented in Pouliot et al. (2015), Giordano et al. (2015) and Brunner et
7 al. (2015).

8 The participating models follow a restrictive protocol concerning the emissions and the
9 meteorological and chemical boundary conditions. In AQMEII-I, meteorological models
10 applied nudging to the NCEP GFS meteorological analysis. In AQMEII-II, the simulations
11 were run more in a way as if they were real forecasts; meteorological boundary conditions for
12 the majority of the models were from the ECMWF operational archive (see Tables 1 and 2 in
13 Brunner et al, 2015) and no nudging or FDDA was applied. However, the driving
14 meteorological data were analysis (but no reanalysis) for all simulations, with exception of the
15 COSMO-MUSCAT run. Hence, the runs from AQMEII-II are more like forecasts than those
16 from AQMEII-I.

17 Recent studies with regional air quality models yielded that the full variability of the
18 ensemble can be retained with only an effective number of models (N_{EFF}) on the order of 5-6
19 (e.g. Solazzo et al., 2013; Kioutsioukis and Galmarini, 2014; Marecal et al., 2015). The
20 minimum number of ensemble members to sample the uncertainty should be well above N_{EFF} ;
21 for this reason, we focus on the European domain (EU) due to its sufficient number of models
22 to form the ensemble.

23 Table 1 summarises the features of the modelling systems analysed in this study with regard to
24 O_3 , NO_2 and PM_{10} concentrations in the EU. The modelling contribution to the two phases of
25 AQMEII consists of 12, 13 and 10 models for O_3 , NO_2 and PM_{10} respectively in AQMEII-I,
26 while 14 members were available for all species in AQMEII-II. Several discrete simulations
27 of WRF-Chem with alternative chemistry and physics configurations are included in
28 AQMEII-II (Forkel et al. 2015, San José et al, 2015, Baró et al., 2015).

29 Following the statements of section 2, each model has been bias-corrected prior to the
30 analysis, i.e. its own mean bias over the examined three-month period has been subtracted
31 from its modelled time-series at each monitoring site. For each modelling system, its long-
32 term systematic error is a known quantity estimated during its validation stage; therefore the

1 subtraction of the seasonal bias does not restrict the generality of the study. Actually, the
2 requirement for bias removal is a necessary condition only for the weighted ensemble mean.
3 In the results section we will address this issue and its effect on the skill of the ensemble
4 estimators.

5 The observational data sets for O_3 , NO_2 and PM_{10} derived from the surface AQ monitoring
6 networks operating in the EU constitutes the same data set used in the first and second phases
7 of AQMEII to support model evaluation. All monitoring stations are rural and have data at
8 least 75% of the time. The network is denser for O_3 (451/450 stations in AQMEII-I/II) for
9 which there are as many monitoring stations as for NO_2 (290/337 stations in AQMEII-I/II)
10 and PM_{10} (126/131 stations in AQMEII-I/II) combined, with PM_{10} having the fewest
11 observations. **Figure 1** compares the statistical distribution of all three species between the two
12 AQMEII phases, through the cumulative density function composed from the mean value at
13 each percentile of the observations. The Kolmogorov-Smirnov test (Massey, 1951) yields that
14 only the PM_{10} distributions differ at the 1% significance level. It results from the
15 unavailability of data for France and UK in AQMEII-II for PM_{10} (station locations are shown
16 in **Figure 3**).

17 **4 Results**

18 In this section we apply the conceptual context briefly presented in section 2 to investigate the
19 effect of the differences in the ensemble properties within each of the two AQMEII phases
20 (Rao et al., 2011) in the skill of the unconditional multi-model mean. The potential for
21 improved estimates through conditional ensemble averages and their robustness is ultimately
22 assessed.

23 From the provided station-based hourly time-series, we analysed one season (three-monthly
24 period) with continuous data and relatively high concentrations; for O_3 , June-July-August was
25 selected while September-October-November is used for NO_2 and PM_{10} .

26 **4.1 Single Models**

27 The distributions of each model's NMSE for O_3 , NO_2 and PM_{10} over all monitoring stations
28 are presented in Figure 2 as box-and-whisker plots. On each box, the central mark indicates the
29 median, and the bottom and top edges of the box indicate the 25th and 75th percentiles,

1 respectively. The whiskers extend to the most extreme data points not considered outliers (i.e.
2 points with distance from the 25th and 75th percentiles smaller than 1.5 times the interquartile
3 range). Among the examined pollutants, the models simulate better the O₃ concentrations, as
4 is evident from the axis scale. The highest variability in the skill between and within the
5 models is observed for NO₂.

6 The distribution of average NMSE at each station ($\langle \text{NMSE} \rangle$) has a median on the order of
7 0.1 for O₃ and 0.5 for NO₂ and PM₁₀ for both phases (Table 2). The application of the
8 Kolmogorov-Smirnov test (Massey, 1951) on the $\langle \text{NMSE} \rangle$ distributions across AQMEII-I
9 and AQMEII-II shows that there are no statistically significant differences in the $\langle \text{NMSE} \rangle$
10 distributions between the two ensemble datasets at the 1% significance level. The same also
11 applies for the statistical distribution of the minimum NMSE at each station (NMSE_{BEST}) at
12 each monitoring station. Hence, despite the different modelling systems and input data, the
13 $\langle \text{NMSE} \rangle$ and NMSE_{BEST} distributions between AQMEII-I and AQMEII-II are
14 indistinguishable for the three examined pollutants.

15 Besides $\langle \text{NMSE} \rangle$ and NMSE_{BEST}, we evaluate the percentage of cases each model has been
16 identified as being ‘best’ and calculate the coefficient of variation ($CoV = \text{std}/\text{mean}$) of this
17 index for each ensemble. If models were behaving like *i.i.d.*, the probabilities of being best
18 would be roughly equal ($\sim 1/M$) for all models and the CoV would generally be well below
19 unity for the examined range of ensemble members. As can be inferred from Table 2, the
20 proportion of *equally good models* is higher for O₃ and NO₂ in the 2nd dataset. Among the
21 pollutants, the CoV of NO₂ exhibits the most dramatic change.

22 4.2 Pitfalls of the unconditional multi-model mean

23 The skill of the multi-model mean has been compared against the skill of the best
24 deterministic model, independently evaluated at each monitoring site (hereafter *bestL*). The
25 geographical distribution of the ratio $\text{RMSE}(mme)/\text{RMSE}_{\text{BESTMODEL}}$ is presented in Figure 3.
26 The indicator does not exhibit any longitudinal or latitudinal dependence. Summary statistics
27 indicate that the *mme* outcores the *bestL* at roughly half of the stations for O₃ (namely 52/49
28 for AQMEII-I/II) and at approximately 40% of the stations for PM₁₀ (38/42). The same
29 statistic for NO₂ varies considerably (39/64). The Kolmogorov-Smirnov test yields that the
30 corresponding distributions (pI/pII) are different at the 1% significance level but the t-test
31 demonstrates that the mean of the distributions differ significantly only for NO₂. The reason

1 behind the skill of *mme* with respect to the *bestL* is investigated next with respect to the skill
2 difference and the error dependence of each ensemble.

3 The skill difference between the best model and the average skill is inferred from the
4 indicator $NMSE_{BEST} / \langle NMSE \rangle$ (Table 2). High values of the indicator correspond to small
5 skill differences between the ensemble members (desirable). The distribution of the
6 $NMSE_{BEST} / \langle NMSE \rangle$ at each station has a median on the order of 0.6-0.8, variable with
7 respect to the dataset and the pollutant. The spread of the indicator, measured by its
8 interquartile range, is higher for NO_2 and lower for O_3 .

9 The eigenvalues of the covariance matrix calculated from the model errors provides
10 information on the members' diversity and the ensemble redundancy (Eq. 5). Following the
11 eigen-analysis of the error covariance matrix at each station separately and converting the
12 eigenvalues to cumulative amount of explained variance, the resulting matrix is presented into
13 box and whisker plot (Figure 4). The error dependence of the ensemble members is deduced
14 from the explained variation by the maximum eigenvalue s_m . Low values of the indicator
15 corresponds to independent members with small error dependence (desirable). The average
16 variation explained by s_m ranges between 65% and 79%, taking the lower values for NO_2 . The
17 spread of the indicator, measured by its interquartile range, is higher for NO_2 and lower for
18 O_3 .

19 All species demonstrate smaller skill difference and higher error dependence in the AQMEII-
20 II dataset. The Kolmogorov-Smirnov test yielded the difference in the corresponding
21 distributions of the indicators between AQMEII-I and AQMEII-II is significant at the 1%
22 level. However, it is the joint distribution of skill difference and error dependence that
23 modulates the *mme* skill with respect to the *bestL*, as seen in Figure 5. Shifts in the
24 distributions of the indicators at opposite directions eventually cancel out, yielding no change
25 in the *mme* skill. This case is observed for O_3 and PM_{10} . For NO_2 , skill difference was
26 improved more than error dependence was worsened, yielding a net improvement of *mme* in
27 AQMEII-II.

28 The area below the diagonal in Figure 5 corresponds to monitoring sites with disproportionately
29 low diversity under the current level of accuracy. This area of the chart indicates high spread
30 in skill difference and relatively highly dependent errors. This situation practically means a
31 limited number of skilled models with correlated errors, which in turn denotes a small N_{EFF}
32 value as demonstrated in Figure 6. The opposite state is true for the area above the diagonal. It

1 corresponds to locations that are constituted from models with comparable skill and relatively
2 independent errors, reflecting a high N_{EFF} value. This matches the desired synthesis for an
3 ensemble.

4 The cumulative distribution of N_{EFF} from the error minimization (i.e. the optimal trade-off
5 between accuracy and diversity) across all possible combinations of M models at each site is
6 also presented in Figure 4 (solid line). At over 90% of the stations, we do not need more than 5
7 members for O_3 , 6 members for PM_{10} and 6-7 members for NO_2 . Further, from a pool of 10-
8 14 models, the benefits of ensemble averaging cease after 5-7 members (but not 5-7 particular
9 members across all stations).

10 4.3 Conditional multi-model mean

11 Following the identification of the weaknesses in the ensemble design, the potential for
12 corrections through more sophisticated schemes is now investigated. We consider the skill of
13 the multi model mean as the starting point and we investigate pathways for further enhancing
14 it through the non-trivial problem of weighting or sub-selecting. The optimal weights (mmW)
15 are estimated from the analytical formulas presented in Potempski and Galmarini, 2009. The
16 sub-selection of members has been built upon the optimization of either the accuracy/diversity
17 trade-off ($mme<$) (Kioutsioukis and Galmarini, 2014) or the spectral representation of 1st
18 order components by different models ($kzFO$) (Galmarini et al., 2013). Another approach
19 built upon higher order (namely, N_{EFF}) spectral components ($kzHO$) is also investigated. In
20 this section we mark the boundaries of the possible improvements for different ensemble
21 mean estimators applicable to the AQMEII datasets and their sensitivity to sub-optimal
22 conditions using cross-validation.

23 The global skill of all the single models and the ensemble estimators, evaluated at all stations,
24 are presented in Figure 7 in the form of Taylor plots. For O_3 , the deterministic models have
25 standard deviations that are smaller compared to observations and a narrow correlation pattern
26 (~ 0.7) that is slightly deteriorated in AQMEII-II. For NO_2 , members with higher variance -as
27 well as lower- than the observed variance exist in the ensemble while the correlation spread is
28 becoming narrower in AQMEII-II and demonstrates a minor improvement. Last, simulated
29 PM_{10} from the deterministic models displays smaller standard deviation compared to
30 observations with a wide correlation spread (0.3-0.6). The multi-model mean is always found
31 closer to the reference point, in an area that incorporates lower error and increased correlation

1 but at the same time generally low variance. The examined ensemble estimators (*mmW*,
2 *mme<*, *kzFO*, *kzHO*) are horizontally shifted from *mme*, hence they demonstrate even lower
3 error and increased correlation and variance. Among them, the highest composite skill was
4 found for *mmW*, followed by *kzHO*.

5 A comparison between the skill of the examined ensemble estimators versus the *mme* and the
6 best single model is now conducted (Table 3). The best single model is evaluated globally
7 (*bestG*: average across all stations) and locally (*bestL*: at each station separately). The former
8 estimates the best average deterministic skill among the candidate models; the latter provides
9 a useful indicator for controlling whether the anticipated benefits of ensemble averaging
10 holds. The skill scores have been evaluated against the guaranteed minimum gain of the
11 ensemble ($\langle \text{MSE} \rangle$), the ensemble mean (*mme*) and the best single model globally (*bestG*).
12 The estimations calculated from the unprecedented AQMEII datasets (2 years of hourly
13 measurements and simulations from 2 different ensembles of 10-14 models each at over 450
14 stations for 3 pollutants) allows the following interpretation:

- 15 - The *mme* always achieves lower error than *bestG*. The advancement is higher for O₃
16 (9-22%), followed by NO₂ (7-9%) while the PM₁₀ demonstrate the least skill
17 improvement (1-3%). With respect to *bestL*, the *mme* generally attains similar or
18 slightly higher MSE. Hence, the average error over multiple stations statistically
19 favours the ensemble mean over the single models but the comparison at each site
20 generally does not as it depends on the skill difference and the error dependence of the
21 models.
- 22 - The skill score of *mme* over $\langle \text{MSE} \rangle$ (i.e., the guaranteed upper ceiling for the MSE of
23 *mme*, from eq. 2) ranges between 15% and 30%, higher for NO₂ and lower for PM₁₀.
24 According to eq. 2, this number also represents the diversity as percentage of the
25 accuracy. Therefore, besides improving the single models, their combination in an
26 ensemble confines the *mme* skill if their diversity is limited.
- 27 - The skill score of the examined ensemble estimators (*mmW*, *mme<*, *kzFO*, *kzHO*) over
28 $\langle \text{MSE} \rangle$ ranges between 25% and 50%, higher for O₃ and NO₂ and lower for PM₁₀.
29 Among them, the improvement is higher for *mmW* and lower for *mme<* and *kzFO*.
30 Thus, the promotion of accuracy and diversity within the ensemble almost doubles the
31 distance to $\langle \text{MSE} \rangle$ compared to *mme* and results in an additional skill over the *mme*
32 between 14% and 31% (for *mmW*).

- 1 - The improvement of the ensemble estimator using selected N_{EFF} members ($mme<$)
- 2 over all members (mme) is illustrated in Figure 8 in the context of skill difference and
- 3 error dependence. The charts demonstrate no points below the diagonal, i.e. the sub-
- 4 selection results in an ensemble constituted from models with comparable skill and
- 5 relatively independent errors (compared to the full ensemble).
- 6 - The theoretical minimum MSE of mme for the case of unbiased and uncorrelated
- 7 models (from eq. 1) is far from being achieved from all ensemble estimators.

8 The statistical distributions of the skill scores of the examined ensemble estimators (mmW ,

9 $mme<$, $kzFO$, $kzHO$) over mme are well bounded from above to lower than unity values

10 (Figure 9). The only exception exists for roughly 10% of the stations, for all pollutants, where

11 $kzFO$ demonstrates higher MSE compared to mme . Unlike the other ensemble estimators,

12 $kzFO$ utilises independent spectral components each obtained from a single model,

13 eliminating the possibility for ‘cancelling out’ of random errors. All cases belonging to this

14 10% of the samples (lower tail of the cdf) demonstrate high N_{EFF} , where the benefits from

15 unconditional ensemble averaging are optimal (Kioutsioukis and Galmarini, 2014). Contrary,

16 for another 10% of the stations (upper tail of the cdf), there is an abrupt improvement from

17 the conditional ensemble estimators. Those cases demonstrate low N_{EFF} , where the benefits

18 from unconditional ensemble averaging are minimal.

19 The ability to simulate extreme values is now examined through the POD and FAR indices.

20 Two thresholds were utilised for each pollutant, being 120 and 180 $\mu\text{g}/\text{m}^3$ for O_3 , 25 and 50

21 $\mu\text{g}/\text{m}^3$ for NO_2 and 50 and 90 $\mu\text{g}/\text{m}^3$ for PM_{10} . The average 90th percentile across the stations

22 was 129/117 $\mu\text{g}/\text{m}^3$ (AQMEII-I/II) for O_3 , 30/26 $\mu\text{g}/\text{m}^3$ for NO_2 and 52/33 $\mu\text{g}/\text{m}^3$ for PM_{10}

23 (Figure 1). Hence, the thresholds fall into the upper 10% of the distributions, being even more

24 extreme for PM_{10} in AQMEII-II. The numbers in Table 4 give rise to the following inferences:

- 25 - for O_3 and NO_2 , mme achieves somewhat higher POD than $bestG$ at the lower
- 26 threshold but the order is reversed at the higher threshold. For PM_{10} , $bestG$ always
- 27 performs better than mme for values exceeding the lower threshold. As we move
- 28 towards the tail, the POD of $bestG$ dominates over the mme . Thus, the ranking of the
- 29 mme and $bestG$ at the extreme percentiles and on average (seen earlier) are opposite.
- 30 - The $mme<$ generally achieves somewhat higher POD than $bestL$ at the lower threshold
- 31 but the order is reversed at the higher threshold. Over that level, $kzFO$ and mmW are
- 32 the only estimators with POD higher than $bestL$.

- 1 - As we move towards higher percentiles, the 1st order spectral model (*kzFO*) has higher
2 POD than the higher-order spectral model (*kzHO*) due to the averaging in the latter. In
3 addition, the frequency domain averaging (*kzHO*) had slightly higher POD compared
4 to the time domain averaging (*mme<*).
- 5 - The *mmW*, besides its lower MSE, has the highest POD among all models and
6 ensemble estimators.
- 7 - The variation of FAR was very small between all examined models and ensemble
8 estimators.

9 The combination of the results from the average error and the extremes identifies *mmW* as the
10 estimator that outscores the others across all percentiles. *kzFO* has high capacity for extremes
11 but requires attention for the limited sites with high N_{EFF} , where its skill is inferior to *mme*.
12 *kzHO* and *mme<* have both high skill across all percentiles (better for *kzHO*) but they could
13 have inferior POD compared to *bestL* at the extreme percentiles. With respect to the
14 pollutants, the advancement of *mmW* skill over *mme* was higher for O_3 .

15 The additional skill over *mme* in the range between 8% and 31% from the statistical
16 approaches applied to a pool of ensemble simulations identifies the upper ceiling of the
17 improvements from the corrections in the skill difference and the error dependence of the
18 ensemble members. The bound results from the removal of the seasonal bias from the time
19 series and the optimal training of the methods. We now proceed with splitting the datasets
20 into training and testing and explore the sensitivity of the *mmW* skill arising from improper
21 bias removal and weights. Both factors are estimated on the training set for variable time-
22 series length that is progressively increasing from 1 to 60 days, for all monitoring stations and
23 pollutants. The evaluation period for all training windows is the same 30-day segment, not
24 available in the training procedure. The analysis will provide a perspective on applying the
25 techniques in a forecasting context, although the examined simulations did not operate in
26 forecasting mode.

27 The interquartile range of the day-to-day difference in the weights is calculated and its range
28 over all stations is displayed in Figure 10. No convergence occurs, however the variability of
29 the *mmW* weights is notably reduced after a certain amount of time. If we set a tolerance level
30 at the second decimal, to be satisfied at all stations, we need at a minimum 20-45 days of
31 hourly time-series. The variability of weights is smaller for O_3 and higher for NO_2 and PM_{10} ,
32 explained by the larger NMSE spread in the latter case. The identification of the necessary

1 training or learning period will be assessed by its effect on the *mmW* skill. Table 5 presents the
2 *mmW* skill obtained from training over time series of different lengths varying from 5 to 60
3 days. For O₃, *mmW* trained over 10 days yields similar results with *mme* while longer periods
4 result in large departures from *mme*. NO₂ and PM₁₀ require larger training periods than O₃.
5 The use of *mmW* is practically of no benefit compared to *mme* if the training period is less than
6 20 days for NO₂ and 30 days for PM₁₀. For all pollutants, the variability of the weights and
7 the bias has no effect in the error after 60 days.

8 The results demonstrate that the ensemble estimators based on the analytical optimization
9 become insensitive to inaccuracies in the bias and weights for training periods exceeding 60
10 days. Other published studies with weighted ensembles using non-analytical optimization
11 though (e.g. linear regression, Monteiro et al., 2012), argue that one month is sufficient for the
12 weights and the bias. The sub-selecting schemes are more robust compared to the optimal
13 weighting scheme in the variations of their parameters (bias, members). Using data from
14 AQMEII-I, training periods in the order of a week were found essential for *mme*<
15 (Kioutsioukis and Galmarini, 2014) and *kzFO* (Galmarini et al., 2013). Therefore, the
16 operational implementation of each ensemble approach requires knowledge of its safety
17 margins for the examined pollutants.

18 5 Conclusions

19 In this paper we analyze two independent suites of chemical weather modelling systems
20 regarding their effect in the skill of the ensemble mean (*mme*). The results are interpreted with
21 respect to the error decomposition of the *mme*. Four ways to extract more information from an
22 ensemble besides the *mme* are ultimately investigated and evaluated. The first approach
23 applies optimal weights to the models of the ensemble (*mmW*) and the other three methods
24 utilise selected members in time (*mme*<) or frequency (*kzFO*, *kzHO*) domain. The study
25 focuses on O₃, NO₂ and PM₁₀, using the unprecedented datasets from two phases of AQMEII
26 over the European domain.

27 The comparison of the *mme* skill versus the globally best single model (*bestG*: identified from
28 the evaluation over all stations), points out that *mme* achieves lower average (across all
29 stations) error compared to *bestG*. The enhancement of accuracy is highest for O₃ (up to 22%)
30 and lowest for PM₁₀ (below 3%). We then investigate whether this benefit of ensemble
31 averaging of air quality time series holds at each station by direct comparison between the

1 *mme* and the locally best single model (*bestL*: identified from the evaluation at each station).
2 Summary statistics indicate that the *mme* outscores the *bestL* at roughly 50% of the stations
3 for O₃ and at approximately 40% of the stations for PM₁₀, while for NO₂ the values were
4 about 40% and 60% for the two datasets. This result indicates that there is a considerable
5 amount of stations (over 40%) where the unconditional averaging is not advantageous
6 because the ensemble does not meet the necessary conditions. A new chart has been
7 introduced in this paper that interprets the skill of the *mme* according to the skill difference
8 and the error dependence of the ensemble members.

9 The four examined ensemble estimators are then assessed for their skill in the average error as
10 well as their capability to correctly identify extreme values (events exceeding threshold
11 value). The key results of the analysis are summarized below:

- 12 - The skill score of *mme* over its guaranteed upper ceiling (case of zero diversity) ranges
13 between 15% and 30%, being lower for PM₁₀. Those percentages also represent the
14 diversity normalized by the accuracy. Therefore, besides improving the single models,
15 their combination in an ensemble confines the *mme* skill if their diversity is limited.
- 16 - The promotion of the right amount of accuracy and diversity in the conditional
17 ensemble estimators almost doubles the distance to the guaranteed upper ceiling. The
18 skill score over *mme* is higher for O₃ (in the range 18%-31%) and lower for NO₂ and
19 PM₁₀ (in the range 8%-25%), associated to the extent of potential changes in the joint
20 distribution of accuracy and diversity in the respective ensembles. The improvement is
21 larger for *mmW* and smaller for *mme<* and *kzFO*.
- 22 - The theoretical minimum MSE of *mme* for the case of unbiased and uncorrelated
23 models is far from being achieved from all ensemble estimators.
- 24 - As we move towards the tail, the probability of detection (POD) of *bestG* (*bestL*)
25 dominates over the *mme* (*mme<*). At the extreme percentiles, *kzFO* and *mmW* are the
26 only estimators with POD higher than *bestL*.
- 27 - The combination of the results from the average error and the extremes identifies
28 *mmW* as the estimator that outscores the others across all percentiles. *kzFO* has high
29 capacity for extremes but requires attention for the limited sites with high N_{EFF}, where
30 its skill is inferior to *mme*. *kzHO* and *mme<* have both high skill across all percentiles
31 (better for *kzHO*) but they could have inferior POD compared to *bestL* at the extreme
32 percentiles.

1 The skill enhancement is superior using the weighting scheme but the required training period
2 to acquire representative weights was longer compared to the sub-selecting schemes. For all
3 pollutants, the variability of the weights and the bias has negligible effect in the error for
4 training periods longer than 60 days. For the schemes relying in member selection, accurate
5 recent representations on the order of a week were sufficient. The learning periods constitute
6 the necessary time to acquire similar prior and posterior distributions in the controlling
7 parameters of samples. The risks of all the statistical learning processes originate from the
8 violation of this assumption, which holds for example in the case of changing weather or
9 chemical regimes. Therefore, the operational implementation of each ensemble approach
10 requires knowledge of its safety margins for the examined pollutants as well as its risks.

11 The improvement of the physical, chemical and dynamical processes in the deterministic
12 models is a continuous procedure that results in better forecasts. Besides that, mathematical
13 optimizations in the input data (e.g. data assimilation) or the model output (e.g. ensemble
14 estimators) have a significant contribution in the accuracy of the whole modelling process.
15 The presented post-simulation advancements were the result of only favourable ensemble
16 design. However, the theoretical minimum MSE of *mme* for the case of unbiased and
17 uncorrelated models is far from being achieved from all ensemble estimators. Further
18 development is underway in the presented ensemble methods that take into account the
19 meteorological and chemical regimes.

20

21

1 **References**

- 2 Baró, R., P. Jiménez-Guerrero, A. Balzarini , G. Curci , R. Forkel, M. Hirtl , L. Honzak, U.
3 Im, C. Lorenz, J.L. Pérez, G. Pirovano, R. San José; P. Tuccella, J. Werhahn, R. Žabkar:
4 Sensitivity analysis of the microphysics scheme in WRF-Chem contributions to AQMEII
5 phase 2, *Atmospheric Environment* 715: 620-629, 2015.
- 6 Bishop, C.H., Abramowitz, G.: Climate model dependence and the replicate earth paradigm.
7 *Clim Dyn* 41(3–4): 885–900, 2013.
- 8 Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M., Bladè, I.: The effective
9 number of spatial degrees of freedom of a time-varying field. *J. Climate* 12(7): 1990-2009,
10 1999.
- 11 Brunner, D., Jorba, O., Savage, N., Eder, B., Makar, P., Giordano, L., Badia, A., Balzarini,
12 A., Baro, R., Bianconi, R., Chemel, C., Forkel, R., Jimenez-Guerrero, P., Hirtl, M., Hodzic,
13 A., Honzak, L., Im, U., Knote, C., Kuenen, J.J.P., Makar, P.A., Manders-Groot, A., Neal, L.,
14 Perez, J.L., Pirovano, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D.,
15 Torian, A., Werhahn, K., Wolke, R., van Meijgaard, E., Yahya, K., Zabkar, R., Zhang, Y.,
16 Zhang, J., Hogrefe, C., Galmarini, S.: Comparative analysis of meteorological performance of
17 coupled chemistry-meteorology models in the context of AQMEII phase 2. *Atmospheric*
18 *Environment* 115: 470-498, 2015.
- 19 Delle Monache, L., T. Nipen, Y. Liu, G. Roux, Stull, R.: Kalman filter and analog schemes to
20 postprocess numerical weather predictions. *Month. Wea. Rev.* 139: 3554-3570, 2011.
- 21 Djalalova, I, J Wilczak, S McKeen, G Grell, S Peckham, M Pagowski, L DelleMonache, J
22 McQueen, Y Tang, P Lee, J McHenry, W Gong, V Bouchet, Mathur, R.: Ensemble and bias-
23 correction techniques for air quality model forecasts of surface O₃ and PM_{2.5} during the
24 TEXAQS-II experiment of 2006. *Atmos. Environ.* 44 (4): 455-467, 2010.
- 25 Eskes, H., van Velthoven, F., Kedler H.: Global ozone forecasting based on ERS-2 GOME
26 observations. *Atmos. Chem. Phys.* 2: 271-278, 2002.
- 27 Forkel R, A. Balzarini, R. Baró, R. Bianconi, G. Curci, P. Jiménez-Guerrero, M. Hirtl, Luka
28 Honzak, C. Lorenz, U. Im, J.L. Pérez, G. Pirovano, R. San José, P. Tuccella, J. Werhahn, R.
29 Žabkar. Analysis of the WRF-Chem contributions to AQMEII phase2 with respect to aerosol
30 radiative feedbacks on meteorology and pollutant distributions, *Atmospheric Environment*

- 1 115: 630–645, 2015.
- 2 Galmarini, S., Kioutsioukis, I., and Solazzo, E.: E pluribus unum: ensemble air quality
3 predictions, *Atmos. Chem. Phys.* 13: 7153–7182, 2013.
- 4 Giordano, L., D. Brunner, J. Flemming, C. Hogrefe, U. Im, R. Bianconi, A. Badia, A.
5 Balzarini, R. Baró, C. Chemel, G. Curci, R. Forkel, P. Jiménez-Guerrero, M. Hirtl, A. Hodzic,
6 L. Honzak, O. Jorba, C. Knote, J.J.P. Kuenen, P.A. Makar, A. Manders-Groot, L. Neal, J.L.
7 Pérez, G. Pirovano, G. Pouliot, R. San José, N. Savage, W. Schröder, R.S. Sokhi, D. Syrakov,
8 A. Torian, P. Tuccella, J. Werhahn, R. Wolke, K. Yahya, R. Žabkar, Y. Zhang, S. Galmarini.
9 Assessment of the MACC reanalysis and its influence as chemical boundary conditions for
10 regional air quality modeling in AQMEII-2, *Atmospheric Environment* 115: 371–388, 2015.
- 11 Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio,
12 R., Brunner, D., Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jimenez-
13 Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Kuenen, J.J.P., Makar,
14 P.A., Manders-Groot, A., Neal, L., Perez, J.L., Piravano, G., Pouliot, G., San Jose, R.,
15 Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R.,
16 Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S.: Evaluation of
17 operational online-coupled regional air quality models over Europe and North America in the
18 context of AQMEII phase 2. Part I: Ozone. *Atmospheric Environment* 115: 404-420, 2015a.
- 19 Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio,
20 R., Brunner, D., Chemel, C., Curci, G., Denier van der Gon, H.A.C., Flemming, J., Forkel, R.,
21 Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C.,
22 Makar, P.A., Manders-Groot, A., Neal, L., Perez, J.L., Piravano, G., Pouliot, G., San Jose, R.,
23 Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R.,
24 Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S.: Evaluation of
25 operational online-coupled regional air quality models over Europe and North America in the
26 context of AQMEII phase 2. Part II: Particulate Matter. *Atmospheric Environment* 115: 421-
27 441, 2015b.
- 28 Kalnay E.: *Atmospheric modelling, data assimilation and predictability*, Cambridge
29 University Press, 341 pp., 2003.
- 30 Kioutsioukis, I. and Galmarini S.: De praeceptis ferendis: good practice in multi-model
31 ensembles, *Atmospheric Chemistry and Physics* 14: 11791-11815, 2014.

- 1 Krogh A. and J. Vedelsby: Neural network ensembles, cross validation, and active learning,
2 In *Advances in Neural Information Processing Systems*, pp.231-238, 1995.
- 3 Mallet, V., Stoltz, G., Mauricette, B.: Ozone ensemble forecast with machine learning
4 algorithms. *J. Geophys. Res.* 114, D05307. doi:10.1029/2008JD009978, 2009.
- 5 Marécal V., V.-H. Peuch, C. Andersson, S. Andersson, J. Arteta, M. Beekmann, A.
6 Benedictow, R. Bergström, B. Bessagnet, A. Cansado, F. Chéroux, A. Colette, A. Coman, R.
7 L. Curier, H. A. C. Denier van der Gon, A. Drouin, H. Elbern, E. Emili, R. J. Engelen, H. J.
8 Eskes, G. Foret, E. Friese, M. Gauss, C. Giannaros, J. Guth, M. Joly, E. Jaumouillé, B. Josse,
9 N. Kadyrov, J. W. Kaiser, K. Krajsek, J. Kuenen, U. Kumar, N. Liora, E. Lopez, L.
10 Malherbe, I. Martinez, D. Melas, F. Meleux, L. Menut1, P. Moinat, T. Morales, J. Parmentier,
11 A. Piacentini, M. Plu, A. Poupkou, S. Queguiner, L. Robertson, L. Rouïl, M. Schaap, A.
12 Segers, M. Sofiev, L. Tarasson, M. Thomas, R. Timmermans, Á. Valdebenito, P. van
13 Velthoven, R. van Versendaal, J. Vira, A. Ung: A regional air quality forecasting system over
14 Europe: the MACC-II daily ensemble production *Geosci. Model Dev.*, 8, 2777-2813, 2015.
- 15 Massey, F.J.: The Kolmogorov-Smirnov Test for Goodness of Fit, *Journal of the American*
16 *Statistical Association*, 46(253): 68-78, 1951.
- 17 Monteiro, A., I. Ribeiro, O. Tchepel, A. Carvalho, H. Martins, E. Sá, J. Ferreira, V. Martins,
18 S. Galmarini, A. I. Miranda, C. Borrego: Ensemble Techniques to Improve Air Quality
19 Assessment: Focus on O₃ and PM, *Environmental Modeling & Assessment*, 18(3): 249–257,
20 2013.
- 21 Pagowski, M., G.A. Grell, S.A. McKeen, D. Devenyi, J.M. Wilczak, V. Bouchet, W. Gong, J.
22 McHenry, S. Peckham, J. McQueen, R. Moffet, Y. Tang: A simple method to improve
23 ensemble-based ozone forecasts, *Geophys. Res. Lett.*, 32, L07814,
24 doi:10.1029/2004GL022305, 2005.
- 25 Pagowski, M., G.A. Grell, D. Devenyi, S. Peckham, S.A. McKeen, W. Gong, L. Delle
26 Monache, J.N. McHenry, J. McQueen, P. Lee: Application of Dynamic Linear Regression to
27 Improve the Skill of Ensemble-Based Deterministic Ozone Forecasts, *Atmos. Environ.* 40:
28 3240-3250, 2006.
- 29 Potempski, S. and Galmarini, S.: Est modus in rebus: analytical properties of multi-model
30 ensembles, *Atmos. Chem. Phys.* 9: 9471-9489, 2009.

- 1 Pouliot, G., Pierce, T., Denier van der Gon, H., Schaap, M., Nopmongcol, U.: Comparing
2 Emissions Inventories and Model-Ready Emissions Datasets between Europe and North
3 America for the AQMEII Project. *Atmospheric Environment* 53: 4–14, 2012.
- 4 Pouliot, G., Hugo Denier van der Gon, J Kuenen, Junhua Zhang, Michael Moran, Paul
5 Makar: Analysis of the Emission Inventories and Model-Ready Emission Datasets of Europe
6 and North America for Phase 2 of the AQMEII Project, *Atmospheric Environment* 115: 345–
7 360, 2015.
- 8 Rao, S.T., Galmarini, S., Puckett, K.: Air quality model evaluation international initiative
9 (AQMEII): Advancing the state of the science in regional photochemical modeling and its
10 applications. *Bulletin of the American Meteorological Society* 92(1): 23-30, 2011.
- 11 Riccio, A., Giunta, G., Galmarini, S.: Seeking for the rational of the median model: the
12 optimal combination of multi- model ensemble. *Atmospheric Chemistry and Physics* 7: 6085–
13 6098, 2007.
- 14 San José, R., J.L. Pérez, A. Balzarini, R. Baró, G. Curci, R. Forkel, S. Galmarini, G. Grell, M.
15 Hirtl, L. Honzak, U. Im, P. Jiménez-Guerrero, M. Langer, G. Pirovano, P. Tuccella, J.
16 Werhahn, R. Žabkar: Sensitivity of feedback effects in CBMZ/MOSAIC chemical
17 mechanism, *Atmos. Environ.* 115: 646–656, 2015.
- 18 Schere, K., Flemming, J., Vautard, R., Chemel, C., Colette, A., Hogrefe, C., Bessagnet, B.,
19 Meleux, F., Mathur, R., Roselle, S., Hu, R.-M., Sokhi, R. S., Rao, S. T., and Galmarini, S.:
20 Trace gas/aerosol concentrations and their impacts on continental-scale AQMEII modelling
21 sub-regions, *Atmos. Environ.*, 53, 38–50, 2012.
- 22 Simmons, A.: From Observations to service delivery: Challenges and opportunities. *WMO*
23 *Bulletin* 60(2): 96-107, 2011.
- 24 Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet,
25 B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., van der Gon, H. D., Ferreira, J., Forkel,
26 R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jerice- vic, A., Kraljevic, L., Miranda,
27 A. I., Nopmongcol, U., Pirovano, G., Prank, M., Riccio, A., Sartelet, K. N., Schaap, M.,
28 Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S.
29 T., and Galmarini, S.: Model evaluation and ensemble modelling and for surface-level ozone
30 in Europe and North America, *Atmos. Environ.*, 53, 60–74, 2012a.

- 1 Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Appel, K.
2 W., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R.,
3 Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Hogrefe, C., Miranda, A. I., Nopmongco,
4 U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J.,
5 Wolke, R., Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Operational model
6 evaluation for particulate matter in Europe and North America, *Atmos. Environ.*, 53, 75–92,
7 2012b.
- 8 Solazzo E., A. Riccio, I. Kioutsioukis, S. Galmarini: Pauci ex tanto numero: reduce
9 redundancy in multi-model ensembles, *Atmos. Chem. Phys.* 13: 8315–8333, 2013.
- 10 Solazzo, E., Galmarini, S.: Error Apportionment for atmospheric chemistry transport models
11 – a new approach to model evaluation. *Atmospheric Chemistry and Physics* 16, 6263-6283,
12 2016.
- 13 Taylor, K.E.: Summarizing multiple aspects of model performance in a simple diagram.
14 *Journal Geophys. Res.* 106, D7, 7183-7192, 2001.
- 15 Ueda N., R. Nakano.: Generalization error of ensemble estimators, In *Proceedings of*
16 *International Conference on Neural Networks*, pages 90–95, 1996.
- 17 Vautard, R., M.D. Moran, E. Solazzo, R.C. Gilliam, V. Matthias, R. Bianconi, C. Chemel, J.
18 Ferreira, B. Geyer, A.B. Hansen, A. Jericevic, M. Prank, A. Segers, J.D. Silver, J. Werhahn,
19 R. Wolke, S.T. Rao, S. Galmarini: Evaluation of the meteorological forcing used for the Air
20 Quality Model Evaluation International Initiative (AQMEII) air quality simulations,
21 *Atmospheric Environment* 53:15-37, 2012.
- 22 Weigel A., R. Knutti, M. Liniger, C. Appenzeller: Risks of model weighting in multimodel
23 climate projections, *Journal of Climate* 23: 4175-4191, 2010.
- 24 Zhang, Y., C. Seigneur, M. Bocquet, V. Mallet, A. Baklanov: Real-Time Air Quality
25 Forecasting, Part II: State of the Science, Current Research Needs, and Future Prospects,
26 *Atmos. Environ.*, 60, 656-676, 2012.
- 27 Zurbenko, I. G.: *The Spectral Analysis of Time Series*, 236 pp., North-Holland, Amsterdam,
28 1986.
- 29
30

1 **Acknowledgements**

2 We gratefully acknowledge the contribution of various groups to the second air Quality
3 Model Evaluation international Initiative (AQMEII) activity: U.S. EPA, Environment
4 Canada, Mexican Secretariat of the Environment and Natural Resources (Secretaría de Medio
5 Ambiente y Recursos Naturales-SEMARNAT) and National Institute of Ecology (Instituto
6 Nacional de Ecología-INE) (North American national emissions inventories); U.S. EPA
7 (North American emissions processing); TNO (European emissions processing);
8 ECMWF/MACC project & Météo-France/CNRM-GAME (Chemical boundary conditions).
9 Ambient North American concentration measurements were extracted from Environment
10 Canada's National Atmospheric Chemistry Database (NAtChem) PM database and provided
11 by several U.S. and Canadian agencies (AQS, CAPMoN, CASTNet, IMPROVE, NAPS,
12 SEARCH and STN networks); North American precipitation-chemistry measurements were
13 extracted from NAtChem's precipitation-chemistry data base and were provided by several
14 U.S. and Canadian agencies (CAPMoN, NADP, NBPMN, NSPSN, and REPQ networks); the
15 WMO World Ozone and Ultraviolet Data Centre (WOUDC) and its data-contributing
16 agencies provided North American and European ozonesonde profiles; NASA's AErosol
17 RObotic NETwork (AeroNet) and its data-contributing agencies provided North American
18 and European AOD measurements; the MOZAIC Data Centre and its contributing airlines
19 provided North American and European aircraft takeoff and landing vertical profiles; for
20 European air quality data the following data centers were used: EMEP European Environment
21 Agency/European Topic Center on Air and Climate Change/AirBase provided European air-
22 and precipitation-chemistry data. The Finish Meteorological Institute is acknowledged for
23 providing biomass burning emission data for Europe. Data from meteorological station
24 monitoring networks were provided by NOAA and Environment Canada (for the US and
25 Canadian meteorological network data) and the National Center for Atmospheric Research
26 (NCAR) data support section. Joint Research Center Ispra/Institute for Environment and
27 Sustainability provided its ENSEMBLE system for model output harmonization and analyses
28 and evaluation. The co-ordination and support of the European contribution through COST
29 Action ES1004 EuMetChem is gratefully acknowledged. The views expressed here are those
30 of the authors and do not necessarily reflect the views and policies of the U.S. Environmental
31 Protection Agency (EPA) or any other organization participating in the AQMEII project. This
32 paper has been subjected to EPA review and approved for publication. The UPM authors
33 thankfully acknowledge the computer resources, technical expertise and assistance provided

1 by the Centro de Supercomputación y Visualización de Madrid (CESVIMA) and the Spanish
2 Supercomputing Network (BSC). GC and PT were supported by the Italian Space Agency
3 (ASI) in the frame of PRIMES project (contract n. I/017/11/0). The same authors are deeply
4 thankful to the Euro Mediterranean Centre on Climate Change (CMCC) for having made
5 available the computational resources.

6

1 **Appendix I**

- 2 The relevant separate scales of motion are defined by means of physical considerations and
3 periodogram analysis (Rao et al., 1997). They are namely the intra-day component (ID), the
4 diurnal component (DU), the synoptic component (SY) and the long-term component (LT).
5 The hourly time series (S) can therefore be decomposed as:

$$S(t) = ID(t) + DU(t) + SY(t) + LT(t) \quad (1)$$

- 6 where:

$$ID(t) = S(t) - KZ_{3,3}$$

$$DU(t) = KZ_{3,3} - KZ_{13,5}$$

$$SY(t) = KZ_{13,5} - KZ_{103,5}$$

$$LT(t) = KZ_{103,5}$$

(2)

7

1 Table 1. The modelling systems participating in the first and second phases of AQMEII for Europe.

	Model		Grid	Emissions	Chemical BC
	Met	AQ			
EU – AQMEII phase I	MM5	DEHM	50 km	Global emission databases, EMEP	Satellite measurements
	MM5	Polyphemus	24 km	Standard [§]	Standard
	MM5	Chimere	25 km	MEGAN, Standard	Standard
	MM5	CAMx	15 km	MEGAN, Standard	Standard
	PARLAM-PS	EMEP	50 km	EMEP model	From ECMWF and forecasts
	WRF	CMAQ	18 km	Standard [§]	Standard
	WRF	Chem	22.5 km	Standard [§]	Fixed
	ECMWF	SILAM	24 km	Standard anthropogenic; In-house biogenic	Standard
	ECMWF	Lotos-EUROS	25 km	Standard [§]	Standard
	GEM	GEM-AQ	25 km	Standard (AQMEII region); EDGAR/GEIA (rest of the global domain)	Global variable grid setup (no boundary conditions)
	COSMO	Muscat	24 km	Standard [§]	Standard
COSMO-CLM	CMAQ	24 km	Standard [§]	Standard	
EU – AQMEII phase II	WRF	Chem	23 km	Standard	Standard
	WRF	CMAQ	18 km	Standard	Standard
	COSMO	Cosmo-ART	0.22°	Standard	Standard
	COSMO	Muscat	0.25°	Standard	Standard
	NMMB	BSCCTM	0.20°	Standard	Standard
	RACMO	LOTOS-EUROS	0.5° x 0.25°	Standard	Standard
	MetUM	UKCA RAQ	0.22°	Standard	Standard

2 AQMEII phase I

3 Standard Boundary conditions: provided from GEMS project (Global and regional Earth-system Monitoring using Satellite and in-situ data). Refer to Schere et al. (2012) for details.

4
5 [§] Standard anthropogenic emissions and biogenic emissions derived from meteorology (temperature and solar radiation) and land use distribution implemented in the meteorological driver. Refer to Solazzo et al. (2012a-b) and references therein for details.

7 AQMEII phase II

8
9 Standard Boundary conditions: 3-D daily chemical boundary conditions were provided by the ECMWF IFS-MOZART model run in the context of the MACC-II project (Monitoring Atmospheric Composition and Climate - Interim Implementation) at 3-hourly and 1.125 spatial resolution. Refer to Im et al. (2015a-b) for details.

1
2
3
4

Standard Emissions: based on the TNO-MACC-II (Netherlands Organization for Applied Scientific Research, Monitoring Atmospheric Composition and Climate - Interim Implementation) framework for Europe. Refer to Im et al. (2015a-b) for details.

1 **Table 2. The statistical distribution of (a) the Normalized Mean Square Error (NMSE) of the best**
 2 **model ($NMSE_{BEST}$), (b) the ensemble average NMSE ($\langle NMSE \rangle$) and (c) the skill difference indicator**
 3 **($NMSE_{BEST} / \langle NMSE \rangle$). In addition, the coefficient of variation (CoV = standard deviation / mean) of**
 4 **the number of cases where each model has been identified as best. All indicators have been**
 5 **evaluated at each monitoring site for the examined species of the two AQMEII phases.**

	O ₃	O ₃	NO ₂	NO ₂	PM ₁₀	PM ₁₀
	(I/II)	(I/II)	(I/II)	(I/II)	(I/II)	(I/II)
	$\langle NMSE \rangle$	$NMSE_{BEST}$	$\langle NMSE \rangle$	$NMSE_{BEST}$	$\langle NMSE \rangle$	$NMSE_{BEST}$
5 th	0.04 / 0.04	0.03 / 0.03	0.28 / 0.23	0.17 / 0.18	0.30 / 0.27	0.20 / 0.20
25 th	0.07 / 0.07	0.05 / 0.05	0.39 / 0.35	0.24 / 0.25	0.40 / 0.39	0.26 / 0.28
50 th	0.10 / 0.10	0.07 / 0.08	0.52 / 0.49	0.33 / 0.34	0.47 / 0.51	0.34 / 0.37
75 th	0.15 / 0.15	0.11 / 0.12	0.82 / 0.76	0.48 / 0.50	0.61 / 0.62	0.46 / 0.50
95 th	0.24 / 0.23	0.18 / 0.18	1.69 / 1.49	0.81 / 0.93	1.02 / 0.98	0.73 / 0.81
$\frac{NMSE_{BEST}}{\langle NMSE \rangle}$	O ₃	O ₃	NO ₂	NO ₂	PM ₁₀	PM ₁₀
	(I)	(II)	(I)	(II)	(I)	(II)
5 th	0.50	0.60	0.36	0.45	0.49	0.63
25 th	0.62	0.70	0.50	0.62	0.61	0.72
50 th	0.70	0.76	0.61	0.72	0.70	0.79
75 th	0.76	0.82	0.72	0.81	0.85	0.85
95 th	0.83	0.88	0.87	0.93	0.92	0.92
mean	0.69	0.75	0.61	0.70	0.72	0.77
N_{BEST}	O ₃	O ₃	NO ₂	NO ₂	PM ₁₀	PM ₁₀
	(I)	(II)	(I)	(II)	(I)	(II)
CoV	1.08	0.70	1.42	0.65	1.16	1.53

1 **Table 3. The MSE from (a) the best deterministic models, globally (*bestG*) and locally (*bestL*), (b)**
2 **the unconditional ensemble mean (*mme*) and (c) the four conditional ensemble estimators (*mme*<**
3 ***kzFO*, *kzHO*, *mmW*). In addition, the bounds for the MSE of the ensemble mean are also presented.**
4 **The maximum value (<MSE>) arises for ensemble members without diversity and the minimum**
5 **value (*mme*MIN) has been estimated from the variance term only (i.e. calculated for unbiased and**
6 **uncorrelated ensemble members). The ability of the estimators is evaluated through their skill**
7 **scores ($SS_{REF}=1-MSE/MSE_{REF}$, $REF=bestG$, <MSE>, *mme*).**

		SS	SS	SS			SS	SS	SS
O3 (I)	MSE	(bestG)	(<MSE>)	(mme)	O3 (II)	MSE	(bestG)	(<MSE>)	(mme)
bestG	641		7%		bestG	499		14%	
bestL	483	25%	30%	3%	bestL	441	12%	24%	3%
mme	498	22%	28%		mme	454	9%	21%	
mme<	398	38%	42%	20%	mme<	374	25%	35%	18%
kzFO	400	38%	42%	20%	kzFO	369	26%	36%	19%
kzHO	367	43%	47%	26%	kzHO	349	30%	40%	23%
mmW	345	46%	50%	31%	mmW	315	37%	45%	31%
<MSE>	690				<MSE>	577			
mmeMIN	58				mmeMIN	41			
		SS	SS	SS			SS	SS	SS
NO2 (I)	MSE	(bestG)	(<MSE>)	(mme)	NO2 (II)	MSE	(bestG)	(<MSE>)	(mme)
bestG	77		25%		bestG	61		20%	
bestL	70	10%	32%	3%	bestL	58	5%	25%	-4%
mme	72	7%	30%		mme	56	9%	27%	
mme<	63	19%	39%	13%	mme<	51	17%	34%	9%
kzFO	62	19%	40%	13%	kzFO	52	16%	33%	8%
kzHO	59	24%	43%	18%	kzHO	48	21%	37%	14%
mmW	56	27%	46%	22%	mmW	46	25%	40%	18%
<MSE>	104				<MSE>	77			
mmeMIN	8				mmeMIN	6			
		SS	SS	SS			SS	SS	SS
PM10 (I)	MSE	(bestG)	(<MSE>)	(mme)	PM10 (II)	MSE	(bestG)	(<MSE>)	(mme)
bestG	341		16%		bestG	141		14%	
bestL	326	5%	20%	1%	bestL	139	2%	15%	0%
mme	330	3%	19%		mme	139	1%	15%	
mme<	303	11%	25%	8%	mme<	121	14%	26%	13%

kzFO	299	13%	27%	10%	kzFO	122	13%	25%	12%
kzHO	294	14%	28%	11%	kzHO	117	17%	29%	16%
mmW	284	17%	30%	14%	mmW	105	26%	36%	25%
<MSE>	407				<MSE>	164			
mmeMIN	41				mmeMIN	12			

- 1 **mme**: unconditional ensemble mean
- 2 **mme<**: conditional ensemble mean (Kioutsioukis and Galmarini, 2014)
- 3 **kzFO**: conditional spectral ensemble mean with 1st order components (Galmarini et al., 2013)
- 4 **kzHO**: conditional spectral ensemble mean with 2nd and higher order components (*kzHO*)
- 5 **mmW**: optimal weighted ensemble (Potemski and Galmarini, 2009)
- 6

1 Table 4. The probability of detection (POD) and false alarm rate (FAR) from (a) the best
 2 deterministic models, globally (*bestG*) and locally (*bestL*), (b) the unconditional ensemble mean
 3 (*mme*) and (c) the four conditional ensemble estimators (*mme<*, *kzFO*, *kzHO*, *mmW*). Two
 4 thresholds were examined for each indicator, corresponding to tail percentiles.

O3 (I)	POD	FAR	POD	FAR	O3 (II)	POD	FAR	POD	FAR
threshold	120		180		threshold	120		180	
bestG	37.9	3.6	11.4	0.0	bestG	19.9	1.2	1.2	0.0
bestL	54.7	3.5	19.5	0.0	bestL	33.2	1.5	5.4	0.0
mme	39.9	2.5	12.0	0.0	mme	22.0	1.2	0.5	0.0
mme<	53.5	2.6	18.3	0.0	mme<	34.9	1.3	2.4	0.0
kzFO	57.7	3.0	19.6	0.0	kzFO	39.1	1.5	4.4	0.0
kzHO	57.1	2.5	19.2	0.0	kzHO	36.9	1.2	2.3	0.0
mmW	60.6	2.6	27.2	0.0	mmW	45.4	1.6	8.6	0.0
NO2 (I)	POD	FAR	POD	FAR	NO2 (II)	POD	FAR	POD	FAR
threshold	25		50		threshold	25		50	
bestG	45.9	4.6	3.8	0.2	bestG	39.3	3.3	4.9	0.1
bestL	48.7	4.2	8.5	0.3	bestL	41.4	3.1	8.1	0.1
mme	49.4	4.6	3.0	0.1	mme	44.4	3.5	5.4	0.1
mme<	52.2	4.1	7.1	0.1	mme<	47.6	3.2	7.6	0.1
kzFO	52.7	4.1	8.4	0.1	kzFO	46.5	3.1	9.5	0.1
kzHO	54.2	4.0	6.8	0.1	kzHO	49.5	3.2	9.3	0.1
mmW	57.0	4.1	14.8	0.2	mmW	50.9	3.1	13.5	0.1
PM10 (I)	POD	FAR	POD	FAR	PM10 (II)	POD	FAR	POD	FAR
threshold	50		90		threshold	50		90	
bestG	25.9	2.7	1.2	0.0	bestG	13.0	0.4	0.0	0.0
bestL	27.8	2.3	6.9	1.2	bestL	14.5	0.4	1.6	0.0
mme	21.6	1.8	0.4	0.0	mme	11.4	0.4	0.0	0.0
mme<	30.6	2.3	5.6	0.1	mme<	13.9	0.4	0.0	0.0

kzFO	31.1	2.3	6.9	0.1	kzFO	14.1	0.3	0.0	0.0
kzHO	33.2	2.4	6.1	0.1	kzHO	13.2	0.3	0.2	0.0
mmW	35.5	2.6	13.3	0.2	mmW	23.9	0.4	20.8	0.0

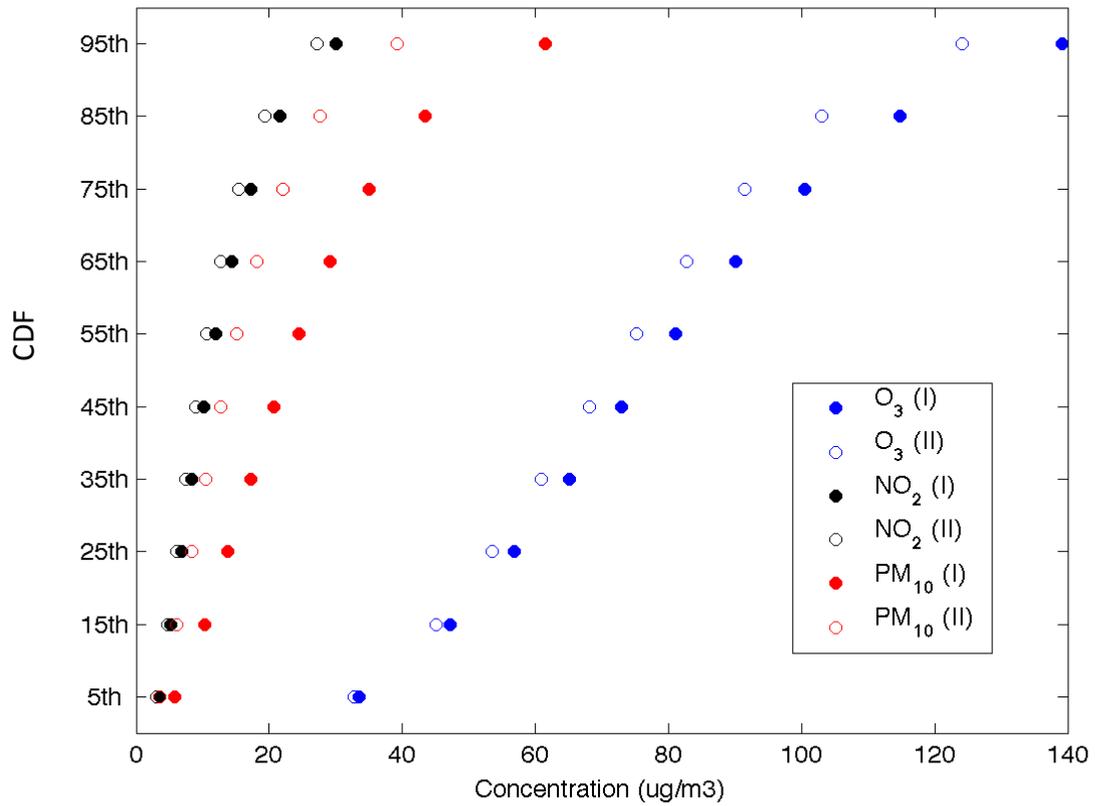
- 1 *mme*: unconditional ensemble mean
- 2 *mme*<: conditional ensemble mean (Kioutsioukis and Galmarini, 2014)
- 3 *kzFO*: conditional spectral ensemble mean with 1st order components (Galmarini et al., 2013)
- 4 *kzHO*: conditional spectral ensemble mean with 2nd and higher order components (*kzHO*)
- 5 *mmW*: optimal weighted ensemble (Potemski and Galmarini, 2009)
- 6
- 7

1 Table 5. The average MSE of *mmW* for various training lengths, calculated for the testing time-
 2 series (i.e. not-used in the training phase) that contains all stations.

Length of training period (days)	O ₃	O ₃	NO ₂	NO ₂	PM ₁₀	PM ₁₀
	(I)	(II)	(I)	(II)	(I)	(II)
5	616	540	90	91	717	210
10	496	441	77	66	443	150
20	400	378	65	56	348	125
30	380	344	62	52	308	109
40	366	334	59	50	300	113
50	357	326	57	48	294	108
60	351	319	56	45	282	102

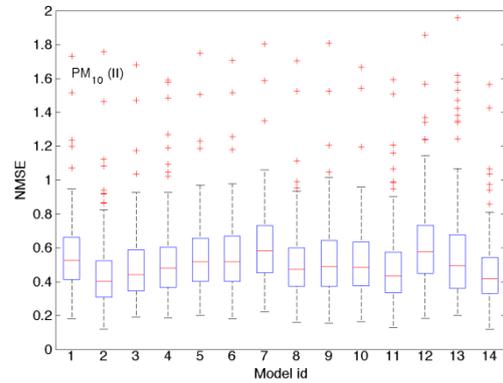
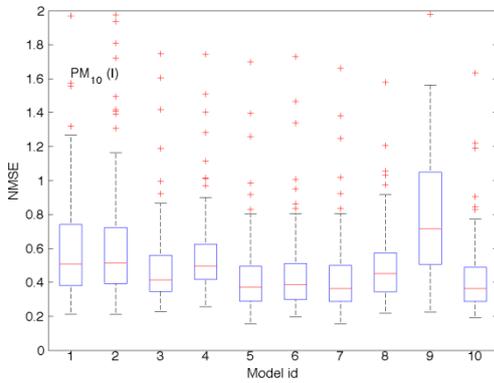
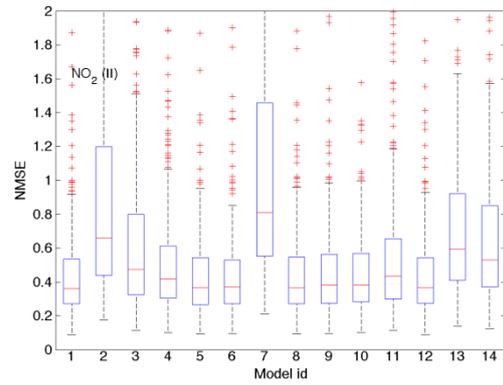
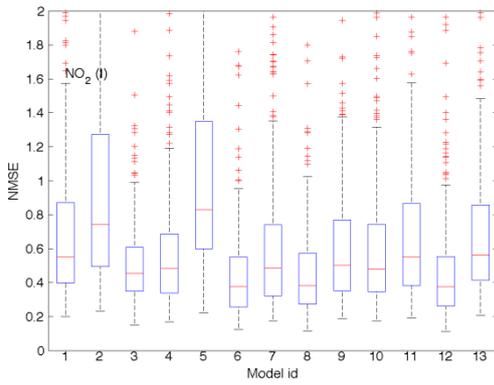
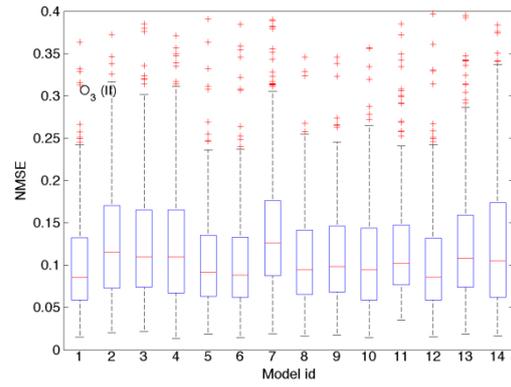
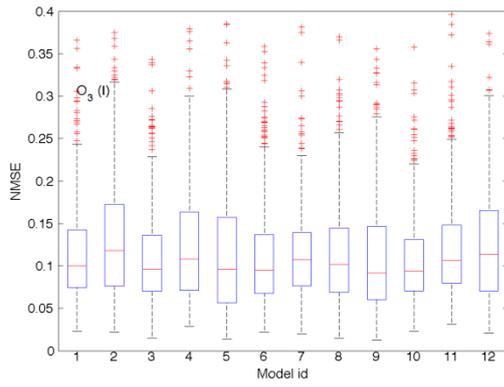
3

4



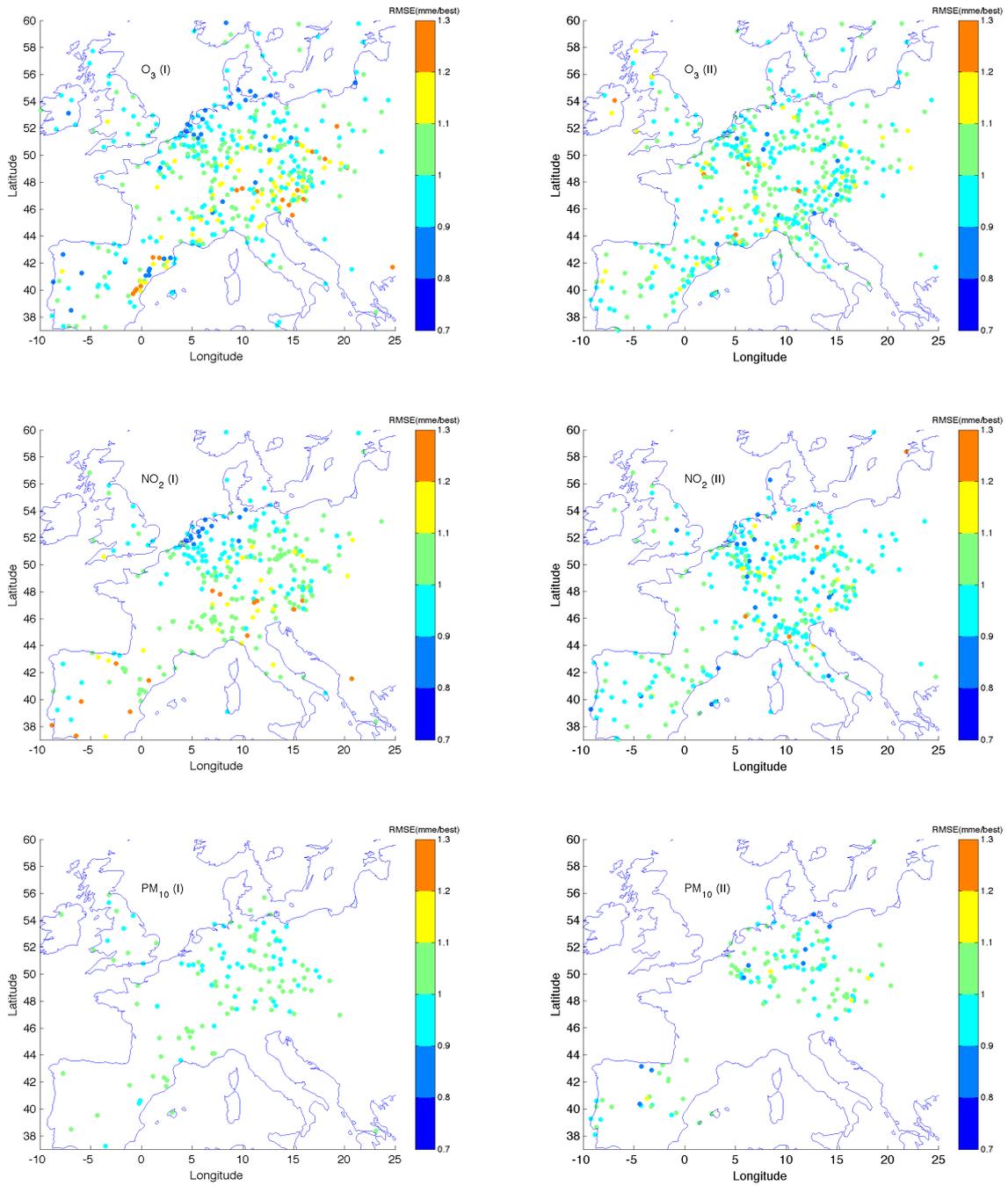
1 Figure 1. The Cumulative density functions of the observations (O₃, NO₂, PM₁₀) in the two AQMEII
 2 phases (Phase I: *filled circles*, Phase II: *non-filled circles*). Each bullet represents the median at the
 3 specific percentile.

4



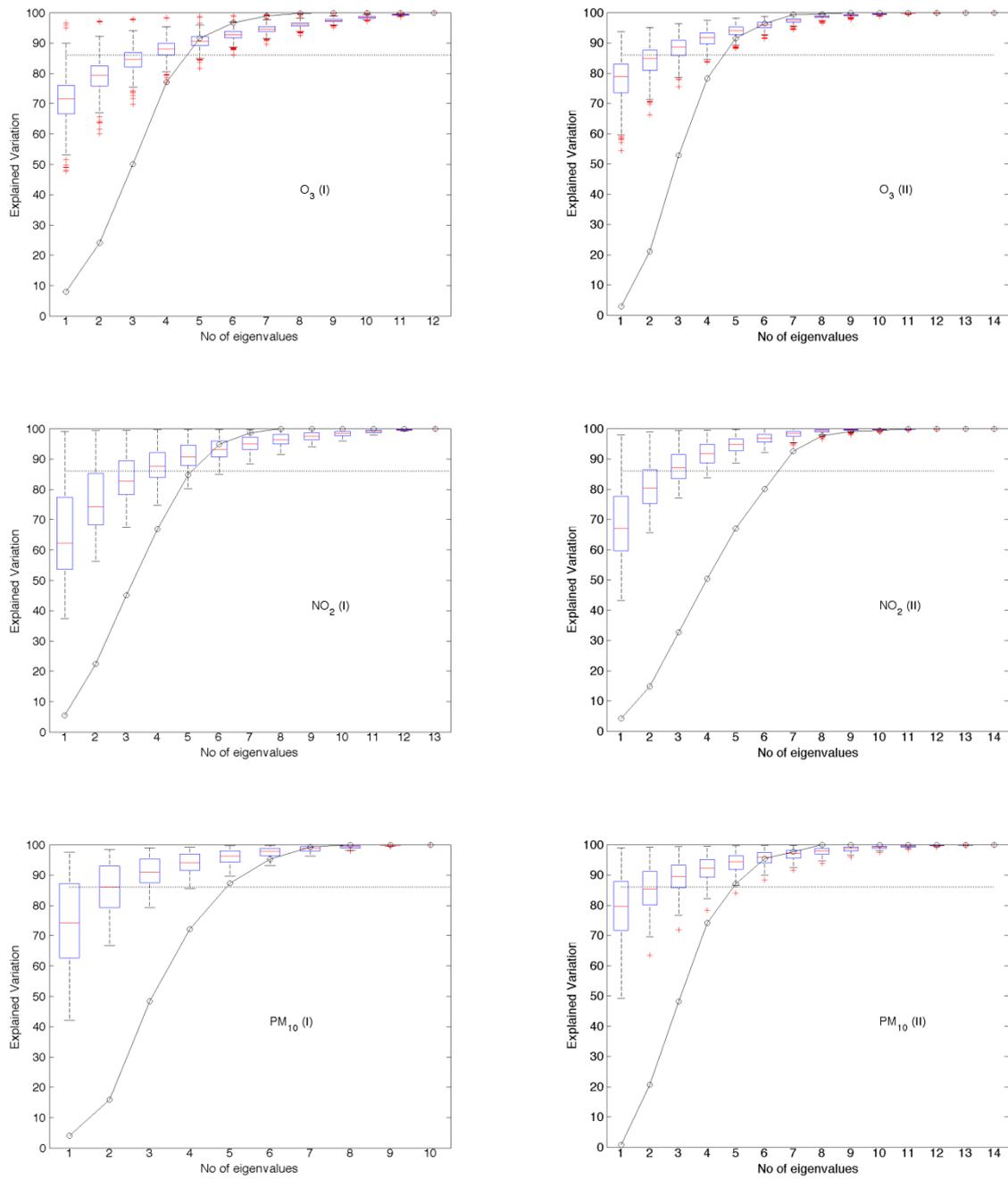
1 **Figure 2. Model skill difference via the NMSE. On each box, the central mark indicates the median,**
 2 **and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The**
 3 **whiskers extend to the most extreme data points not considered outliers and the outliers (points**
 4 **with distance from the 25th and 75th percentiles larger than 1.5 times the interquartile range) are**
 5 **plotted individually using the '+' symbol.**

6

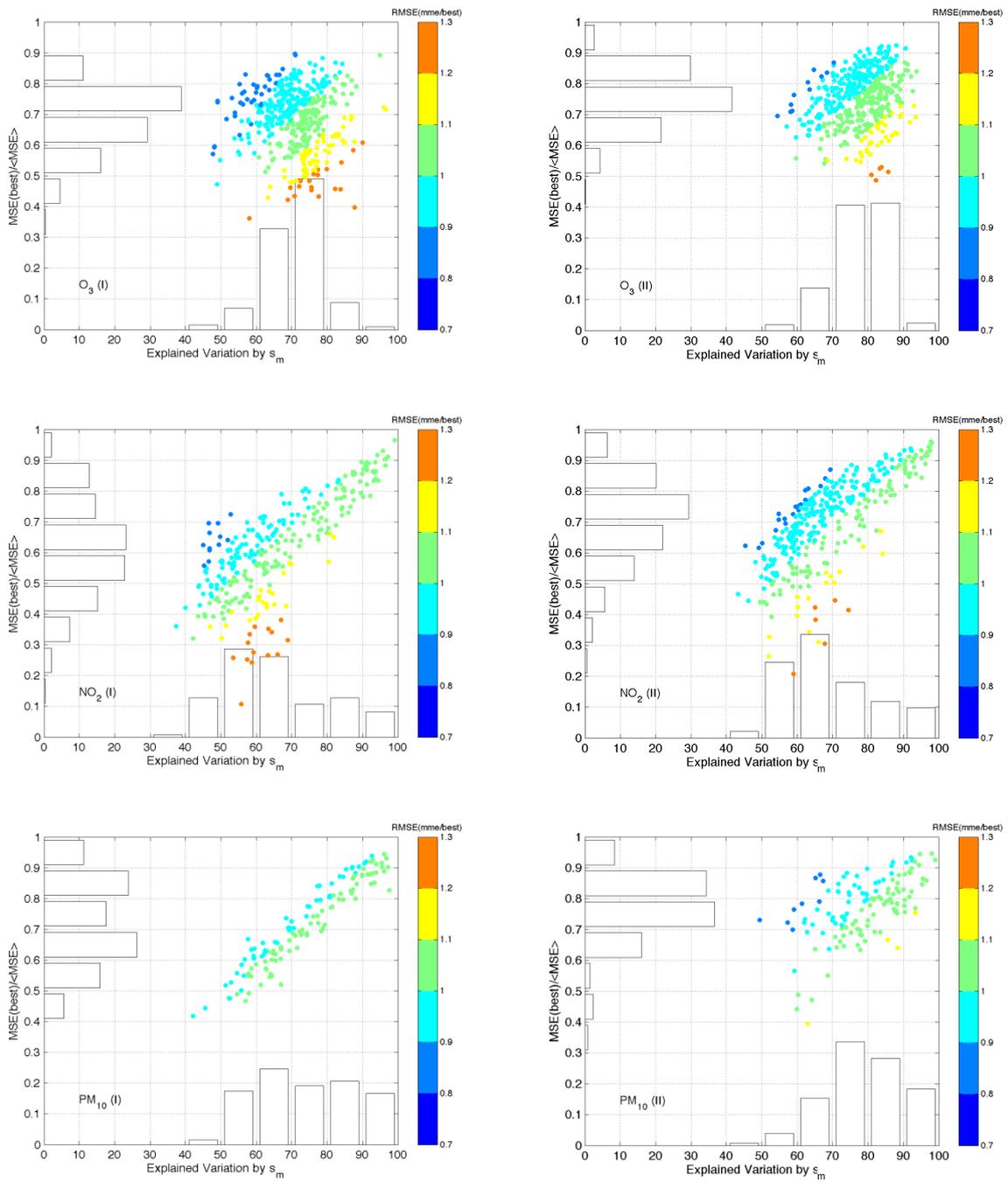


1 **Figure 3. Comparison of the *mme* skill against the best local deterministic model by means of the**
 2 **indicator $RMSE_{MME}/RMSE_{BEST}$.**

3

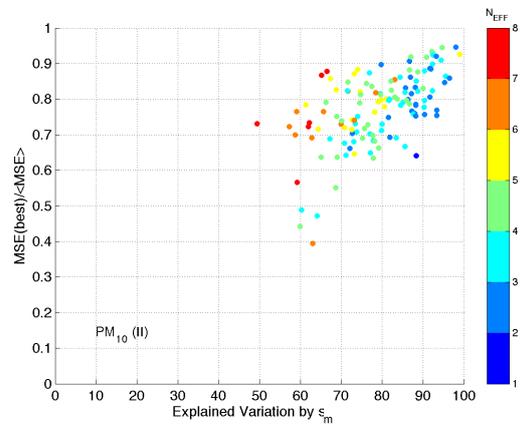
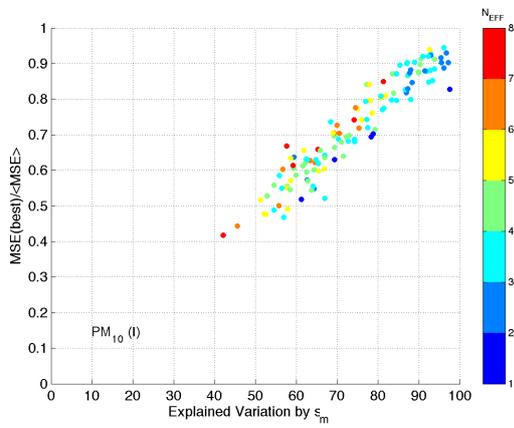
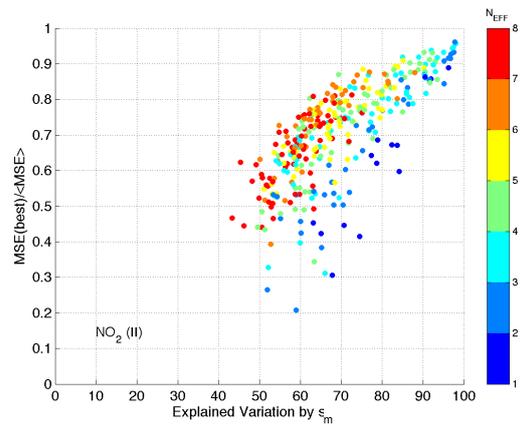
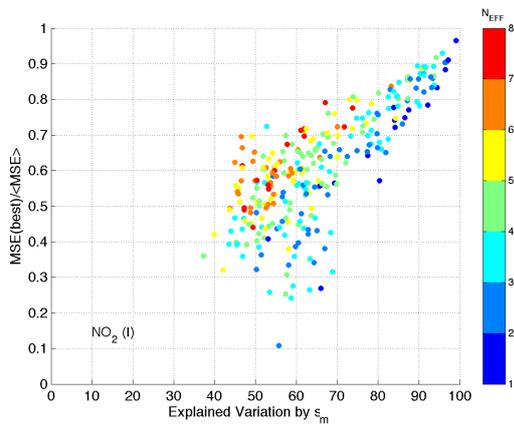
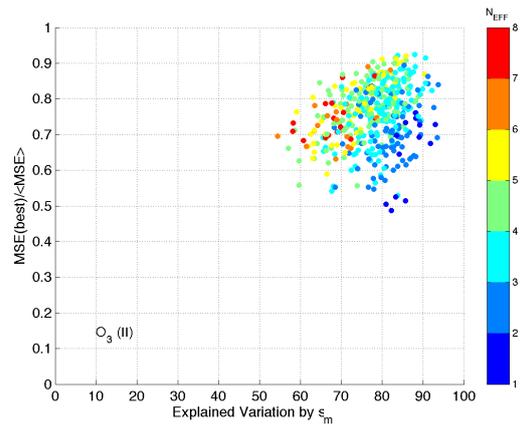
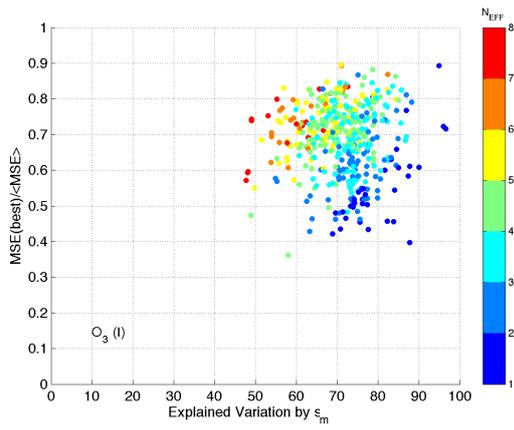


1 **Figure 4. Model error dependence through the eigenvalues spectrum. The average explained**
 2 **variation from the maximum eigenvalue is 71/78 (phase I/II) for O_3 , 65/69 for NO_2 and 74/79 for**
 3 **PM_{10} . On the same graph, the cumulative density function of N_{EFF} calculated from all possible**
 4 **ensemble combinations is presented with the black line.**



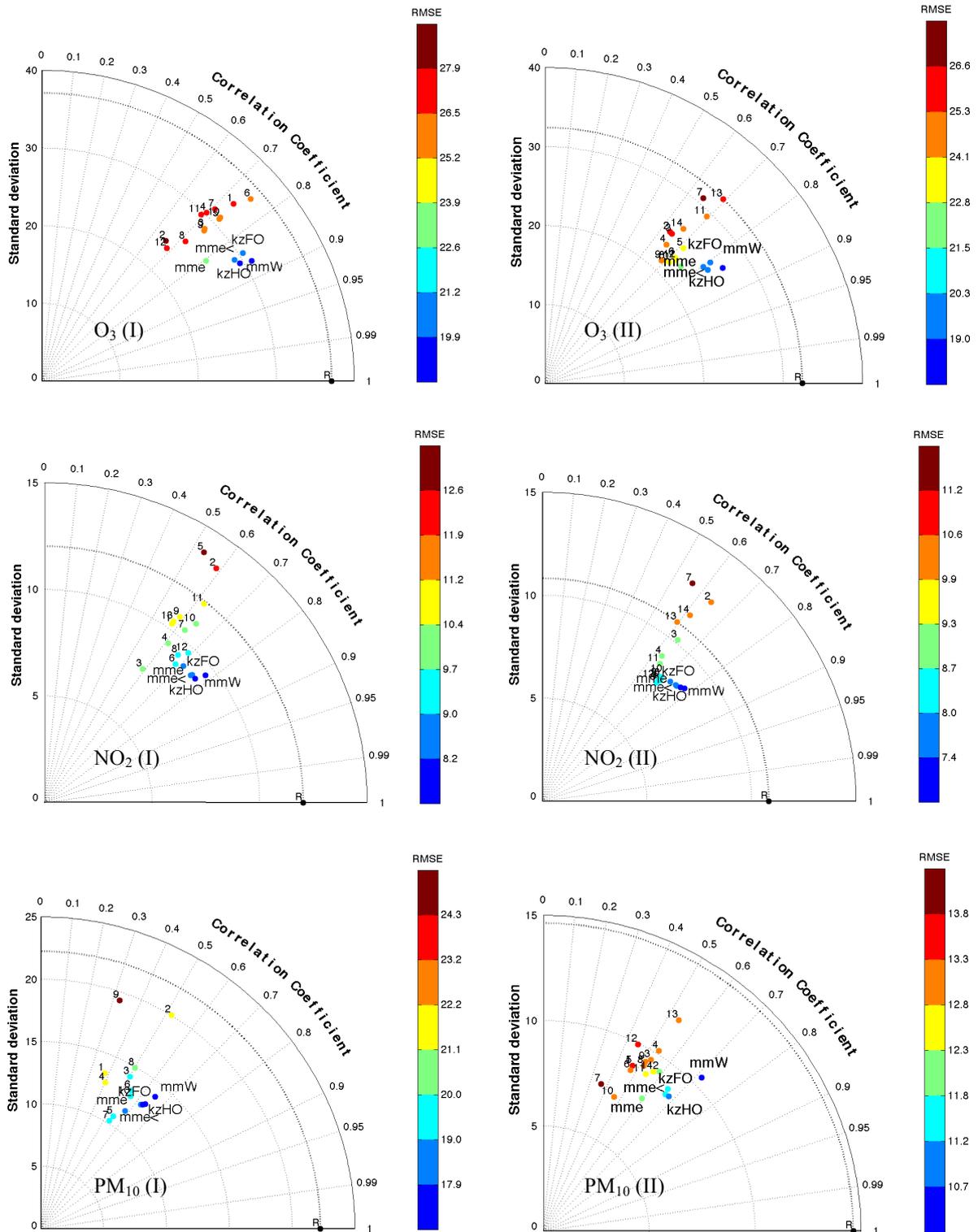
1 **Figure 5. Interpretation of Figure 4: the explanation of the mme skill against the best local**
 2 **deterministic model with respect to skill difference (evaluated from $MSE_{BEST}/\langle MSE \rangle$) and error**
 3 **dependence (evaluated from the explained variation by the highest eigenvalue).**

4



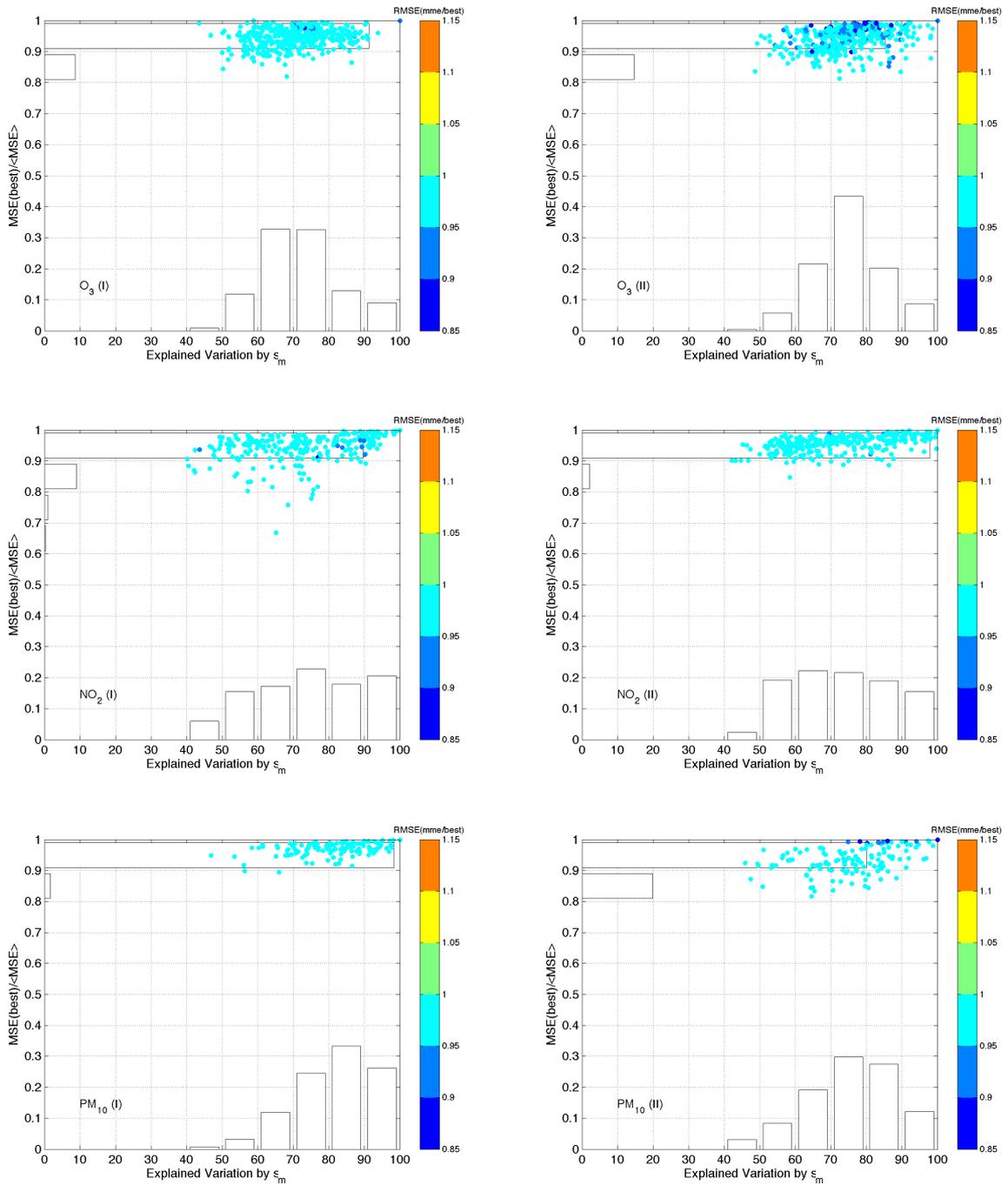
1 Figure 6. Like Figure 5 but showing the N_{EFF} with respect to skill difference and error dependence.

2



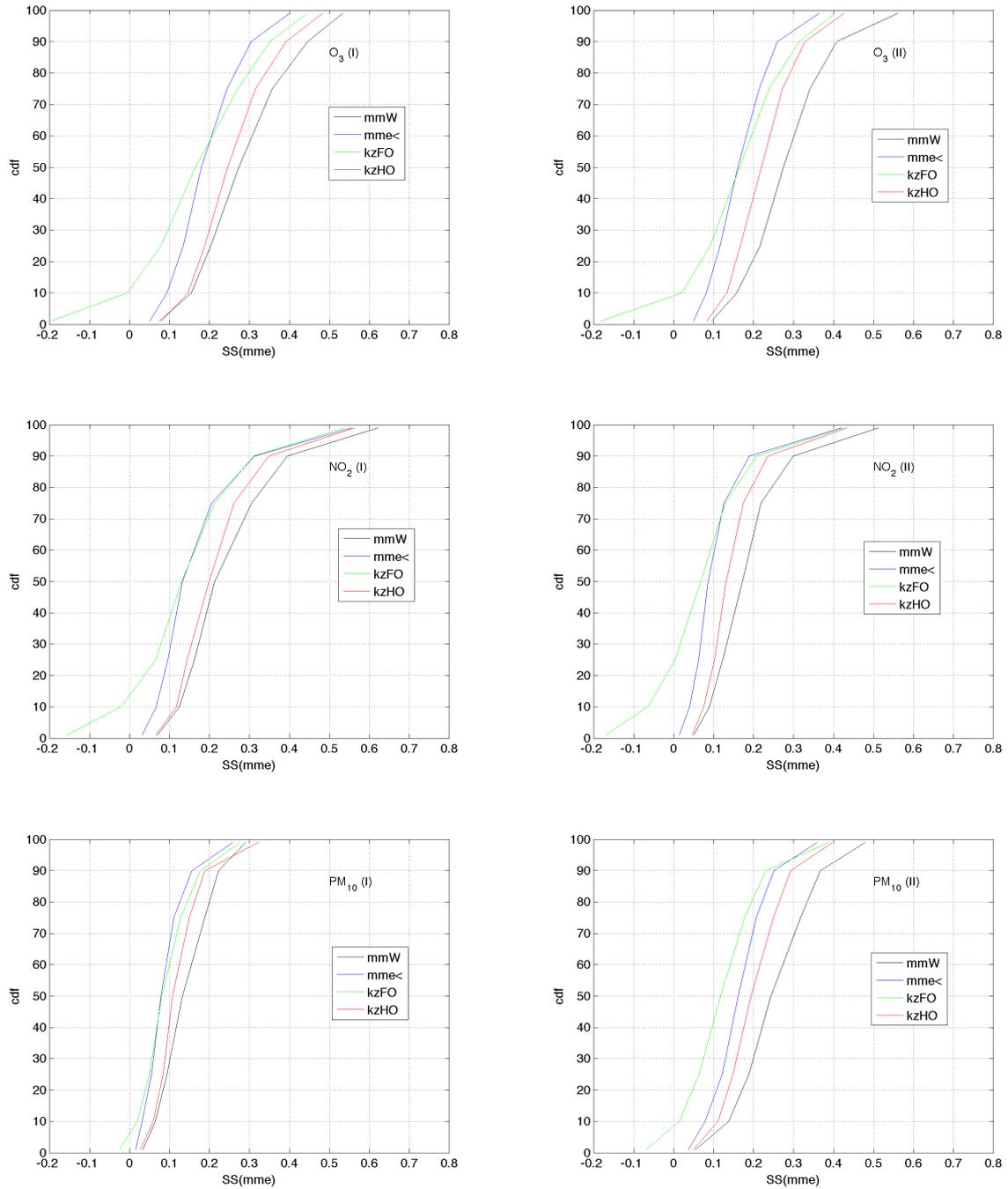
1 **Figure 7. Composite skill of all deterministic models and ensemble estimators (*mme*, *mme<*, *kzFO*,
 2 *kzHO*, *mmW*) through Taylor plots. The point R represents the reference point (i.e. observations).**

3



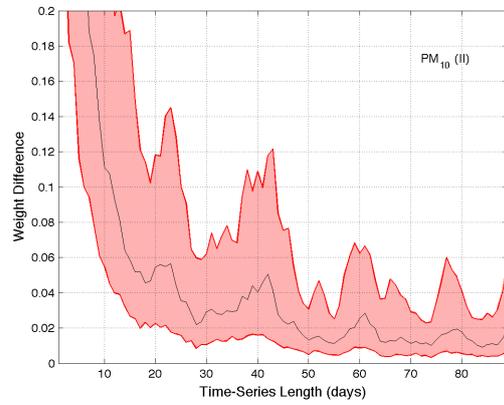
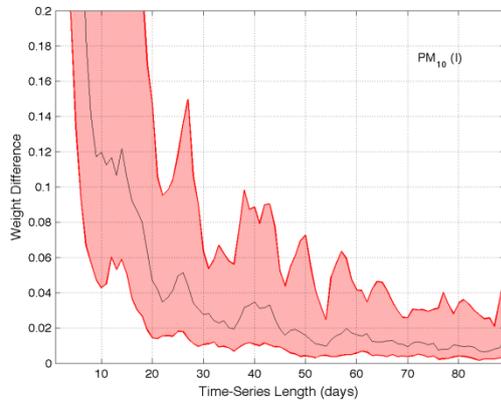
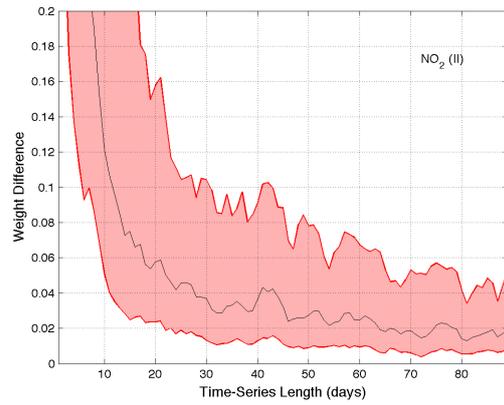
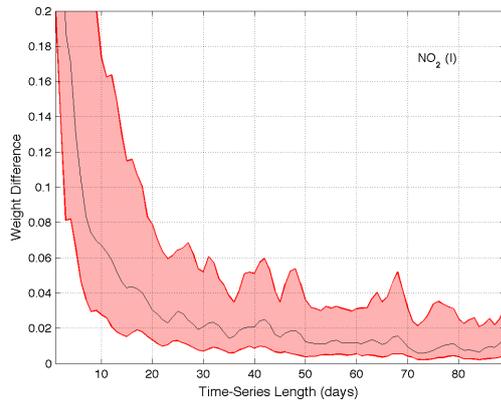
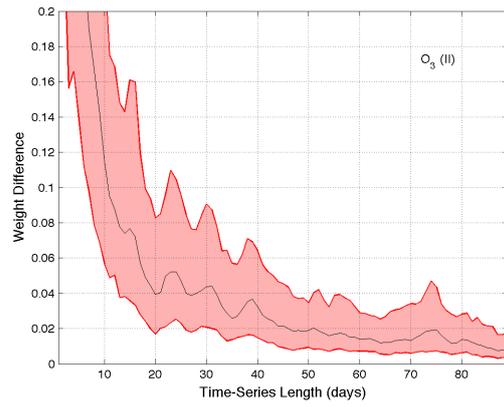
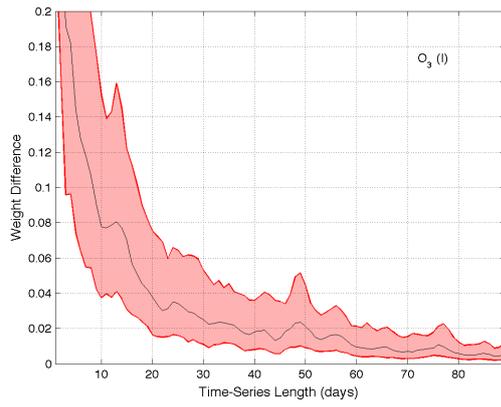
1 **Figure 8.** Like Figure 5 but for the *mme*< skill in the reduced ensemble. Please note the change in
 2 **the colorscale.**

- 3
- 4
- 5



1 Figure 9. The cumulative density function of the Skill Score ($1-MSE_X/MSE_{MME}$, $X = mmW, mme<, kzFO, kzHO$) over mme , evaluated at each monitoring site for the examined species of the two AQMEII phases.

4



1 **Figure 10. The interquartile range over all stations of the day-to-day difference in the weights**
 2 **arising from variable time-series length.**

3