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 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 Table2 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 \( N_{\text{Eff}} \) 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.
Insights into the deterministic skill of air quality ensembles from the analysis of AQMEII data

Ioannis Kioutsioukis\textsuperscript{a,b}, Ulas Im\textsuperscript{c}, Efisio Solazzo\textsuperscript{b}, Roberto Bianconi\textsuperscript{d}, Alba Badía\textsuperscript{e}, Alessandra Balzarini\textsuperscript{f}, Rocio Baró\textsuperscript{l}, Roberto Bellasio\textsuperscript{d}, Dominik Brunner\textsuperscript{g}, Charles Chemel\textsuperscript{h}, Gabriele Curci\textsuperscript{i}, Hugo Denier van der Gon\textsuperscript{k}, Johannes Flemming\textsuperscript{m}, Renate Forkel\textsuperscript{n}, Lea Giordano\textsuperscript{g}, Pedro Jiménez-Guerrero\textsuperscript{l}, Marcus Hirtl\textsuperscript{o}, Oriol Jorba\textsuperscript{e}, Astrid Manders-Groot\textsuperscript{k}, Lucy Neal\textsuperscript{p}, Juan L. Pérez\textsuperscript{q}, Guidio Pirovano\textsuperscript{f}, Roberto San Jose\textsuperscript{q}, Nicholas Savage\textsuperscript{p}, Wolfram Schroder\textsuperscript{r}, Ranjeet S Sokhi\textsuperscript{h}, Dimiter Syrakov\textsuperscript{s}, Paolo Tuccella\textsuperscript{ij}, Johannes Werhahn\textsuperscript{n}, Ralf Wolke\textsuperscript{e}, Christian Hogrefe\textsuperscript{t}, Stefano Galmarini\textsuperscript{b}

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

Simulations from chemical weather models are subject to uncertainties in the input data (e.g., emission inventory, initial and boundary conditions) as well as those intrinsic to the model (e.g. physical parameterization, chemical mechanism). Multi-model ensembles can improve the forecast skill provided that certain mathematical conditions are fulfilled. In this work, four ensemble methods were applied to two different datasets and their performance was compared for ozone (O$_3$), nitrogen dioxide (NO$_2$) and particulate matter (PM$_{10}$). Apart from the unconditional ensemble average, the approach behind the other three methods relies on adding optimum weights to members or constraining the ensemble to those members that meet certain conditions in time or frequency domain. The two different datasets were created for the first and second phase of the Air Quality Model Evaluation International Initiative (AQMEII). The methods are evaluated against ground level observations collected from the EMEP and Airbase databases. The goal of the study is to quantify to what extent we can extract predictable signals from an ensemble with superior skill over the single models and the ensemble mean. Verification statistics shows that the deterministic models simulate better O$_3$ than NO$_2$ and PM$_{10}$, linked to different levels of complexity in the represented processes. The unconditional ensemble mean achieves higher skill compared to each station’s best deterministic model at no more than 60% of the sites, indicating for the rest a combination of members with unbalanced skill difference and error dependence. The promotion of the right amount of accuracy and diversity within the ensemble results in an average additional skill up to 31% compared to using the full ensemble in an unconditional way. The skill improvements were higher for O$_3$ and lower for PM$_{10}$, associated to the extent of potential changes in the joint distribution of accuracy and diversity in the respective ensembles. The skill enhancement was superior using the weighting scheme but the training period required to acquire representative weights was longer compared to the sub-selecting schemes. Further development of the method is discussed in the conclusion.

Keywords: AQMEII, multi-model ensembles, air quality model, error decomposition, verification.
1 Introduction

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

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

Ideally, for continuous and unbiased variables, the multi-model ensemble mean outscores the skill of the deterministic models provided that the members have similar skill and independent errors (Potempski and Galmarini, 2009; Weigel et al., 2010). Practically, the multi-model ensemble mean usually outscores the skill of the deterministic models if the evaluation is performed over multiple observation sites and times. This occurs because over a network of stations, there are some where the essential conditions (e.g. the skill difference between the models is not too large) for the ensemble members are fulfilled, favouring the ensemble mean; for the remaining stations, where the conditions are not fulfilled, local verification identifies the best model but generally no single model is the best at all sites. Hence, although the skill of the numerical models varies in space (latitude, longitude, altitude) and time (e.g., hour of the day, month, season), the ensemble mean is usually the most accurate spatio-temporal representation.
One of the challenges in multi-model ensemble forecasting is the processing of the deterministic models datasets prior to averaging in order to construct another dataset for which its members ideally constitute an independent and identically distributed (i.i.d.) sample (Kioutsioukis and Galmarini, 2014; Bishop and Abramowitz, 2013). This statistical process favours the ensemble mean at each observation site. Two basic pathways exist to achieve this goal: model weighting or model sub-selecting. There are several methods to assign weights to ensemble members such as the singular value decomposition (Pagowski et al., 2005), dynamic linear regression (Pagowski et al., 2006; Djalalova et al., 2010), Kalman filtering (Delle Monache et al., 2011), Bayesian model averaging (Riccio et al., 2007; Monteiro et al., 2013) and analytical optimization (Potempski and Galmarini, 2009) while model selection usually relies on the quadratic error or its proxies, in time (e.g. Solazzo et al., 2013; Kioutsioukis and Galmarini, 2014) or frequency space (Galmarini et al., 2013). The majority of those ensemble studies focuses on $O_3$ and only recently the studies also involve particulate matter (Djalalova et al., 2010; Monteiro et al., 2013).

In this work, we apply and intercompare both approaches (weighting and sub-selecting) using the Air Quality Model Evaluation International Initiative (AQMEII) datasets from phase I and phase II. The ensemble approaches are evaluated against ground level observations from the EMEP and Airbase databases, focusing on the pollutants $O_3$, NO$_2$ and PM$_{10}$ that exhibit different levels of forecast skill. The differences between the multi-model ensembles of phase I (hereafter AQMEII-I) and phase II (hereafter AQMEII-II) originate from many sources, related to both the input data and the models: (a) the simulated years are different (2006 vs. 2010), therefore the meteorological conditions are different; (b) emission methodologies have changed; (c) boundary conditions are very different; (d) the composition of the ensembles is different; (e) the models in AQMEII-II use on-line coupling between meteorology and chemistry; (f) the models may have been updated with new science processes apart from feedback processes. The uncertainties arising from observational errors are not taken into consideration.

In spite of these differences we consider the analysis of the two sets of ensembles revealing. In detail, the objectives of the paper are (a) to interpret the skill of the unconditional multi-model mean within AQMEII-I and AQMEII-II (b) to calculate the maximum expectations in the skill of alternative ensemble estimators and (c) to evaluate the operational implementation of the approaches using cross-validation. The originality 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. Therefore we carry out an analysis of the performance of different ensemble techniques rather than a comparison of the results from the two phases of the AQMEII activity.

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

2 Minimization of the ensemble error

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

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

\[
\text{MSE}(\bar{f}) = \left( \frac{1}{M} \sum_{i=1}^{M} (f_i - \mu) \right)^2 + \frac{1}{M} \left( \sum_{i=1}^{M} (f_i - \mu)^2 \right) + \left( 1 - \frac{1}{M} \right) \frac{1}{M(M-1)} \sum_{i=1}^{M-1} \sum_{j=i+1}^{M} (f_i - \mu)(f_j - \mu)
\]

Eq.1

The decomposition provides the reasoning behind ensemble averaging: as we include more ensemble members, the variance factor is monotonically decreasing and the MSE converges towards the covariance factor. Covariance, unlike the other two positive definite terms, can be
either positive or negative; its minimization requires an ensemble composed by independent
or even better, negatively correlated members. In addition, bias correction should be a
necessary step prior to any ensemble manipulation. More details regarding this decomposition
within the air quality ensembles context can be found in Kioutsioukis and Galmarini, 2014.
In a similar fashion, the squared error of the multi-model ensemble mean can be decomposed
into the difference of two positive-definite components, with their expectations characterized
as accuracy and diversity (Krogh and Vedelsby, 1995):

\[
MSE(\bar{f}) = \frac{1}{M} \sum_{i=1}^{M} (f_i - \mu)^2 - \frac{1}{M} \sum_{i=1}^{M} (f_i - \bar{f})^2
\]

Eq. 2

This decomposition proves that the error of the ensemble mean is guaranteed to be less than
or equal to the average quadratic error of the component models. The minimum ensemble
error depends on the right trade-off between accuracy (1\textsuperscript{st} term on the r.h.s. of Eq. 2) and
diversity (2\textsuperscript{nd} term on the r.h.s. of Eq. 2). If the evaluation is applied on multiple sites, then the
equations 1 and 2 should be replaced with their expectations over the stations.
An error decomposition approach can also be applied on the spectral components (SC) of the
observed and modelled time-series. The data can be spectrally decomposed with the
Kolmogorov-Zurbenko (kz) filter (Zurbenko, 1986) while the original time-series can be
obtained with the linear combination of the spectral components. Assuming the pollution data
at the frequency domain yields N principal spectral bands, the squared error of the multi-
model ensemble mean can be broken down into N\textsuperscript{2} components (Galmarini et al., 2013;
Solazzo and Galmarini, 2016):

\[
MSE(\bar{f}) = \sum_{i=1}^{N} MSE(SC_{f_i}) + \sum_{i \neq j} Cov(SC_{f_i}, SC_{f_j})
\]

Eq. 3

This decomposition shows that the error of the ensemble mean could be split into the sum of
N errors associated with different parts of the spectrum (1\textsuperscript{st} term), provided the spectral
components are independent (the covariance term is zero). The minimization of the error at
each spectral band can be achieved with another approach such as the decompositions
presented in Eq.1 and Eq.2.
The three decompositions presented assume uniform ensembles, i.e. all members receive equal weight. For the case of a non-uniform ensemble, the MSE of the multi-model ensemble mean can be analytically minimized to yield the optimal weights, provided that the participating models are bias-corrected (Potempski and Galmarini, 2009):

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

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

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

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

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

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

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

• Unconditional ensemble mean ($mme$)

• Conditional (on selected members) ensemble mean in time domain ($mme^<$): the optimal trade-off between accuracy and diversity (equation 2) is identified across all possible combinations of the available M models (Kioutsioukis and Galmarini, 2014).
The number of members in the ensemble combination that gives the minimum error will be used as the effective number of models \( (N_{\text{EFF}}) \) rather than its estimate based on the independent components of the ensemble (eq. 5).

- Conditional (on selected members) ensemble mean in frequency domain \((k_{\text{FO}})\): following equation 3, an ensemble estimator is synthesized from the best member at each spectral band (Galmarini et al., 2013). The original time-series are decomposed into four spectral components (see Appendix I), namely the intra-diurnal, diurnal, synoptic and long-term component, using the Kolmogorov-Zurbenko filter (Zurbenko, 1986).

- Conditional (on selected members) ensemble mean in frequency domain \((k_{\text{HO}})\): it is an extension of the \(k_{\text{FO}}\), where the spectral components of the ensemble estimator are averaged from \(N_{\text{EFF}}\) members at each spectral band (rather than the best).

- Conditional (optimally weighted) ensemble mean \((mmW)\): according to equation 4 (Potempski and Galmarini, 2009).

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

## 3 Setup: experiments, models and observations

The two AQMEII ensemble datasets have simulated the air quality for Europe \([-10.39]W; (30,65]N\) and North America \([-125,-55]W; (26,51]N\). Despite the common domains, the modelling systems across the two phases have profound differences. The simulation year was 2006 for AQMEII-I and 2010 for AQMEII-II, therefore the two sets are dissimilar with respect to the input data (emissions, chemical boundary conditions, meteorology). Boundary conditions are obtained from GEMS (Global and Regional Earth-System Monitoring using Satellite and in-situ data) in AQMEII-I and MACC (Monitoring Atmospheric Composition &
Climate) in AQMEII-II. The air quality models of the second phase are coupled with their meteorological driver (chemistry feedbacks on meteorology), while those of the first phase are not. The participating models are also different. Detailed analysis of the emissions, boundary conditions and meteorology for the modelled year 2006 (AQMEII-I) is presented in Pouliot et al. (2012), Schere et al. (2012) and Vautard et al. (2012). For 2010 (AQMEII-II), similar information is presented in Pouliot et al. (2015), Giordano et al. (2015) and Brunner et al. (2015).

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

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

Table 1 summarises the features of the modelling systems analysed in this study with regard to \( \text{O}_3 \), \( \text{NO}_2 \) and \( \text{PM}_{10} \) concentrations in the EU. The modelling contribution to the two phases of AQMEII consists of 12, 13 and 10 models for \( \text{O}_3 \), \( \text{NO}_2 \) and \( \text{PM}_{10} \) respectively in AQMEII-I, while 14 members were available for all species in AQMEII-II. Several discrete simulations of WRF-Chem with alternative chemistry and physics configurations are included in AQMEII-II (Forkel et al. 2015, San José et al, 2015, Baró et al., 2015).

Following the statements of section 2, each model has been bias-corrected prior to the analysis, i.e. its own mean bias over the examined three-month period has been subtracted from its modelled time-series at each monitoring site. For each modelling system, its long-term systematic error is a known quantity estimated during its validation stage; therefore the
subtraction of the seasonal bias does not restrict the generality of the study. Actually, the requirement for bias removal is a necessary condition only for the weighted ensemble mean. In the results section we will address this issue and its effect on the skill of the ensemble estimators.

The observational data sets for O$_3$, NO$_2$ and PM$_{10}$ derived from the surface AQ monitoring networks operating in the EU constitutes the same data set used in the first and second phases of AQMEII to support model evaluation. All monitoring stations are rural and have data at least 75% of the time. The network is denser for O$_3$ (451/450 stations in AQMEII-I/II) for which there are as many monitoring stations as for NO$_2$ (290/337 stations in AQMEII-I/II) and PM$_{10}$ (126/131 stations in AQMEII-I/II) combined, with PM$_{10}$ having the fewest observations. Figure 1 compares the statistical distribution of all three species between the two AQMEII phases, through the cumulative density function composed from the mean value at each percentile of the observations. The Kolmogorov-Smirnov test (Massey, 1951) yields that only the PM$_{10}$ distributions differ at the 1% significance level. It results from the unavailability of data for France and UK in AQMEII-II for PM$_{10}$ (station locations are shown in Figure 3).

4 Results

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

From the provided station-based hourly time-series, we analysed one season (three-monthly period) with continuous data and relatively high concentrations; for O$_3$, June-July-August was selected while September-October-November is used for NO$_2$ and PM$_{10}$.

4.1 Single Models

The distributions of each model’s NMSE for O$_3$, NO$_2$ and PM$_{10}$ over all monitoring stations are presented in Figure 2 as box-and-whisker plots. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25$^{th}$ and 75$^{th}$ percentiles,
respectively. The whiskers extend to the most extreme data points not considered outliers (i.e. points with distance from the 25\textsuperscript{th} and 75\textsuperscript{th} percentiles smaller than 1.5 times the interquartile range). Among the examined pollutants, the models simulate better the O\textsubscript{3} concentrations, as is evident from the axis scale. The highest variability in the skill between and within the models is observed for NO\textsubscript{2}.

The distribution of average NMSE at each station (\langle NMSE \rangle) has a median on the order of 0.1 for O\textsubscript{3} and 0.5 for NO\textsubscript{2} and PM\textsubscript{10} for both phases (Table 2). The application of the Kolmogorov-Smirnov test (Massey, 1951) on the \langle NMSE \rangle distributions across AQMEII-I and AQMEII-II shows that there are no statistically significant differences in the \langle NMSE \rangle distributions between the two ensemble datasets at the 1\% significance level. The same also applies for the statistical distribution of the minimum NMSE at each station (NMSE\textsubscript{BEST}) at each monitoring station. Hence, despite the different modelling systems and input data, the \langle NMSE \rangle and NMSE\textsubscript{BEST} distributions between AQMEII-I and AQMEII-II are indistinguishable for the three examined pollutants.

Besides \langle NMSE \rangle and NMSE\textsubscript{BEST}, we evaluate the percentage of cases each model has been identified as being ‘best’ and calculate the coefficient of variation (CoV=\text{std/mean}) of this index for each ensemble. If models were behaving like i.i.d., the probabilities of being best would be roughly equal (~1/M) for all models and the CoV would generally be well below unity for the examined range of ensemble members. As can be inferred from Table 2, the proportion of equally good models is higher for O\textsubscript{3} and NO\textsubscript{2} in the 2\textsuperscript{nd} dataset. Among the pollutants, the CoV of NO\textsubscript{2} exhibits the most dramatic change.

\subsection*{4.2 Pitfalls of the unconditional multi-model mean}

The skill of the multi-model mean has been compared against the skill of the best deterministic model, independently evaluated at each monitoring site (hereafter bestL). The geographical distribution of the ratio RMSE(mme)/RMSE\textsubscript{BESTMODEL} is presented in Figure 3. The indicator does not exhibit any longitudinal or latitudinal dependence. Summary statistics indicate that the mme outscores the bestL at roughly half of the stations for O\textsubscript{3} (namely 52/49 for AQMEII-I/II) and at approximately 40\% of the stations for PM\textsubscript{10} (38/42). The same statistic for NO\textsubscript{2} varies considerably (39/64). The Kolmogorov-Smirnov test yields that the corresponding distributions (pI/pII) are different at the 1\% significance level but the t-test demonstrates that the mean of the distributions differ significantly only for NO\textsubscript{2}. The reason
behind the skill of mme with respect to the bestL is investigated next with respect to the skill difference and the error dependence of each ensemble.

The skill difference between the best model and the average skill is inferred from the indicator $\text{NMSE}_{\text{BEST}} / \langle \text{NMSE} \rangle$ (Table 2). High values of the indicator correspond to small skill differences between the ensemble members (desirable). The distribution of the $\text{NMSE}_{\text{BEST}} / \langle \text{NMSE} \rangle$ at each station has a median on the order of 0.6-0.8, variable with respect to the dataset and the pollutant. The spread of the indicator, measured by its interquartile range, is higher for NO$_2$ and lower for O$_3$.

The eigenvalues of the covariance matrix calculated from the model errors provides information on the members’ diversity and the ensemble redundancy (Eq. 5). Following the eigen-analysis of the error covariance matrix at each station separately and converting the eigenvalues to cumulative amount of explained variance, the resulting matrix is presented into box and whisker plot (Figure 4). The error dependence of the ensemble members is deduced from the explained variation by the maximum eigenvalue $s_m$. Low values of the indicator corresponds to independent members with small error dependence (desirable). The average variation explained by $s_m$ ranges between 65% and 79%, taking the lower values for NO$_2$. The spread of the indicator, measured by its interquartile range, is higher for NO$_2$ and lower for O$_3$.

All species demonstrate smaller skill difference and higher error dependence in the AQMEII-II dataset. The Kolmogorov-Smirnov test yielded the difference in the corresponding distributions of the indicators between AQMEII-I and AQMEII-II is significant at the 1% level. However, it is the joint distribution of skill difference and error dependence that modulates the mme skill with respect to the bestL, as seen in Figure 5. Shifts in the distributions of the indicators at opposite directions eventually cancel out, yielding no change in the mme skill. This case is observed for O$_3$ and PM$_{10}$. For NO$_2$, skill difference was improved more than error dependence was worsened, yielding a net improvement of mme in AQMEII-II.

The area below the diagonal in Figure 5 corresponds to monitoring sites with disproportionally low diversity under the current level of accuracy. This area of the chart indicates high spread in skill difference and relatively highly dependent errors. This situation practically means a limited number of skilled models with correlated errors, which in turn denotes a small $N_{\text{EFF}}$ value as demonstrated in Figure 6. The opposite state is true for the area above the diagonal. It
corresponds to locations that are constituted from models with comparable skill and relatively
independent errors, reflecting a high $N_{\text{EFF}}$ value. This matches the desired synthesis for an
ensemble.

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

4.3 Conditional multi-model mean

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

The global skill of all the single models and the ensemble estimators, evaluated at all stations,
are presented in Figure 7 in the form of Taylor plots. For $O_3$, the deterministic models have
standard deviations that are smaller compared to observations and a narrow correlation pattern
(~0.7) that is slightly deteriorated in AQMEII-II. For $NO_2$, members with higher variance -as
well as lower- than the observed variance exist in the ensemble while the correlation spread is
becoming narrower in AQMEII-II and demonstrates a minor improvement. Last, simulated
$PM_{10}$ from the deterministic models displays smaller standard deviation compared to
observations with a wide correlation spread (0.3-0.6). The multi-model mean is always found
closer to the reference point, in an area that incorporates lower error and increased correlation
but at the same time generally low variance. The examined ensemble estimators (mmW, mme<, kzFO, kzHO) are horizontally shifted from mme, hence they demonstrate even lower error and increased correlation and variance. Among them, the highest composite skill was found for mmW, followed by kzHO.

A comparison between the skill of the examined ensemble estimators versus the mme and the best single model is now conducted (Table 3). The best single model is evaluated globally (bestG: average across all stations) and locally (bestL: at each station separately). The former estimates the best average deterministic skill among the candidate models; the latter provides a useful indicator for controlling whether the anticipated benefits of ensemble averaging holds. The skill scores have been evaluated against the guaranteed minimum gain of the ensemble (<MSE>), the ensemble mean (mme) and the best single model globally (bestG).

The estimations calculated from the unprecedented AQMEII datasets (2 years of hourly measurements and simulations from 2 different ensembles of 10-14 models each at over 450 stations for 3 pollutants) allows the following interpretation:

- The mme always achieves lower error than bestG. The advancement is higher for O₃ (9-22%), followed by NO₂ (7-9%) while the PM₁₀ demonstrate the least skill improvement (1-3%). With respect to bestL, the mme generally attains similar or slightly higher MSE. Hence, the average error over multiple stations statistically favours the ensemble mean over the single models but the comparison at each site generally does not as it depends on the skill difference and the error dependence of the models.

- The skill score of mme over <MSE> (i.e., the guaranteed upper ceiling for the MSE of mme, from eq. 2) ranges between 15% and 30%, higher for NO₂ and lower for PM₁₀. According to eq. 2, this number also represents the diversity as percentage of the accuracy. Therefore, besides improving the single models, their combination in an ensemble confines the mme skill if their diversity is limited.

- The skill score of the examined ensemble estimators (mmW, mme<, kzFO, kzHO) over <MSE> ranges between 25% and 50%, higher for O₃ and NO₂ and lower for PM₁₀. Among them, the improvement is higher for mmW and lower for mme< and kzFO. Thus, the promotion of accuracy and diversity within the ensemble almost doubles the distance to <MSE> compared to mme and results in an additional skill over the mme between 14% and 31% (for mmW).
The improvement of the ensemble estimator using selected $N_{\text{EFF}}$ members ($mme<^i$) over all members ($mme$) is illustrated in Figure 8 in the context of skill difference and error dependence. The charts demonstrate no points below the diagonal, i.e. the sub-selection results in an ensemble constituted from models with comparable skill and relatively independent errors (compared to the full ensemble).

The theoretical minimum MSE of $mme$ for the case of unbiased and uncorrelated models (from eq. 1) is far from being achieved from all ensemble estimators.

The statistical distributions of the skill scores of the examined ensemble estimators ($mmW$, $mme<^i$, $kzFO$, $kzHO$) over $mme$ are well bounded from above to lower than unity values (Figure 9). The only exception exists for roughly 10% of the stations, for all pollutants, where $kzFO$ demonstrates higher MSE compared to $mme$. Unlike the other ensemble estimators, $kzFO$ utilises independent spectral components each obtained from a single model, eliminating the possibility for ‘cancelling out’ of random errors. All cases belonging to this 10% of the samples (lower tail of the cdf) demonstrate high $N_{\text{EFF}}$, where the benefits from unconditional ensemble averaging are optimal (Kioutsioukis and Galmarini, 2014). Contrary, for another 10% of the stations (upper tail of the cdf), there is an abrupt improvement from the conditional ensemble estimators. Those cases demonstrate low $N_{\text{EFF}}$, where the benefits from unconditional ensemble averaging are minimal.

The ability to simulate extreme values is now examined through the POD and FAR indices. Two thresholds were utilised for each pollutant, being 120 and 180 $\mu g/m^3$ for $O_3$, 25 and 50 $\mu g/m^3$ for $NO_2$ and 50 and 90 $\mu g/m^3$ for $PM_{10}$. The average 90$^{\text{th}}$ percentile across the stations was 129/117 $\mu g/m^3$ (AQMEII-I/II) for $O_3$, 30/26 $\mu g/m^3$ for $NO_2$ and 52/33 $\mu g/m^3$ for $PM_{10}$ (Figure 1). Hence, the thresholds fall into the upper 10% of the distributions, being even more extreme for $PM_{10}$ in AQMEII-II. The numbers in Table 4 give rise to the following inferences:

- For $O_3$ and $NO_2$, $mme$ achieves somewhat higher POD than $bestG$ at the lower threshold but the order is reversed at the higher threshold. For $PM_{10}$, $bestG$ always performs better than $mme$ for values exceeding the lower threshold. As we move towards the tail, the POD of $bestG$ dominates over the $mme$. Thus, the ranking of the $mme$ and $bestG$ at the extreme percentiles and on average (seen earlier) are opposite.

- The $mme<^i$ generally achieves somewhat higher POD than $bestL$ at the lower threshold but the order is reversed at the higher threshold. Over that level, $kzFO$ and $mmW$ are the only estimators with POD higher than $bestL$. 


As we move towards higher percentiles, the 1st order spectral model ($kzFO$) has higher POD than the higher-order spectral model ($kzHO$) due to the averaging in the latter. In addition, the frequency domain averaging ($kzHO$) had slightly higher POD compared to the time domain averaging ($mme<$).

The $mmW$, besides its lower MSE, has the highest POD among all models and ensemble estimators.

The variation of FAR was very small between all examined models and ensemble estimators.

The combination of the results from the average error and the extremes identifies $mmW$ as the estimator that outscores the others across all percentiles. $kzFO$ has high capacity for extremes but requires attention for the limited sites with high $N_{EFF}$, where its skill is inferior to $mme$.

$kzHO$ and $mme<$ have both high skill across all percentiles (better for $kzHO$) but they could have inferior POD compared to best$L$ at the extreme percentiles. With respect to the pollutants, the advancement of $mmW$ skill over $mme$ was higher for $O_3$.

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

The interquartile range of the day-to-day difference in the weights is calculated and its range over all stations is displayed in Figure 10. No convergence occurs, however the variability of the $mmW$ weights is notably reduced after a certain amount of time. If we set a tolerance level at the second decimal, to be satisfied at all stations, we need at a minimum 20-45 days of hourly time-series. The variability of weights is smaller for $O_3$ and higher for $NO_2$ and $PM_{10}$, explained by the larger NMSE spread in the latter case. The identification of the necessary
training or learning period will be assessed by its effect on the $mmW$ skill. Table 5 presents the $mmW$ skill obtained from training over time series of different lengths varying from 5 to 60 days. For $O_3$, $mmW$ trained over 10 days yields similar results with $mme$ while longer periods result in large departures from $mme$. NO$_2$ and PM$_{10}$ require larger training periods than $O_3$. The use of $mmW$ is practically of no benefit compared to $mme$ if the training period is less than 20 days for NO$_2$ and 30 days for PM$_{10}$. For all pollutants, the variability of the weights and the bias has no effect in the error after 60 days.

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

5 Conclusions

In this paper we analyze two independent suites of chemical weather modelling systems regarding their effect in the skill of the ensemble mean ($mme$). The results are interpreted with respect to the error decomposition of the $mme$. Four ways to extract more information from an ensemble besides the $mme$ are ultimately investigated and evaluated. The first approach applies optimal weights to the models of the ensemble ($mmW$) and the other three methods utilise selected members in time ($mme<$) or frequency ($kzFO$, $kzHO$) domain. The study focuses on $O_3$, NO$_2$ and PM$_{10}$, using the unprecedented datasets from two phases of AQMEII over the European domain.

The comparison of the $mme$ skill versus the globally best single model ($bestG$: identified from the evaluation over all stations), points out that $mme$ achieves lower average (across all stations) error compared to $bestG$. The enhancement of accuracy is highest for $O_3$ (up to 22%) and lowest for PM$_{10}$ (below 3%). We then investigate whether this benefit of ensemble averaging of air quality time series holds at each station by direct comparison between the
Summary statistics indicate that the mme outscores the bestL at roughly 50% of the stations for O₃ and at approximately 40% of the stations for PM₁₀, while for NO₂ the values were about 40% and 60% for the two datasets. This result indicates that there is a considerable amount of stations (over 40%) where the unconditional averaging is not advantageous because the ensemble does not meet the necessary conditions. A new chart has been introduced in this paper that interprets the skill of the mme according to the skill difference and the error dependence of the ensemble members.

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

- The skill score of mme over its guaranteed upper ceiling (case of zero diversity) ranges between 15% and 30%, being lower for PM₁₀. Those percentages also represent the diversity normalized by the accuracy. Therefore, besides improving the single models, their combination in an ensemble confines the mme skill if their diversity is limited.

- The promotion of the right amount of accuracy and diversity in the conditional ensemble estimators almost doubles the distance to the guaranteed upper ceiling. The skill score over mme is higher for O₃ (in the range 18%-31%) and lower for NO₂ and PM₁₀ (in the range 8%-25%), associated to the extent of potential changes in the joint distribution of accuracy and diversity in the respective ensembles. The improvement is larger for mmW and smaller for mme< and kzFO.

- The theoretical minimum MSE of mme for the case of unbiased and uncorrelated models is far from being achieved from all ensemble estimators.

- As we move towards the tail, the probability of detection (POD) of bestG (bestL) dominates over the mme (mme<). At the extreme percentiles, kzFO and mmW are the only estimators with POD higher than bestL.

- The combination of the results from the average error and the extremes identifies mmW as the estimator that outscores the others across all percentiles. kzFO has high capacity for extremes but requires attention for the limited sites with high N_{EFF}, where its skill is inferior to mme. kzHO and mme< have both high skill across all percentiles (better for kzHO) but they could have inferior POD compared to bestL at the extreme percentiles.
The skill enhancement is superior using the weighting scheme but the required training period
to acquire representative weights was longer compared to the sub-selecting schemes. For all
pollutants, the variability of the weights and the bias has negligible effect in the error for
training periods longer than 60 days. For the schemes relying in member selection, accurate
recent representations on the order of a week were sufficient. The learning periods constitute
the necessary time to acquire similar prior and posterior distributions in the controlling
parameters of samples. The risks of all the statistical learning processes originate from the
violation of this assumption, which holds for example in the case of changing weather or
chemical regimes. Therefore, the operational implementation of each ensemble approach
requires knowledge of its safety margins for the examined pollutants as well as its risks.

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


Acknowledgements

We gratefully acknowledge the contribution of various groups to the second air Quality Model Evaluation international Initiative (AQMEII) activity: U.S. EPA, Environment Canada, Mexican Secretariat of the Environment and Natural Resources (Secretaría de Medio Ambiente y Recursos Naturales-SEMARNAT) and National Institute of Ecology (Instituto Nacional de Ecología-INE) (North American national emissions inventories); U.S. EPA (North American emissions processing); TNO (European emissions processing); ECMWF/MACC project & Météo-France/CNRM-GAME (Chemical boundary conditions). Ambient North American concentration measurements were extracted from Environment Canada's National Atmospheric Chemistry Database (NAtChem) PM database and provided by several U.S. and Canadian agencies (AQS, CAPMoN, CASTNet, IMPROVE, NAPS, SEARCH and STN networks); North American precipitation-chemistry measurements were extracted from NAtChem's precipitation-chemistry data base and were provided by several U.S. and Canadian agencies (CAPMoN, NADP, NBPMN, NSPSN, and REPQ networks); the WMO World Ozone and Ultraviolet Data Centre (WOUDC) and its data-contributing agencies provided North American and European ozone sonde profiles; NASA's AErosol RObotic NETwork (AeroNet) and its data-contributing agencies provided North American and European AOD measurements; the MOZAIC Data Centre and its contributing airlines provided North American and European aircraft takeoff and landing vertical profiles; for European air quality data the following data centers were used: EMEP European Environment Agency/European Topic Center on Air and Climate Change/AirBase provided European air-and precipitation-chemistry data. The Finish Meteorological Institute is acknowledged for providing biomass burning emission data for Europe. Data from meteorological station monitoring networks were provided by NOAA and Environment Canada (for the US and Canadian meteorological network data) and the National Center for Atmospheric Research (NCAR) data support section. Joint Research Center Ispra/Institute for Environment and Sustainability provided its ENSEMBLE system for model output harmonization and analyses and evaluation. The co-ordination and support of the European contribution through COST Action ES1004 EuMetChem is gratefully acknowledged. The views expressed here are those of the authors and do not necessarily reflect the views and policies of the U.S. Environmental Protection Agency (EPA) or any other organization participating in the AQMEII project. This paper has been subjected to EPA review and approved for publication. The UPM authors thankfully acknowledge the computer resources, technical expertise and assistance provided...
by the Centro de Supercomputación y Visualización de Madrid (CESVIMA) and the Spanish
Supercomputing Network (BSC). GC and PT were supported by the Italian Space Agency
(ASI) in the frame of PRIMES project (contract n. I/017/11/0). The same authors are deeply
thankful to the Euro Mediterranean Centre on Climate Change (CMCC) for having made
available the computational resources.
The relevant separate scales of motion are defined by means of physical considerations and periodogram analysis (Rao et al., 1997). They are namely the intra-day component (ID), the diurnal component (DU), the synoptic component (SY) and the long-term component (LT).

The hourly time series (S) can therefore be decomposed as:

\[
S(t) = ID(t) + DU(t) + SY(t) + LT(t)
\]  \hspace{1cm} (1)

where:

\[
ID(t) = S(t) - KZ_{3,3}
\]

\[
DU(t) = KZ_{3,3} - KZ_{13,5}
\]

\[
SY(t) = KZ_{13,5} - KZ_{103,5}
\]

\[
LT(t) = KZ_{103,5}
\]  \hspace{1cm} (2)
Table 1. The modelling systems participating in the first and second phases of AQMEII for Europe.

<table>
<thead>
<tr>
<th>Model</th>
<th>Grid</th>
<th>Emissions</th>
<th>Chemical BC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Met</td>
<td>AQ</td>
<td>Grid</td>
<td></td>
</tr>
<tr>
<td>MM5</td>
<td>DEHM</td>
<td>50 km</td>
<td>Global emission databases, EMEP</td>
</tr>
<tr>
<td>MM5</td>
<td>Polyphemus</td>
<td>24 km</td>
<td>Standard§</td>
</tr>
<tr>
<td>MM5</td>
<td>Chimere</td>
<td>25 km</td>
<td>MEGAN, Standard§</td>
</tr>
<tr>
<td>MM5</td>
<td>CAMx</td>
<td>15 km</td>
<td>MEGAN, Standard§</td>
</tr>
<tr>
<td>PARLAM-PS</td>
<td>EMEP</td>
<td>50 km</td>
<td>EMEP model</td>
</tr>
<tr>
<td>WRF</td>
<td>CMAQ</td>
<td>18 km</td>
<td>From ECMWF, and forecasts</td>
</tr>
<tr>
<td>WRF</td>
<td>Chem</td>
<td>22.5 km</td>
<td>Fixed</td>
</tr>
<tr>
<td>ECMWF</td>
<td>SILAM</td>
<td>24 km</td>
<td>Standard anthropogenic; In-house biogenic</td>
</tr>
<tr>
<td>ECMWF</td>
<td>Lotos-EUROS</td>
<td>25 km</td>
<td>Standard§</td>
</tr>
<tr>
<td>GEM</td>
<td>GEM-AQ</td>
<td>25 km</td>
<td>Standard (AQMEII region); EDGAR/GEIA</td>
</tr>
<tr>
<td>COSMO</td>
<td>Muscat</td>
<td>24 km</td>
<td>Global variable grid setup (no boundary conditions)</td>
</tr>
<tr>
<td>COSMO-CLM</td>
<td>CMAQ</td>
<td>24 km</td>
<td>Standard§</td>
</tr>
<tr>
<td>WRF</td>
<td>Chem</td>
<td>23 km</td>
<td>Standard</td>
</tr>
<tr>
<td>WRF</td>
<td>CMAQ</td>
<td>18 km</td>
<td>Standard</td>
</tr>
<tr>
<td>COSMO</td>
<td>Cosmo-ART</td>
<td>0.22°</td>
<td>Standard</td>
</tr>
<tr>
<td>COSMO</td>
<td>Muscat</td>
<td>0.25°</td>
<td>Standard</td>
</tr>
<tr>
<td>NMMB</td>
<td>BSCCTM</td>
<td>0.20°</td>
<td>Standard</td>
</tr>
<tr>
<td>RACMO</td>
<td>LOTOS-EUROS</td>
<td>0.5° x 0.25°</td>
<td>Standard</td>
</tr>
<tr>
<td>MetUM</td>
<td>UKCA RAQ</td>
<td>0.22°</td>
<td>Standard</td>
</tr>
</tbody>
</table>

AQMEII phase I

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.

§ 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.

AQMEII phase II

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

<table>
<thead>
<tr>
<th></th>
<th>O\textsubscript{3}</th>
<th>O\textsubscript{3}</th>
<th>NO\textsubscript{2}</th>
<th>NO\textsubscript{2}</th>
<th>PM\textsubscript{10}</th>
<th>PM\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(I/II)</td>
<td>(I/II)</td>
<td>(I/II)</td>
<td>(I/II)</td>
<td>(I/II)</td>
<td>(I/II)</td>
</tr>
<tr>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
</tr>
<tr>
<td>5\textsuperscript{th}</td>
<td>0.04 / 0.04</td>
<td>0.03 / 0.03</td>
<td>0.28 / 0.23</td>
<td>0.17 / 0.18</td>
<td>0.30 / 0.27</td>
<td>0.20 / 0.20</td>
</tr>
<tr>
<td>25\textsuperscript{th}</td>
<td>0.07 / 0.07</td>
<td>0.05 / 0.05</td>
<td>0.39 / 0.35</td>
<td>0.24 / 0.25</td>
<td>0.40 / 0.39</td>
<td>0.26 / 0.28</td>
</tr>
<tr>
<td>50\textsuperscript{th}</td>
<td>0.10 / 0.10</td>
<td>0.07 / 0.08</td>
<td>0.52 / 0.49</td>
<td>0.33 / 0.34</td>
<td>0.47 / 0.51</td>
<td>0.34 / 0.37</td>
</tr>
<tr>
<td>75\textsuperscript{th}</td>
<td>0.15 / 0.15</td>
<td>0.11 / 0.12</td>
<td>0.82 / 0.76</td>
<td>0.48 / 0.50</td>
<td>0.61 / 0.62</td>
<td>0.46 / 0.50</td>
</tr>
<tr>
<td>95\textsuperscript{th}</td>
<td>0.24 / 0.23</td>
<td>0.18 / 0.18</td>
<td>1.69 / 1.49</td>
<td>0.81 / 0.93</td>
<td>1.02 / 0.98</td>
<td>0.73 / 0.81</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>O\textsubscript{3}</th>
<th>O\textsubscript{3}</th>
<th>NO\textsubscript{2}</th>
<th>NO\textsubscript{2}</th>
<th>PM\textsubscript{10}</th>
<th>PM\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
</tr>
<tr>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
</tr>
<tr>
<td>5\textsuperscript{th}</td>
<td>0.50</td>
<td>0.60</td>
<td>0.36</td>
<td>0.45</td>
<td>0.49</td>
<td>0.63</td>
</tr>
<tr>
<td>25\textsuperscript{th}</td>
<td>0.62</td>
<td>0.70</td>
<td>0.50</td>
<td>0.62</td>
<td>0.61</td>
<td>0.72</td>
</tr>
<tr>
<td>50\textsuperscript{th}</td>
<td>0.70</td>
<td>0.76</td>
<td>0.61</td>
<td>0.72</td>
<td>0.70</td>
<td>0.79</td>
</tr>
<tr>
<td>75\textsuperscript{th}</td>
<td>0.76</td>
<td>0.82</td>
<td>0.72</td>
<td>0.81</td>
<td>0.85</td>
<td>0.85</td>
</tr>
<tr>
<td>95\textsuperscript{th}</td>
<td>0.83</td>
<td>0.88</td>
<td>0.87</td>
<td>0.93</td>
<td>0.92</td>
<td>0.92</td>
</tr>
<tr>
<td>mean</td>
<td>0.69</td>
<td>0.75</td>
<td>0.61</td>
<td>0.70</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>O\textsubscript{3}</th>
<th>O\textsubscript{3}</th>
<th>NO\textsubscript{2}</th>
<th>NO\textsubscript{2}</th>
<th>PM\textsubscript{10}</th>
<th>PM\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
</tr>
<tr>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
<td>&lt;NMSE&gt;</td>
<td>NMSE\textsubscript{BEST}</td>
</tr>
<tr>
<td>5\textsuperscript{th}</td>
<td>0.69</td>
<td>0.75</td>
<td>0.61</td>
<td>0.70</td>
<td>0.72</td>
<td>0.77</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>O\textsubscript{3}</th>
<th>O\textsubscript{3}</th>
<th>NO\textsubscript{2}</th>
<th>NO\textsubscript{2}</th>
<th>PM\textsubscript{10}</th>
<th>PM\textsubscript{10}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
<td>(I)</td>
<td>(II)</td>
</tr>
<tr>
<td>CoV</td>
<td>1.08</td>
<td>0.70</td>
<td>1.42</td>
<td>0.65</td>
<td>1.16</td>
<td>1.53</td>
</tr>
</tbody>
</table>
Table 3. The MSE from (a) the best deterministic models, globally (bestG) and locally (bestL), (b) the unconditional ensemble mean (mme) and (c) the four conditional ensemble estimators (mme<, kzFO, kzHO, mmW). In addition, the bounds for the MSE of the ensemble mean are also presented. The maximum value (<MSE>) arises for ensemble members without diversity and the minimum value (mmeMIN) has been estimated from the variance term only (i.e. calculated for unbiased and uncorrelated ensemble members). The ability of the estimators is evaluated through their skill scores (SS<sub>REF</sub>=1-MSE/MSE<sub>REF</sub>, REF=bestG, <MSE>, mme).

<table>
<thead>
<tr>
<th>O3 (I)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
<th>O3 (II)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
</tr>
<tr>
<td>bestG</td>
<td>641</td>
<td>7%</td>
<td>499</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestL</td>
<td>483</td>
<td>25%</td>
<td>30%</td>
<td>3%</td>
<td></td>
<td>441</td>
<td>12%</td>
<td>24%</td>
<td>3%</td>
</tr>
<tr>
<td>mme</td>
<td>498</td>
<td>22%</td>
<td>28%</td>
<td>454</td>
<td>9%</td>
<td>21%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mme&lt;</td>
<td>398</td>
<td>38%</td>
<td>42%</td>
<td>30%</td>
<td>25%</td>
<td>35%</td>
<td>18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzFO</td>
<td>400</td>
<td>38%</td>
<td>20%</td>
<td>369</td>
<td>26%</td>
<td>36%</td>
<td>19%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzHO</td>
<td>367</td>
<td>43%</td>
<td>47%</td>
<td>26%</td>
<td>30%</td>
<td>40%</td>
<td>23%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmW</td>
<td>345</td>
<td>46%</td>
<td>50%</td>
<td>31%</td>
<td>31%</td>
<td>45%</td>
<td>31%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;MSE&gt;</td>
<td>690</td>
<td>mmeMIN</td>
<td>58</td>
<td></td>
<td></td>
<td>577</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmeMIN</td>
<td>58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NO2 (I)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
<th>NO2 (II)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
</tr>
<tr>
<td>bestG</td>
<td>77</td>
<td>25%</td>
<td>61</td>
<td>20%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestL</td>
<td>70</td>
<td>10%</td>
<td>32%</td>
<td>3%</td>
<td></td>
<td>58</td>
<td>5%</td>
<td>25%</td>
<td>-4%</td>
</tr>
<tr>
<td>mme</td>
<td>72</td>
<td>7%</td>
<td>30%</td>
<td>454</td>
<td>9%</td>
<td>21%</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mme&lt;</td>
<td>63</td>
<td>19%</td>
<td>39%</td>
<td>13%</td>
<td>17%</td>
<td>34%</td>
<td>9%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzFO</td>
<td>62</td>
<td>19%</td>
<td>40%</td>
<td>13%</td>
<td>16%</td>
<td>33%</td>
<td>8%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzHO</td>
<td>59</td>
<td>24%</td>
<td>43%</td>
<td>18%</td>
<td>21%</td>
<td>37%</td>
<td>14%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmW</td>
<td>56</td>
<td>27%</td>
<td>46%</td>
<td>22%</td>
<td>25%</td>
<td>40%</td>
<td>18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;MSE&gt;</td>
<td>104</td>
<td>mmeMIN</td>
<td>8</td>
<td></td>
<td></td>
<td>77</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmeMIN</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PM10 (I)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
<th>PM10 (II)</th>
<th>MSE</th>
<th>SS</th>
<th>SS</th>
<th>SS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
<td></td>
<td></td>
<td>bestG</td>
<td>&lt;MSE&gt;</td>
<td>(mme)</td>
</tr>
<tr>
<td>bestG</td>
<td>341</td>
<td>16%</td>
<td>141</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestL</td>
<td>326</td>
<td>5%</td>
<td>20%</td>
<td>1%</td>
<td></td>
<td>139</td>
<td>2%</td>
<td>15%</td>
<td>0%</td>
</tr>
<tr>
<td>mme</td>
<td>330</td>
<td>3%</td>
<td>19%</td>
<td>3%</td>
<td></td>
<td>139</td>
<td>1%</td>
<td>15%</td>
<td>0%</td>
</tr>
<tr>
<td>mme&lt;</td>
<td>303</td>
<td>11%</td>
<td>25%</td>
<td>8%</td>
<td></td>
<td>121</td>
<td>14%</td>
<td>26%</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>kzFO</td>
<td>kzHO</td>
<td>mmW</td>
<td>&lt;MSE&gt;</td>
<td>mmeMIN</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>------</td>
<td>------</td>
<td>------</td>
<td>-------</td>
<td>--------</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>299</td>
<td>294</td>
<td>284</td>
<td>407</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>%</td>
<td>13%</td>
<td>14%</td>
<td>17%</td>
<td>17%</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>27%</td>
<td>28%</td>
<td>30%</td>
<td>30%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10%</td>
<td>11%</td>
<td>14%</td>
<td>14%</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 (Potempski and Galmarini, 2009)

6.
Table 4. The probability of detection (POD) and false alarm rate (FAR) from (a) the best deterministic models, globally (bestG) and locally (bestL), (b) the unconditional ensemble mean (mme) and (c) the four conditional ensemble estimators (mme<, kzFO, kzHO, mmW). Two thresholds were examined for each indicator, corresponding to tail percentiles.

<table>
<thead>
<tr>
<th></th>
<th>POD</th>
<th>FAR</th>
<th>POD</th>
<th>FAR</th>
<th></th>
<th>POD</th>
<th>FAR</th>
<th>POD</th>
<th>FAR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>O3 (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestG</td>
<td>37.9</td>
<td>3.6</td>
<td>11.4</td>
<td>0.0</td>
<td>bestG</td>
<td>19.9</td>
<td>1.2</td>
<td>1.2</td>
<td>0.0</td>
</tr>
<tr>
<td>bestL</td>
<td>54.7</td>
<td>3.5</td>
<td>19.5</td>
<td>0.0</td>
<td>bestL</td>
<td>33.2</td>
<td>1.5</td>
<td>5.4</td>
<td>0.0</td>
</tr>
<tr>
<td>mme</td>
<td>39.9</td>
<td>2.5</td>
<td>12.0</td>
<td>0.0</td>
<td>mme</td>
<td>22.0</td>
<td>1.2</td>
<td>0.5</td>
<td>0.0</td>
</tr>
<tr>
<td>mme&lt;</td>
<td>53.5</td>
<td>2.6</td>
<td>18.3</td>
<td>0.0</td>
<td>mme&lt;</td>
<td>34.9</td>
<td>1.3</td>
<td>2.4</td>
<td>0.0</td>
</tr>
<tr>
<td>kzFO</td>
<td>57.7</td>
<td>3.0</td>
<td>19.6</td>
<td>0.0</td>
<td>kzFO</td>
<td>39.1</td>
<td>1.5</td>
<td>4.4</td>
<td>0.0</td>
</tr>
<tr>
<td>kzHO</td>
<td>57.1</td>
<td>2.5</td>
<td>19.2</td>
<td>0.0</td>
<td>kzHO</td>
<td>36.9</td>
<td>1.2</td>
<td>2.3</td>
<td>0.0</td>
</tr>
<tr>
<td>mmW</td>
<td>60.6</td>
<td>2.6</td>
<td>27.2</td>
<td>0.0</td>
<td>mmW</td>
<td>45.4</td>
<td>1.6</td>
<td>8.6</td>
<td>0.0</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>NO2 (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestG</td>
<td>45.9</td>
<td>4.6</td>
<td>3.8</td>
<td>0.2</td>
<td>bestG</td>
<td>39.3</td>
<td>3.3</td>
<td>4.9</td>
<td>0.1</td>
</tr>
<tr>
<td>bestL</td>
<td>48.7</td>
<td>4.2</td>
<td>8.5</td>
<td>0.3</td>
<td>bestL</td>
<td>41.4</td>
<td>3.1</td>
<td>8.1</td>
<td>0.1</td>
</tr>
<tr>
<td>mme</td>
<td>49.4</td>
<td>4.6</td>
<td>3.0</td>
<td>0.1</td>
<td>mme</td>
<td>44.4</td>
<td>3.5</td>
<td>5.4</td>
<td>0.1</td>
</tr>
<tr>
<td>mme&lt;</td>
<td>52.2</td>
<td>4.1</td>
<td>7.1</td>
<td>0.1</td>
<td>mme&lt;</td>
<td>47.6</td>
<td>3.2</td>
<td>7.6</td>
<td>0.1</td>
</tr>
<tr>
<td>kzFO</td>
<td>52.7</td>
<td>4.1</td>
<td>8.4</td>
<td>0.1</td>
<td>kzFO</td>
<td>46.5</td>
<td>3.1</td>
<td>9.5</td>
<td>0.1</td>
</tr>
<tr>
<td>kzHO</td>
<td>54.2</td>
<td>4.0</td>
<td>6.8</td>
<td>0.1</td>
<td>kzHO</td>
<td>49.5</td>
<td>3.2</td>
<td>9.3</td>
<td>0.1</td>
</tr>
<tr>
<td>mmW</td>
<td>57.0</td>
<td>4.1</td>
<td>14.8</td>
<td>0.2</td>
<td>mmW</td>
<td>50.9</td>
<td>3.1</td>
<td>13.5</td>
<td>0.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>PM10 (I)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>threshold</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>bestG</td>
<td>25.9</td>
<td>2.7</td>
<td>1.2</td>
<td>0.0</td>
<td>bestG</td>
<td>13.0</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>bestL</td>
<td>27.8</td>
<td>2.3</td>
<td>6.9</td>
<td>1.2</td>
<td>bestL</td>
<td>14.5</td>
<td>0.4</td>
<td>1.6</td>
<td>0.0</td>
</tr>
<tr>
<td>mme</td>
<td>21.6</td>
<td>1.8</td>
<td>0.4</td>
<td>0.0</td>
<td>mme</td>
<td>11.4</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mme&lt;</td>
<td>30.6</td>
<td>2.3</td>
<td>5.6</td>
<td>0.1</td>
<td>mme&lt;</td>
<td>13.9</td>
<td>0.4</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>mme</td>
<td>unconditional ensemble mean</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----</td>
<td>-----------------------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mme&lt;:</td>
<td>conditional ensemble mean (Kioutsioukis and Galmarini, 2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzFO</td>
<td>conditional spectral ensemble mean with 1st order components (Galmarini et al., 2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kzHO</td>
<td>conditional spectral ensemble mean with 2nd and higher order components (kzHO)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mmW</td>
<td>optimal weighted ensemble (Potempski and Galmarini, 2009)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5. The average MSE of $mmW$ for various training lengths, calculated for the testing time-series (i.e. not-used in the training phase) that contains all stations.

<table>
<thead>
<tr>
<th>Length of training period (days)</th>
<th>O$_3$ (I)</th>
<th>O$_3$ (II)</th>
<th>NO$_2$ (I)</th>
<th>NO$_2$ (II)</th>
<th>PM$_{10}$ (I)</th>
<th>PM$_{10}$ (II)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>616</td>
<td>540</td>
<td>90</td>
<td>91</td>
<td>717</td>
<td>210</td>
</tr>
<tr>
<td>10</td>
<td>496</td>
<td>441</td>
<td>77</td>
<td>66</td>
<td>443</td>
<td>150</td>
</tr>
<tr>
<td>20</td>
<td>400</td>
<td>378</td>
<td>65</td>
<td>56</td>
<td>348</td>
<td>125</td>
</tr>
<tr>
<td>30</td>
<td>380</td>
<td>344</td>
<td>62</td>
<td>52</td>
<td>308</td>
<td>109</td>
</tr>
<tr>
<td>40</td>
<td>366</td>
<td>334</td>
<td>59</td>
<td>50</td>
<td>300</td>
<td>113</td>
</tr>
<tr>
<td>50</td>
<td>357</td>
<td>326</td>
<td>57</td>
<td>48</td>
<td>294</td>
<td>108</td>
</tr>
<tr>
<td>60</td>
<td>351</td>
<td>319</td>
<td>56</td>
<td>45</td>
<td>282</td>
<td>102</td>
</tr>
</tbody>
</table>
Figure 1. The Cumulative density functions of the observations (O₃, NO₂, PM₁₀) in the two AQMEII phases (Phase I: filled circles, Phase II: non-filled circles). Each bullet represents the median at the specific percentile.
Figure 2. Model skill difference via the NMSE. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers and the outliers (points with distance from the 25th and 75th percentiles larger than 1.5 times the interquartile range) are plotted individually using the ‘+’ symbol.
Figure 3. Comparison of the mme skill against the best local deterministic model by means of the indicator $\text{RMSE}_{\text{MME}}/\text{RMSE}_{\text{BEST}}$. 
Figure 4. Model error dependence through the eigenvalues spectrum. The average explained variation from the maximum eigenvalue is $71/78$ (phase I/II) for $O_3$, $65/69$ for $NO_2$ and $74/79$ for $PM_{10}$. On the same graph, the cumulative density function of $N_{EFF}$ calculated from all possible ensemble combinations is presented with the black line.
Figure 5. Interpretation of Figure 4: the explanation of the mme skill against the best local deterministic model with respect to skill difference (evaluated from $\text{MSE}_{\text{BEST}}/\langle\text{MSE}\rangle$) and error dependence (evaluated from the explained variation by the highest eigenvalue).
Figure 6. Like Figure 5 but showing the $N_{\text{eff}}$ with respect to skill difference and error dependence.
Figure 7. Composite skill of all deterministic models and ensemble estimators ($mme$, $mme<$, $kzFO$, $kzHO$, $mmW$) through Taylor plots. The point R represents the reference point (i.e. observations).
Figure 8. Like Figure 5 but for the mme< skill in the reduced ensemble. Please note the change in the colorscale.
Figure 9. The cumulative density function of the Skill Score \((1 - \text{MSE}_X / \text{MSE}_{\text{MME}})\), where \(X = \text{mmW, mme<, kzFO, kzHO}\) over \(\text{mme}\), evaluated at each monitoring site for the examined species of the two AQMEII phases.
Figure 10. The interquartile range over all stations of the day-to-day difference in the weights arising from variable time-series length.