Improving the deterministic skill of air quality ensembles

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

a. University of Patras, Physics Department, University Campus 26504 Rio, Greece.
b. Institute for Environment and Sustainability, Joint Research Centre, European Commission, Ispra, 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.
k. Netherlands Organization for Applied Scientific Research (TNO), Utrecht, The Netherlands.
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

Forecasts from chemical weather models are subject to uncertainties in the input data (e.g. emission inventory, initial and boundary conditions) as well as the model itself (e.g. physical parameterization, chemical mechanism). Multi-model ensemble forecasts can improve the forecast skill provided that certain mathematical conditions are fulfilled. We demonstrate through an intercomparison of two dissimilar air quality ensembles that unconditional raw forecast averaging, although generally successful, is far from optimum. One way to achieve an optimum ensemble is also presented. The basic idea is to either add optimum weights to members or constrain the ensemble to those members that meet certain conditions in time or frequency domain. The methods are evaluated against ground level observations collected from the EMEP and Airbase databases.

The two ensembles were created for the first and second phase of the Air Quality Model Evaluation International Initiative (AQMEII). 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 ensemble mean achieves higher skill compared to each station's best deterministic model at 39%-63% of the sites. The skill gained from the favourable ensemble averaging has at least double the forecast skill compared to using the full ensemble. The method proved robust for the 3-monthly examined time-series if the training phase comprises 60 days. 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 unceasing improvement in the understanding of the physical, chemical and dynamical processes result in better forecasts (Zhang et al., 2012). 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. The benefits from ensemble forecasting arise from the averaging out of the unpredictable components (Kalnay, 2003).

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. Recent 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 enlarged scientific understanding, more accurate and denser observations as well as ensemble forecasting, an extra gain of similar magnitude can be achieved for ensemble-based deterministic forecasting 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 rest, where the conditions are not accomplished, local verification highlights one or another atmospheric model but none particularly. 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 ensemble forecasting is the processing of the deterministic models datasets prior to averaging in order to construct another dataset where 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), the dynamic linear regression (Pagowski et al., 2006; Djalalova et al., 2010), the Kalman filtering (Delle

Monache et al., 2011), the Bayesian model averaging (Riccio et al., 2007) and the analytical optimization (Potempski and Galmarini, 2009) while model selection usually relies on the quadratic error or its proxies (e.g. Solazzo et al., 2013; Kioutsioukis and Galmarini., 2014). In this work, we apply both approaches in an inter-comparison study of two air quality ensemble systems (hereafter, Phase I and Phase II), generated within the Air Quality Model Evaluation International Initiative (AQMEII). The differences between the ensembles of Phase I and Phase II originate from many sources, related to both the input data and the models: (a) the year is different (2006 vs. 2010), therefore the meteorological conditions are different; (b) emission methodologies have changed (see Table 3 in Pouliot et al. 2015); (c) boundary conditions are very different (obtained from GEMS in Phase I, MACC in Phase II); (d) the composition of the ensembles is different; (e) the models in Phase II use on-line coupling between meteorology and chemistry; (f) the models may have been updated with new science processes apart from feedback processes. 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_{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_{EFF}; for this reason, we focus on the European domain due to its sufficient number of models to form the ensemble. The uncertainties arising from observational errors are not taken into consideration.

The objectives of the paper are (a) to interpret the skill of the unconditional multi-model mean within the phase I and II of AQMEII, (b) to calculate the maximum expectations in the skill of alternative ensemble estimators and (c) to evaluate the operational implementation of the approach using cross-validation. 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, a comparison of the skill of the deterministic models and the unconditional ensemble mean across phase I and phase II is performed. In Section 4, the skill of the alternative ensemble estimators is demonstrated. Conclusions are given 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 $\bar{f}=\sum_{i=1}^{M} w_i f_i, \sum w_i = 1$. The weights ($w_i$) sum up to one and can be either equal (uniform ensemble) or unequal (nonuniform ensemble). The desired value (measurement) is μ.

Assuming a uniform ensemble, the squared error (MSE) of the multi-model ensemble mean can be broken down into three components, namely, bias, error variance and error covariance (Ueda and Nakano, 1996):

$$\boldsymbol{MSE(\bar{f}) = \overline{bias}^2 + \frac{1}{M}\overline{var} + \left(1-\frac{1}{M}\right)\overline{cov}} \qquad \textbf{Eq.1}$$

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 factors, 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 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):

$$\boldsymbol{MSE(\bar{f})} = E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i-\mu)^2\right) - E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i-\bar{f})^2\right) \qquad \textbf{Eq.2}$$

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 ideal ensemble error depends on the right trade-off between accuracy (1st term on the r.h.s. of Eq. 2) and diversity (2nd term on the r.h.s. of Eq. 2).

The two 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):

$$\overline{\boldsymbol{w}} = \frac{\boldsymbol{K}^{-1}\boldsymbol{l}}{(\boldsymbol{K}^{-1}\boldsymbol{l},\boldsymbol{l})} \quad \textbf{Eq.3}$$

where, $\boldsymbol{w}$ is the vector of optimal weights, $\boldsymbol{K}$ is the error covariance matrix and $\boldsymbol{l}$ 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.

It appears that the skill of the unconditional ensemble mean (*mme*) has the potential for certain advantages over the single members, provided some properties are satisfied. As those properties are not systematically met in practice, better ensemble skill can be achieved through sub-selecting schemes such as the ideal trade-off between accuracy and diversity (*mme*<) or the optimal weighting (*mmW*). Another sub-selecting scheme is also considered that is derived from ensemble optimization at selected spectral bands with the Kolmogorov-Zurbenko (*kz*) filter (Zurbenko, 1986) and combining them either linearly (*kzFO*) or non-linearly (*kzHO*) (Galmarini et al., 2013). An inter-comparison of all those approaches in ensemble averaging is explored in this work using observed and simulated air quality time-series.

2.1 Reducing dimensionality

The combination of redundant models (i.e., models with highly correlated errors) results in loss of valuable information due to the dependent biases (Solazzo et al., 2013). To improve the accuracy of the ensemble, redundant information in the sub-selecting schemes is discarded by mean of the effective number of models (N_{EFF}) sufficient to reproduce the variability of the full ensemble. N_{EFF} is calculated as (Bretherton et al., 1999):

$$N_{EFF} = \frac{\left(\sum_{i=1}^{M} \boldsymbol{s_i}\right)^2}{\sum_{i=1}^{M} \boldsymbol{s_i^2}} \quad \textbf{Eq.4}$$

where s_i is eigenvalue of the error covariance matrix. 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.

2.2 Verification metrics

The skill of the forecasts have been measured with the following statistical parameters: (1) normalised mean square error (NMSE), i.e. the mean square error (MSE) divided by $\bar{O}\bar{M}$, where $\bar{O}$ and $\bar{M}$ are the mean value of the observation and the model respectively, (2) hit rate (HR), i.e. the proportion of occurrences (e.g. events exceeding threshold value) that were correctly identified, (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 Results

In this section we apply the conceptual context briefly presented in section 2 to investigate the differences and commonalities of the ensembles across the two AQMEII phases (Rao et al., 2011). As mentioned in the introduction, the two ensembles are dissimilar with respect to their input data (emissions, boundary conditions) and their participating coupled models (off-line/on-line) apart from the different meteorology/photochemistry due to the different simulation year. The model settings and input data for phase I are described in Solazzo et al. (2012a, b), Schere et al. (2012), Pouliot et al. (2012); for phase II, similar information is presented in Im et al. (2015a, b), Brunner et al. (2015), Baro et al. (2015), Pouliot et al. (2015). In both cases, the modelling communities simulated annual air quality over Europe and North America for the years 2006 (I) and 2010 (II). From the provided station-based hourly time-series, we analysed the three-monthly period with relatively high concentrations; for O_3, June-July-August was selected while September-October-November is used for NO_2 and PM_{10}. All monitoring stations are rural and have data at least 75% of the time.

We start the analysis with a presentation of the ensemble properties in the two phases, originating from variations in the components (observations, models and their interactions). Only the unconditional full ensemble average (i.e. *mme*) is assessed in this section.

3.1 Observations

The observation networks across the two phases of AQMEII have similar characteristics per species like the number of stations and the fraction of missing data (Table 1). The network is denser for O_3 for which there are as many monitoring stations as for NO_2 and PM_{10} 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. All three pollutants demonstrate a decrease from 2006 to 2010, in line with the emissions reductions, as already documented (European Environmental Agency, 2013). However, we should mention that the decline is unrealistically larger for PM_{10} due to the different spatial coverage of the sampling stations. Unlike the other pollutants, no valid data for France and UK were available in phase II for PM_{10} (station locations are shown in Figure 4).

3.2 Models

The number of ensemble members available from Phase I ranges from 10 (PM_{10}) to 12 (O_3) and 13 (NO_2) while in Phase II 14 members were available for all species (Table 1). 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.

The boxplots of NMSE over all monitoring stations is presented in Figure 2. The aggregated mean skill of the individual models across the two phases appears similar for O_3, shows an improvement for NO_2 (median <NMSE> shifted from 0.53 to 0.49) and a worsening for PM_{10} (median <NMSE> shifted from 0.47 to 0.50) (Table 2). At the same time, the best model at each monitoring station has similar behaviour for O_3 and NO_2 across the two phases and experiences degradation for PM_{10} (median <NMSE> shifted from 0.34 to 0.37). In summary, (a) many models improved their skill for NO_2 in the Phase II simulations although no improvement occurred in the prediction capacity of the best model, (b) the model skill was generally deteriorated for PM_{10} in Phase II, shifting the NMSE distribution towards higher values, (c) no notable changes were seen for O_3. The indirect feedback mechanisms available in phase II generally improved the simulation of meteorological drivers such as temperature, radiation and precipitation, which in turn improved the forecast of many atmospheric gases while particulate matter and cloud processes require updated parameterizations (Brunner et al. (2015), Makar et al. (2015)).

3.3 Multi-model mean

As shown above, the differences between Phase 1 and Phase 2 in terms of individual accuracy of the models varied between the three examined species. We examine now the consequences in the behaviour of the multi-model mean and interpret the results with respect to the presented error decompositions. As suggested from equations 1 and 2, the error of the multi-model mean relies on the skill difference of its members and their error dependence.

Skill difference

Despite the different changes in individual model skill for the different species, when they are combined to form an ensemble, the skill difference between the best model and the average skill has decreased for all species from phase I to II. This is inferred from the values of the indicator $NMSE_{BEST}$ /<NMSE> that increase (Table 2). This increase occurs because of more good models in phase II. To explain this, we evaluate the percentage of cases each model has been identified as being 'best' and record the number of models exceeding specific percentage thresholds. If models were behaving like *i.i.d.*, the probabilities of being best would be roughly equal (~1/M) for all models. As can be inferred from Table 2, the proportion of *equally good models* has increased in phase II for O_3 and NO_2, since the number of models exceeding the 1/M percentage contains half of the models compared to one third in phase I. This is not however true for the Phase II PM_{10} simulations, where one model outscores the others at roughly 40% (~6/M) of the stations, implying a missing process in the majority of the models. It turned out that this model was erroneously running with off-line coupling between meteorology and chemistry.

Error dependence

The combination of models with correlated errors brings redundant information in the ensemble and reduces the benefits of ensemble averaging. The eigenvalues of the covariance matrix calculated from the model errors provides information for the members' diversity and the ensemble redundancy. 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 3). The number of necessary eigenvalues to capture 86% of the variation is referred as effective number of models (N_{EFF}). In phase I, the maximum value of N_{EFF} across *all stations* is 6 for O_3 and NO_2 and 4 for PM_{10}. In phase II, this number is approximately 5 for all species. Hence, 5±1 models are sufficient for all species at both phases. Therefore, from a pool of 10-14 models, the benefits of ensemble averaging cease after 6 members (but not 6 particular members). Further, the average explained variation by the maximum eigenvalue (s_m) has increased for all species in phase II, indicating a decrease in ensemble diversity.

Similar values across the two phases for the effective number of models are found from an estimation based on the optimal trade-off between accuracy and diversity, shown in the same figure. Rather than using a benchmark for the error dependence (i.e., the error covariance matrix), the N_{EFF} is estimated from the error minimization across all possible combinations of M models at each site. At 50% of the stations, the optimum number of ensemble members is less or equal to 3 while at 95% of the stations the maximum optimum number of models becomes 6. In other words, we do need more than 6 members at most stations. The only exception is the NO_2 (II) case, where N_{EFF} across the two phases defer by 1 (higher in phase II). As we will see later, this is due to the fact that only for NO_2 (II), there is imbalance in the relative changes of skill difference and error dependence.

Multi-model mean skill

The phase II ensemble consists of models with, compared to phase I, generally improved skill for NO_2, worse skill for PM_{10} and similar skill for O_3. The phase II ensemble as a whole demonstrates smaller skill differences between models for all species. Last, increased error dependence is evidenced in phase II, arising primarily from the fact that 50% of the ensemble members run the same model with differences arising only from the choice of different physical or chemical parameterizations. The modulation of the ensemble mean skill owing to the changes in its properties across the two phases is now examined.

The skill of the multi-model mean has been compared against the skill of the best available deterministic model, independently evaluated at each monitoring site. The geographical distribution of the ratio RMSE(*mme*)/$RMSE_{BESTMODEL}$ is presented in Figure 4. The indicator does not exhibit any longitudinal or latitudinal dependence. We also observe that the number of extreme cases where the *mme* skill was notably inferior to the best model has dropped from phase I to II. Specifically, the percentage of stations where the RMSE(*mme*) was 10-30% higher than the $RMSE_{BESTMODEL}$ dropped from 17.2% to 9.3% for O_3 and from 10.0% to 5.6% for NO_2. As presented in more detail in Table 3 for the statistical distribution of the indicator:

- no major differences exist for O_3, with the *mme* outscoring the best model at half of the stations. Extreme values of the indicator at both tails are trimmed in phase II;

- a clear improvement is evident for NO_2, with the *mme* providing more skilled forecasts at 63% of the sites, compared to 38% in the previous phase. All ranges exhibit improvement, indicating a distribution shift;
- a mild improvement is also evident for PM_{10}, where the number of stations where *mme* performs better increased from 38% to 42%. Extreme values of the indicator at both tails are increased in phase II.

The reason behind the behaviour of *mme* is given in Figure 5 and emerges from the joint distribution of skill difference and error dependence. Skill difference decreased for all species and error dependence increased for all species, from phase I to II. It is their relative change that modulates *mme* skill. For O_3, both are altered by a comparable amount, resulting in similar *mme* skill across phase I and II. For NO_2, skill difference was improved more than error dependence was worsened, yielding a net improvement of *mme*. For PM_{10}, the situation is similar to NO_2 though with a milder relative difference.

The area below the diagonal in Figure 5 corresponds to monitoring sites with disproportionally low diversity under the current level of accuracy. Seen from another angle, 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_{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_{EFF} value. This is the desired synthesis for an ensemble. In the next section we will examine some approaches that are able to put all points in the area above the diagonal. Figure 7 demonstrates such a case with an ensemble build with selected members (*mme*<).

4 Ensemble improvements

Following the identification of the weaknesses in the ensemble design, the potential for corrections through more sophisticated schemes is now investigated. Given the observations, optimal weights or members can be estimated or selected. In this section we mark the boundaries of the possible improvements for different ensemble mean estimators applicable to the AQMEII datasets and in the next subsection we investigate the actual forecast skill for sub-optimal conditions using cross-validation.

The average error across all the monitoring stations was lower for *mme* compared to the single models in both phases. The spatio-temporal robustness of *mme* skill has increased in phase II, for different reasons per species as analysed in the previous section. 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 1st and higher order components by different models (*kzFO*, *kzHO*) (Galmarini et al., 2013).

The results evaluated at all stations are presented in Figure 8 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 phase 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 phase 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 improvements versus *mme*, at each station separately, is now conducted. The cumulative density function of the indicator MSE_X/MSE_{MME} (X = mmW, mme<, kzFO, kzHO) evaluated at each monitoring is shown in Figure 9. For O_3, the median improvement was 27% for *mmW*, 22-25% for *kzHO* and 17% for *kzFO* and *mme*<, relatively equal across the two phases. At ten percent of the stations, the improvement can be over 41%. For NO_2, the median improvement for phase I (phase II) was 21% (17%) for *mmW*, 20% (13%) for *kzHO* and 13% (7-9%) for *kzFO* and *mme*<. The magnitude of improvement can exceed 39% (30%) at roughly ten percent of the stations. Unlike NO_2, PM_{10} shows higher improvement rates for phase II simulations; the median improvement for was 13-24% for *mmW*, 11-19% for *kzHO*, 8-16% for *mme*< and 8-12% for *kzFO*. The magnitude of improvement surpasses 22% (37% in phase II) at ten percent of the stations.

The statistical distributions of all MSE_X/MSE_{MME} indicators (X = *mmW, mme<, kzFO, kzHO*) are well bounded from above to lower than unity values. 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 demonstrate high N_{EFF}, where the benefits from unconditional ensemble averaging are optimal (Kioutsioukis and Galmarini, 2014).

The ability to forecast extreme values is now examined through the hit rate indicator (probability of detecting events exceeding a certain threshold). Due to the lowering of the concentrations from phase I to II, a percentile threshold is more appropriate for the comparison rather than a fixed threshold. Therefore, a threshold reflecting the average 90th percentile across the stations has been selected, being 129/117 $\mu g/m^3$ (phaseI/II) for O_3, 30/26 $\mu g/m^3$ for NO_2 and 52/33 $\mu g/m^3$ for PM_{10}. The ability of the models at the tail simulation was similar to the <NMSE> change from phase I to II. For O_3, the percentage of successful events exceeding the 90th percentile for *mme* was 29% (25%) for phase I (II). The major improvement occurred for *mmW*, where the aggregated hit rate was 51% (48%), and the smaller improvement was for *mme<*, with value 42% (38%). The spectral estimators yielded values of 47% (42%) and 46% (40%) for *kzFO* and *kzHO* respectively. For NO_2, the successful hits for *mme* was 35% (42%) and reached 45% (49%) for *mmW*. For the other ensemble averages, the result was 39% (45%) for *mme<*, 39% (44%) for *kzFO* and 40% (47%) for *kzHO*. For PM_{10}, the total percentage of successful hits for *mme* was 19% (16%) and became 33% (42%) for *mmW,* while the other estimators yielded 28% (27%), 29% (30%) and 31% (28%) for *mme<*, *kzFO* and *kzHO* respectively.

The range of forecast error, from the worst deterministic model to the optimum ensemble-based average is presented in Table 4. Statistics were calculated for the 3-monthly evaluation period and averaged over all monitoring sites. All values have been normalized with the error of the best deterministic model in order to quantify the potential extent of improvement that each method can achieve as a function of species and feedbacks. We observe that the benefits from ensemble averaging in the form of *mme* range from 1% to 12% when compared to the best numerical model. Under proper weighting, this distance is, at a minimum, doubled. The range of improvement for *mmW* over the best single model was from 9% to 27%.

To summarize:

- [Error] The analytical optimization of the error through non-uniform weighting (*mmW*) achieved lower MSE compared to the sub-selecting schemes. Among species, improvements over *mme* are larger for O_3 and smaller for PM_{10}, i.e. proportional to the skill of the deterministic models.
- [Extremes] The ranking of the methods with respect to their capability for extremes was inline with the skill of the methods for the mean error. The ability of all models to capture levels exceeding a fixed threshold was better for O_3 and PM_{10} in phase I and for NO_2 in phase II. Among species, *mme* performed best for NO_2 and worst for PM_{10}. The total percentage of successfully modelled extreme values from using the statistical treatments increased by up to 10% for NO_2, 23% for O_3 and 26% for PM_{10}.

4.1 Forecasting performance

The statistical treatments applied to a pool of ensemble simulations generated results with improved skill in diagnostic mode. To provide a perspective on applying these techniques in a forecasting context, we explore the temporal robustness of the weighting scheme, i.e. their predictability window. For this reason, the weights have been re-calculated for variable time-series length that is progressively increasing from 1 to 60 days, for all monitoring stations across the two phases. The evaluation period for all training windows is the same 30-day segment, not available in the training procedure. 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 20 days of hourly time-series for O_3 and NO_2 and 30 days for PM_{10} (phase I). This period can be thought of as the necessary training or learning period. In phase II, those periods are increased and they become 25 days for O_3, 45 days for NO_2 and PM_{10}. Weights are unpredictable for smaller periods. In practice, even safer margins should be employed. Using half of the tolerance applied, we need an approximate learning period of 50 days for phase I and 60 days for phase II. Last, the sub-selecting schemes, unlike the analytical optimization, are quite robust even for very small training periods (e.g. 1 week), whether in the form of *mme<* (Kioutsioukis and Galmarini, 2014) or *kzFO*/*kzHO* (Galmarini et al., 2013).

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 traning period is less than 20 days for NO_2 and 30 days for PM_{10}. For all pollutants, the variability of the weights has no effect in the error after 60 days.

5 Conclusions

In this paper we give an overview of the performance of the forecast systems in the two phases of AQMEII and their effect in the skill of the ensemble mean. The results are interpreted with respect to the error decomposition of the ensemble. Ways to extract more information from an ensemble besides the ensemble mean are ultimately investigated and evaluated.

Air Quality models simulate the atmospheric composition through a series of complex physical, chemical and dynamical processes. In the hypothetical scenario where a simulation experiment with an ensemble of chemical weather models is performed twice, with the only difference being off-line or on-line coupling among meteorological and chemical modules, the increased non-linearity in the latter case is expected to enhance the model independence and hence generate more diverse results between models. Assuming the accuracy of the models remains the same, the increased diversity in the latter case favours the skill of the multi-model mean in the simulation with feedbacks compared to models without interactions. However, maintaining the same level of accuracy when we incorporate feedbacks in the models is not granted. Besides feedbacks, the varying factors between the two AQMEII experiments included also different models, emissions, boundary conditions and simulation year.

The indirect contrast assessed demonstrated that the ensembles of phase I and phase II have several key differences. The average accuracy in phase II has improved for NO_2, decreased for PM_{10} and remained the same for O_3. At the same time, the accuracy of the best model remained the same for NO_2 and O_3 and decreased for PM_{10}. In other words, without pushing further the predictability limits, many models simulate better NO_2 in phase II. The opposite is true for PM_{10}, where phase II modelling accuracy was deteriorated. In terms of redundancy, despite the expected increase in variability, the ensemble diversity was reduced in phase II, mainly due to the fact that half of the ensemble members were originating from the same model using only different physical or chemical parameterizations. The combined effect for the multi-model mean, in terms of the NMSE was neutral, regardless of the idealized theoretical expectations. However, the relative changes in the accuracy and diversity in phase II, favoured always the multi-model mean over the best local deterministic model, enhancing further its spatiotemporal robustness. This raises the topic of ensemble design and supports again the critical importance of having the right amount of accuracy and diversity within an ensemble.

Several improvements in the multi-model mean skill were also examined in the form of weighting or sub-selecting. The skill enhancement was superior using the weighting scheme but the required training phase to acquire representative weights was higher compared to the sub-selecting schemes. For all pollutants, the variability of the weights has negligible effect in the error for training periods longer than 60 days. The range of improvement for the optimal multi-model mean over the best single model was from 9% (PM_{10}) to 27% (O_3), when the corresponding range for the traditional unconditional multi-model average was from 1% to 12%. The advancement from the other approaches that use reduced-size ensembles closely follows the skill of the optimal scheme. The presented post-simulation advancements were the result of only favourable ensemble design. The combined skill earned from conditional versus unconditional ensemble averaging is comparable with the one obtained each decade as a result of the aggregated advancements in numerical prediction due to more and better assimilated observations, higher computing power and progress in our understanding of dynamics and physics.

The improvement of the physical, chemical and dynamical processes in the deterministic models is a ceaseless 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. Further development is underway in the presented ensemble methods that take into account the meteorological and chemical regimes.

References

Air pollution fact sheet 2013, European Environmental Agency, 20 pp, 2013.

Baró, R., P. Jiménez-Guerrero, A. Balzarini , G. Curci , R. Forkel, M. Hirtl , L. Honzak, U. Im, C. Lorenz, J.L. Pérez, G. Pirovano, R. San José; P. Tuccella, J. Werhahn, R. Žabkar: Sensitivity analysis of the microphysics scheme in WRF-Chem contributions to AQMEII phase 2, Atmospheric Environment 715: 620-629, 2015.

Bishop, C.H., Abramowitz, G.: Climate model dependence and the replicate earth paradigm. Clim Dyn 41(3–4): 885–900, 2013.

Bretherton, C.S., Widmann, M., Dymnikov, V.P., Wallace, J.M., Bladè, I.: The effective number of spatial degrees of freedom of a time-varying field. J. Climate 12(7): 1990-2009, 1999.

Brunner, D., Jorba, O., Savage, N., Eder, B., Makar, P., Giordano, L., Badia, A., Balzarini, A., Baro, R., Bianconi, R., Chemel, C., Forkel, R., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Im, U., Knote, C., Kuenen, J.J.P., Makar, P.A., Manders-Groot, A., Neal, L., Perez, J.L., Pirovano, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R., van Meijgaard, E., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S.: Comparative analysis of meteorological performance of coupled chemistry-meteorology models in the context of AQMEII phase 2. Atmospheric Environment 115: 470-498, 2015.

Delle Monache, L., T. Nipen, Y. Liu, G. Roux, Stull, R.: Kalman filter and analog schemes to postprocess numerical weather predictions. Month. Wea. Rev. 139: 3554-3570, 2011.

Djalalova, I, J Wilczak, S McKeen, G Grell, S Peckham, M Pagowski, L DelleMonache, J McQueen, Y Tang, P Lee, J McHenry, W Gong, V Bouchet, Mathur, R.: Ensemble and bias-correction techniques for air quality model forecasts of surface O-3 and PM2.5 during the TEXAQS-II experiment of 2006. Atmos. Environ. 44 (4): 455-467, 2010.

Eskes, H., van Velthoven, F., Kedler H.: Global ozone forecasting based on ERS-2 GOME observations. Atmos. Chem. Phys. 2: 271-278, 2002.

Galmarini, S., Kioutsioukis, I., and Solazzo, E.: E pluribus unum*: ensemble air quality predictions, Atmos. Chem. Phys. 13: 7153–7182, 2013.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Flemming, J., Forkel, R., Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Kuenen, J.J.P., Makar, P.A., Manders-Groot, A., Neal, L., Perez, J.L., Piravano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S.: Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part I: Ozone. Atmospheric Environment 115: 404-420, 2015a.

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baro, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., Denier van der Gon, H.A.C., Flemming, J., Forkel, R., Giordano, L., Jimenez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Makar, P.A., Manders-Groot, A., Neal, L., Perez, J.L., Piravano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Werhahn, K., Wolke, R., Yahya, K., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C., Galmarini, S.: Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: Particulate Matter. Atmospheric Environment 115: 421-441, 2015b.

Kalnay E.: Atmospheric modelling, data assimilation and predictability, Cambridge University Press, 341 pp., 2003.

Kioutsioukis, I. and Galmarini S.: De praeceptis ferendis: good practice in multi-model ensembles, Atmospheric Chemistry and Physics 14: 11791-11815, 2014.

Krogh A. and J. Vedelsby: Neural network ensembles, cross validation, and active learning, In Advances in Neural Information Processing Systems, pp.231-238, 1995.

Makar, P.A., Gong, W., Hogrefe, C., Zhang, Y., Curci, G., Zabkar, R., Milbrandt, J., Im, U., Balzarini, A., Baro, R., Bianconi, R., Cheung, P., Forkel, R., Gravel, S., Hirtl, M., Honzak, L., Hou, A., Jimenez-Guerero, P., Langer, M., Moran, M.D., Pabla, B., Perez, J.L., Pirovano, G., San Jose, R., Tucella, P., Werhahn, J., Zhang, J., Galmarini, S.: Feedbacks between air pollution and weather, part 2: Effects on chemistry. Atmospheric Environment 115: 499-526, 2015.

Mallet, V., Stoltz, G., Mauricette, B.: Ozone ensemble forecast with machine learning algorithms. J. Geophys. Res. 114, D05307. doi:10.1029/2008JD009978, 2009.

Marécal V., V.-H. Peuch, C. Andersson, S. Andersson, J. Arteta, M. Beekmann, A. Benedictow, R. Bergström, B. Bessagnet, A. Cansado, F. Chéroux, A. Colette, A. Coman, R. L. Curier, H. A. C. Denier van der Gon, A. Drouin, H. Elbern, E. Emili, R. J. Engelen, H. J. Eskes, G. Foret, E. Friese, M. Gauss, C. Giannaros, J. Guth, M. Joly, E. Jaumouillé, B. Josse, N. Kadygrov, J. W. Kaiser, K. Krajsek, J. Kuenen, U. Kumar, N. Liora, E. Lopez, L. Malherbe, I. Martinez, D. Melas, F. Meleux, L. Menut1, P. Moinat, T. Morales, J. Parmentier, A. Piacentini, M. Plu, A. Poupkou, S. Queguiner, L. Robertson, L. Rouïl, M. Schaap, A. Segers, M. Sofiev, L. Tarasson, M. Thomas, R. Timmermans, Á. Valdebenito, P. van Velthoven, R. van Versendaal, J. Vira, A. Ung: A regional air quality forecasting system over Europe: the MACC-II daily ensemble production Geosci. Model Dev., 8, 2777-2813, 2015.

Pagowski, M., G.A. Grell, S.A. McKeen, D. Devenyi, J.M. Wilczak, V. Bouchet, W. Gong, J. McHenry, S. Peckham, J. McQueen, R. Moffet, Y. Tang: A simple method to improve ensembel-based ozone forecasts, Geophy. Res. Lett., 32, L07814, doi:10.1029/2004GL022305, 2005.

Pagowski, M., G.A. Grell, D. Devenyi, S. Peckham, S.A. McKeen, W. Gong, L. Delle Monache, J.N. McHenry, J. McQueen, P. Lee: Application of Dynamic Linear Regression to Improve the Skill of Ensemble-Based Deterministic Ozone Forecasts, Atmos. Environ. 40: 3240-3250, 2006.

Potempski, S. and Galmarini, S.: Est modus in rebus: analytical properties of multi-model ensembles, Atmos. Chem. Phys. 9: 9471-9489, 2009.

Pouliot, G., Pierce, T, Denier van der Gon, H., Schaap, M., Nopmongcol, U.: Comparing Emissions Inventories and Model-Ready Emissions Datasets between Europe and North America for the AQMEII Project. Atmospheric Environment 53: 4–14, 2012.

Pouliot, G., Hugo Denier van der Gon, J Kuenen, Junhua Zhang, Michael Moran, Paul Makar: Analysis of the Emission Inventories and Model-Ready Emission Datasets of Europe and North America for Phase 2 of the AQMEII Project, Atmospheric Environment 115: 345-360, 2015.

Rao, S.T., Galmarini, S., Puckett, K.: Air quality model evaluation international initiative (AQMEII): Advancing the state of the science in regional photochemical modeling and its applications. Bulletin of the American Meteorological Society 92(1): 23-30, 2011.

Riccio, A., Giunta, G., Galmarini, S.: Seeking for the rational of the median model: the optimal combination of multi- model ensemble. Atmospheric Chemistry and Physics 7: 6085–6098, 2007.

Simmons, A.: From Observations to service delivery: Challenges and opportunities. WMO Bulletin 60(2): 96-107, 2011.

Solazzo E., A. Riccio, I. Kioutsioukis, S. Galmarini: Pauci ex tanto numero: reduce redundancy in multi-model ensembles, Atmos. Chem. Phys. 13: 8315–8333, 2013.

Taylor, K.E.: Summarizing multiple aspects of model performance in a simple diagram. Journal Geophys. Res. 106, D7, 7183-7192, 2001.

Ueda N., R. Nakano.: Generalization error of ensemble estimators, In Proceedings of International Conference on Neural Networks, pages 90–95, 1996.

Weigel A., R. Knutti, M. Liniger, C. Appenzeller: Risks of model weighting in multimodel climate projections, Journal of Climate 23: 4175-4191, 2010.

Zhang, Y., C. Seigneur, M. Bocquet, V. Mallet, A. Baklanov: Real-Time Air Quality Forecasting, Part II: State of the Science, Current Research Needs, and Future Prospects, Atmos. Environ., 60, 656-676, 2012.

Zurbenko, I. G.: The Spectral Analysis of Time Series, 236 pp., North-Holland, Amsterdam, 1986.

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 ozonesonde 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.

Table 1. The forecasting systems and the evaluation network in Europe in the inter-comparison exercise of the AQMEII phases I and II: simulation models, number of rural stations and data coverage per species.

	O_3	NO_2	PM_{10}
	(I / II)	(I / II)	(I / II)
Models	12 / 14	13 / 14	10 / 14
Stations	451 / 450	290 / 337	126 / 131
Missing Data (%)		Fraction of stations	
0-5	0.67 / 0.76	0.52 / 0.59	0.72 / 0.78
5-10	0.24 / 0.16	0.28 / 0.29	0.13 / 0.14
10-15	0.05 / 0.05	0.09 / 0.07	0.09 / 0.05
15-20	0.02 / 0.02	0.06 / 0.01	0.03 / 0.01
20-25	0.02 / 0.01	0.04 / 0.04	0.02 / 0.01

Table 2. The statistical distribution of the NMSE of the best model ($NMSE_{BEST}$) and the ensemble average NMSE (<NMSE>), evaluated at each monitoring site for the examined species of the two AQMEII phases. In addition, the average value of the ratio ACCN=$NMSE_{BEST}$ /<NMSE> and the number of best models (N_{BEST}) exceeding specific percentage thresholds is also displayed. For example, for PM_{10} (II) there are 4 out of 14 models that scored the least NMSE across at least the 7% of stations (1/M), 2 models (of those 4) which scored the least NMSE across at least the 14% of stations (2/M), etc, pointing that one model outscored the others at over 36% (5/M) of the stations.

	O_3	O_3	NO_2	NO_2	PM_{10}	PM_{10}
	(I/II)	(I/II)	(I/II)	(I/II)	(I/II)	(I/II)
	<NMSE>	$NMSE_{BEST}$	<NMSE>	$NMSE_{BEST}$	<NMSE>	$NMSE_{BEST}$
5th	0.04 / 0.04	0.03 / 0.03	0.28 / 0.23	0.17 / 0.17	0.30 / 0.28	0.20 / 0.20
25th	0.07 / 0.07	0.05 / 0.05	0.39 / 0.35	0.24 / 0.25	0.40 / 0.39	0.26 / 0.28
50th	0.10 / 0.10	0.07 / 0.08	0.53 / 0.49	0.34 / 0.35	0.47 / 0.50	0.34 / 0.37
75th	0.15 / 0.15	0.11 / 0.11	0.82 / 0.76	0.48 / 0.50	0.60 / 0.62	0.46 / 0.50
95th	0.24 / 0.24	0.18 / 0.18	1.69 / 1.49	0.81 / 0.93	1.02 / 0.98	0.73 / 0.81
	O_3	O_3	NO_2	NO_2	PM_{10}	PM_{10}
	(I)	(II)	(I)	(II)	(I)	(II)
ACCN	0.68	0.76	0.60	0.70	0.70	0.77
N_{BEST} (1/M)	4	6	3	7	3	4
N_{BEST} (2/M)	3	1	3	0	1	2
N_{BEST} (3/M)	1	0	2	0	1	1
N_{BEST} (4/M)	0	0	0	0	0	1
N_{BEST} (5/M)	0	0	0	0	0	1
N_{BEST} (6/M)	0	0	0	0	0	0

1 **Table 3. The percentage of stations lying at various bins of the indicator $RMSE_{MME}/RMSE_{BEST}$,**
2 **evaluated at each monitoring site for the examined species of the two AQMEII phases.**

$RMSE_{MME}/RMSE_{BEST}$	O_3	O_3	NO_2	NO_2	PM_{10}	PM_{10}
	(I)	(II)	(I)	(II)	(I)	(II)
0.7 - 0.8	0	0	0	0	0	0
0.8 - 0.9	8.4	2.4	4.1	6.2	0	6.9
0.9 - 1.0	43.7	46.7	34.5	57.3	38.1	35.1
1.0 - 1.1	29.7	41.6	48.6	30.0	61.9	55.0
1.1 - 1.2	13.7	8.2	7.9	4.7	0	3.0
1.2 - 1.3	3.5	1.1	2.1	0.9	0	0.0
<1	*52.1*	*49.1*	*38.6*	*63.5*	*38.1*	*42.0*

3

Table 4. The RMSE from the worst deterministic model to the optimum ensemble average, averaged over all stations. The worst and the best model have been evaluated at each site. The worst (best) deterministic model is the set containing the worst (best) time-series at each station. All values have been normalized with the RMSE of the composite best deterministic model.

Model	O_3	O_3	NO_2	NO_2	PM_{10}	PM_{10}
	(I)	(II)	(I)	(II)	(I)	(II)
Worst deterministic	1.10	1.19	1.43	1.43	1.31	1.16
Average RMSE	1.04	1.07	1.15	1.11	1.09	1.08
Best deterministic	1.00	1.00	1.00	1.00	1.00	1.00
mme	0.88	0.95	0.96	0.95	0.98	0.99
mme<	0.79	0.87	0.90	0.91	0.94	0.93
kzFO	0.79	0.86	0.90	0.92	0.94	0.93
kzHO	0.76	0.84	0.87	0.89	0.93	0.91
mmW	0.73	0.79	0.85	0.87	0.91	0.86

mme: unconditional ensemble mean

mme<: conditional ensemble mean (Kioutsioukis and Galmarini, 2014)

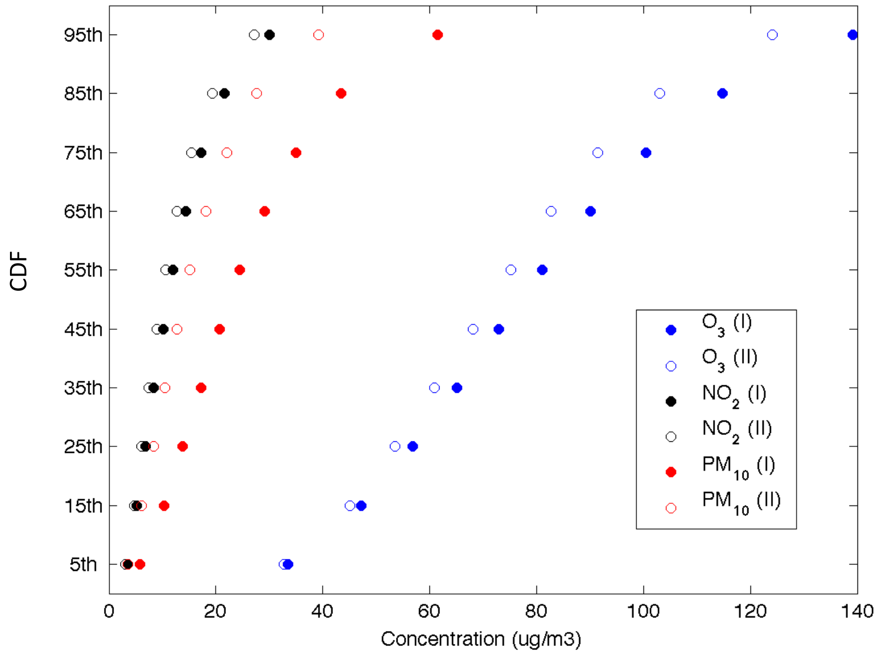
kzFO: conditional spectral ensemble mean with 1st order components (Galmarini et al., 2013)

kzHO: conditional spectral ensemble mean with 2nd and higher order components (*kzHO*)

mmW: optimal weighted ensemble (Potempski and Galmarini, 2009)

Table 5. The RMSE of *mmW* for various training lengths, calculated for the testing time-series (i.e. not-used in the training phase) that contains all stations. All values have been normalized with the RMSE of the composite best deterministic model.

Length of training period (days)	O_3 (I)	O_3 (II)	NO_2 (I)	NO_2 (II)	PM_{10} (I)	PM_{10} (II)
5	0.98	1.04	1.10	1.26	1.55	1.21
10	0.88	0.94	1.01	1.06	1.14	1.05
20	0.79	0.87	0.93	0.96	1.02	0.95
30	0.77	0.83	0.91	0.92	0.96	0.90
60	0.73	0.80	0.85	0.87	0.91	0.86



1 **Figure 1. Comparison of the Cumulative density functions of the observations (O_3, NO_2, PM_{10})**
2 **between the two AQMEII phases (Phase I: *filled circles*, Phase II: *non-filled circles*). Each bullet**
3 **represents the median at the specific percentile.**

4

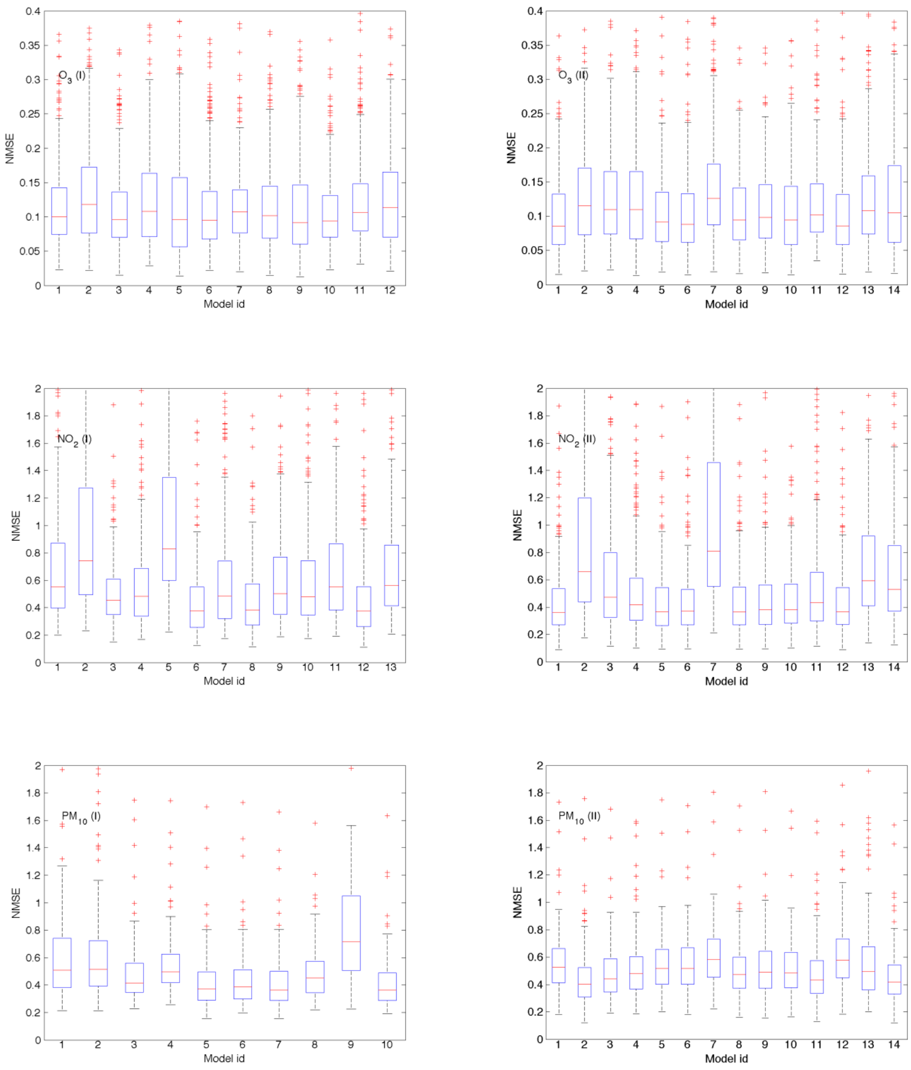


Figure 2. Model skill difference via the NMSE. On each box, the central mark indicates the median, and the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme data points not considered outliers and the outliers (points with distance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

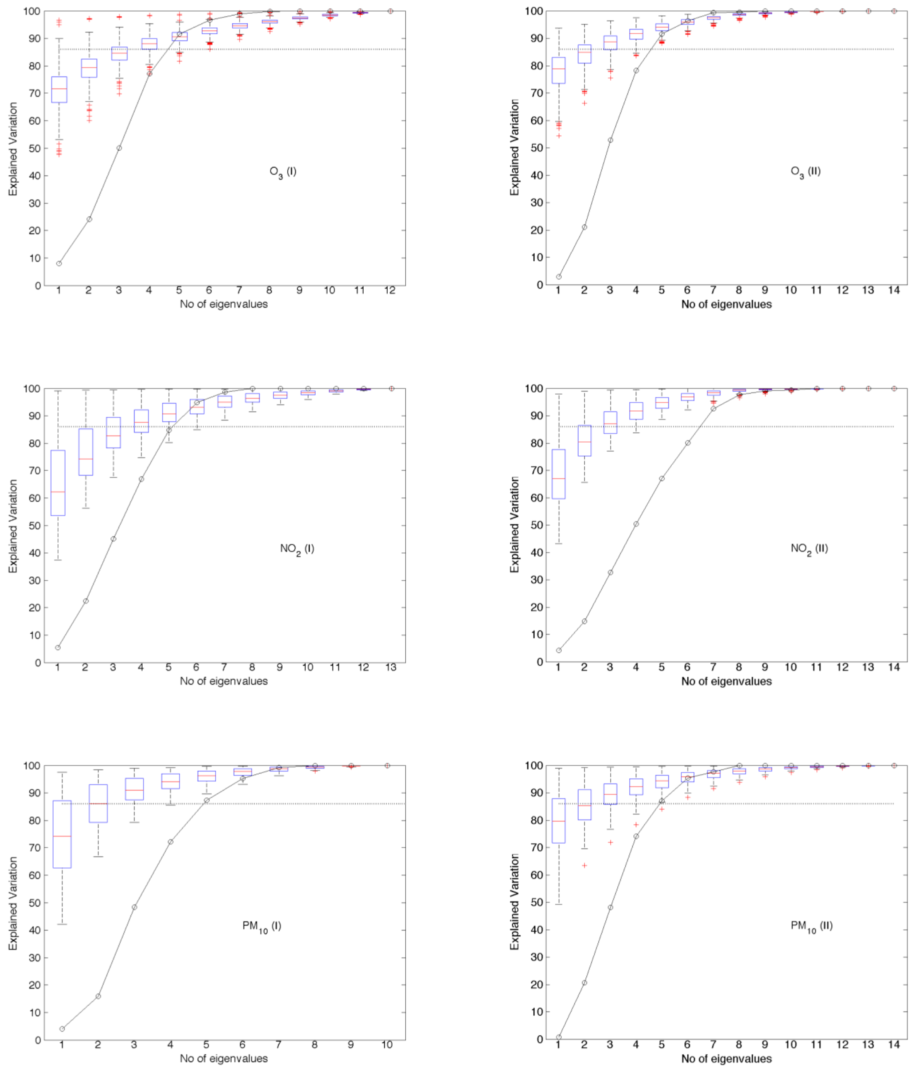
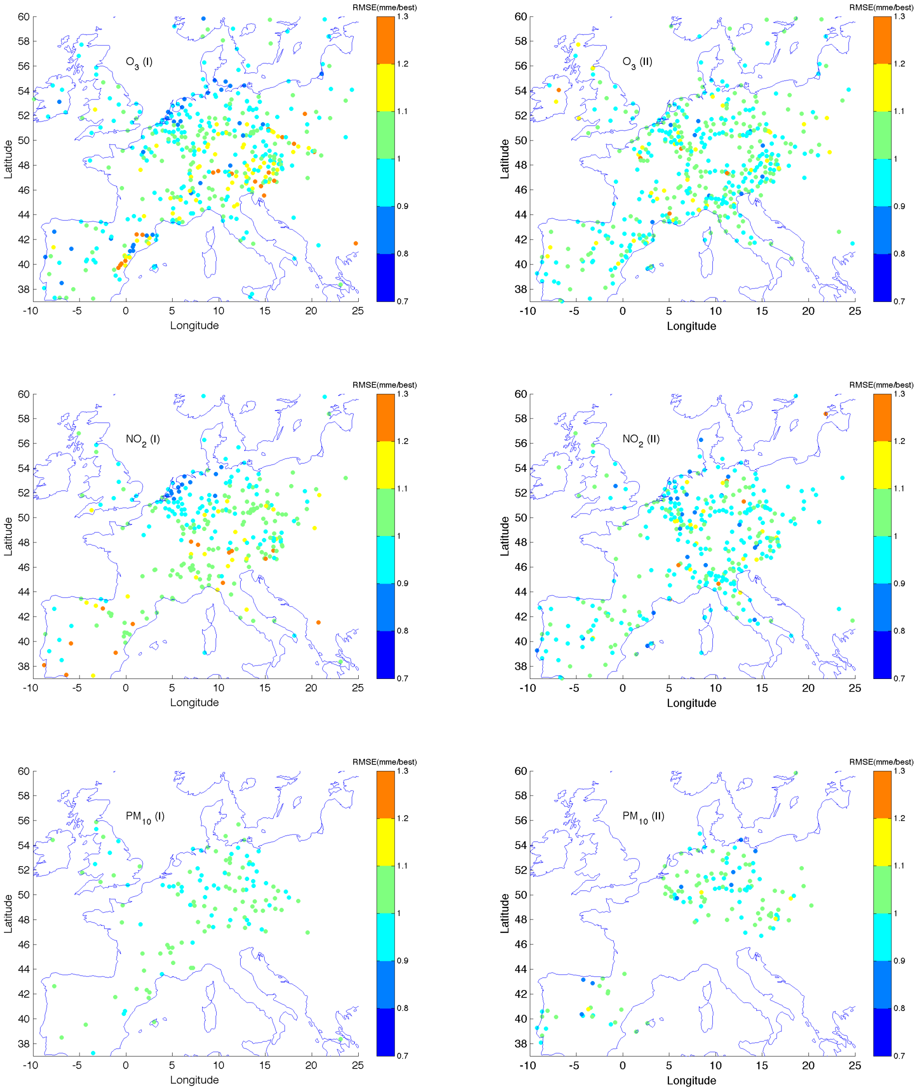


Figure 3. Model error dependence through the eigenvalues spectrum. The average explained variation from the maximum eigenvalue is 71/78 (phase I/II) for O_3, 65/69 for NO_2 and 74/79 for PM_{10}. On the same graph, the cumulative density function of N_{EFF} calculated from all possible ensemble combinations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

Atmos. Chem. Phys. Discuss., doi:10.5194/acp-2016-513, 2016
Manuscript under review for journal Atmos. Chem. Phys.
Published: 30 June 2016



1 **Figure 4. Comparison of the *mme* skill against the best local deterministic model by means of the**
2 **indicator RMSE$_{MME}$/RMSE$_{BEST}$.**

3

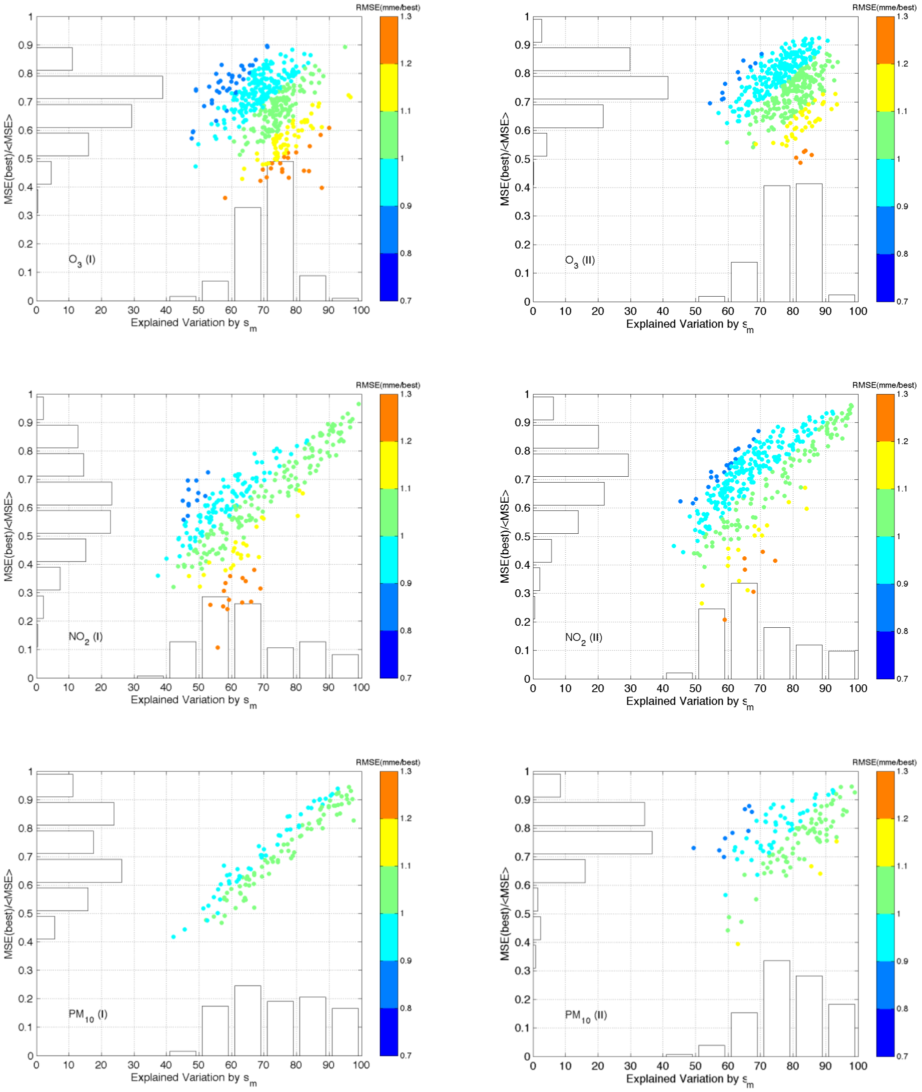


Figure 5. Interpretation of Figure 4: the explanation of the mme skill against the best local deterministic model with respect to skill difference (evaluated from MSE_{BEST}/<MSE>) and error dependence (evaluated from the explained variation by the highest eigenvalue).

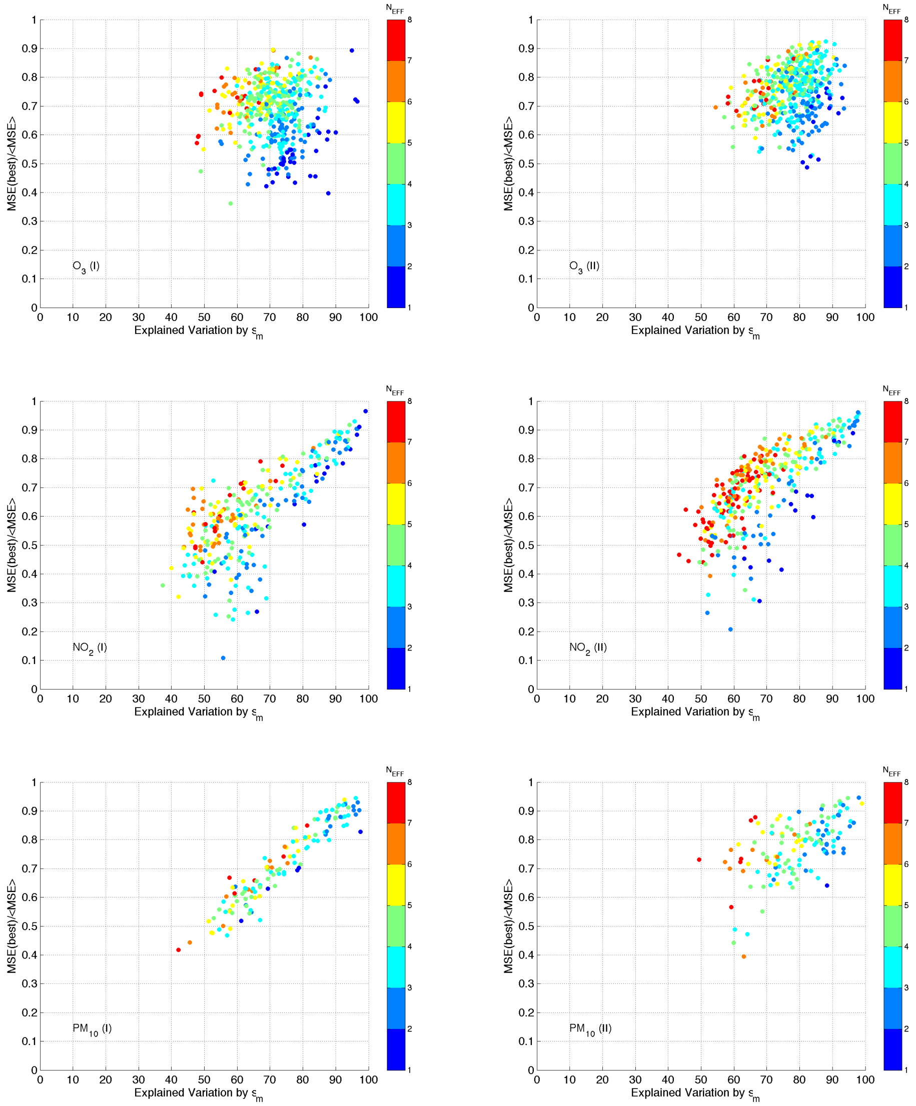
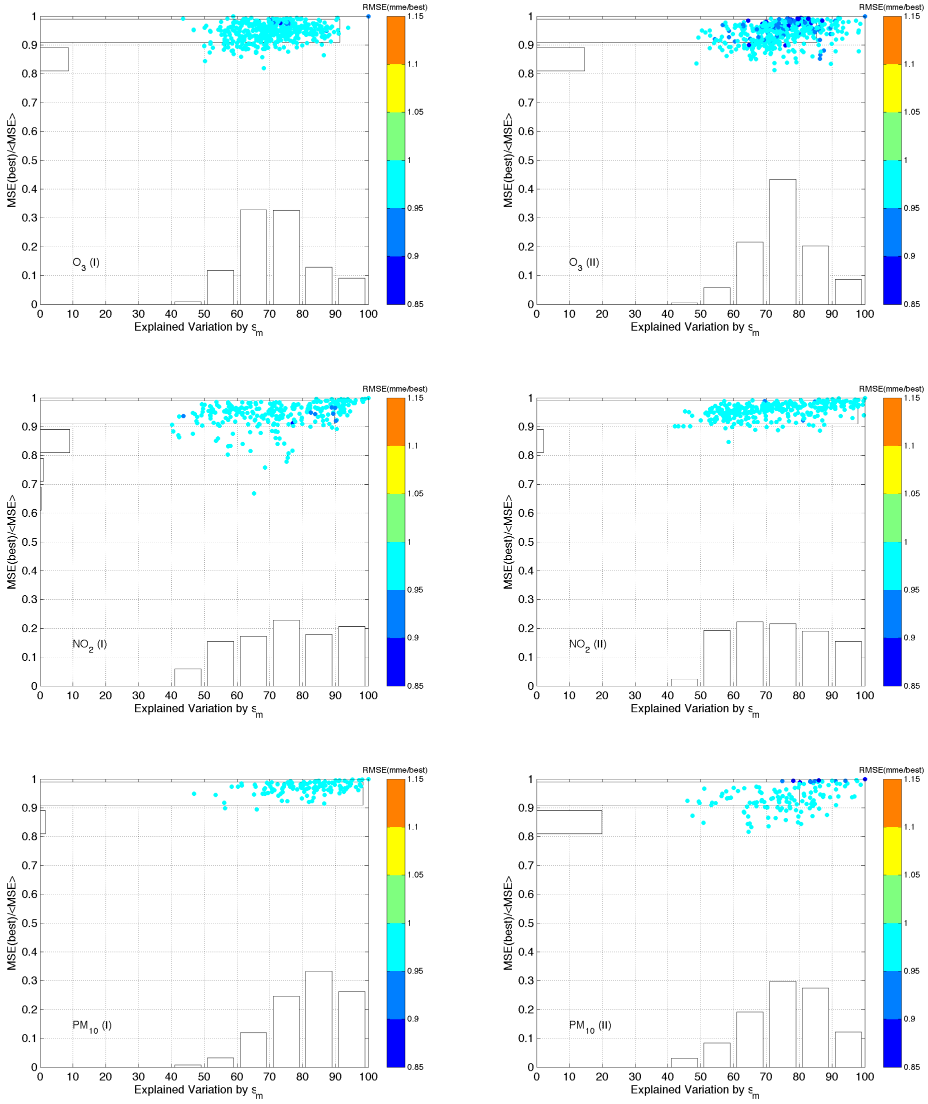


Figure 6. Like Figure 5 but showing the N_{EFF} with respect to skill difference and error dependence.



1 **Figure 7. Like Figure 5 but for the *mme<* skill in the reduced ensemble. Please note the change in**
2 **the colorscale.**

3

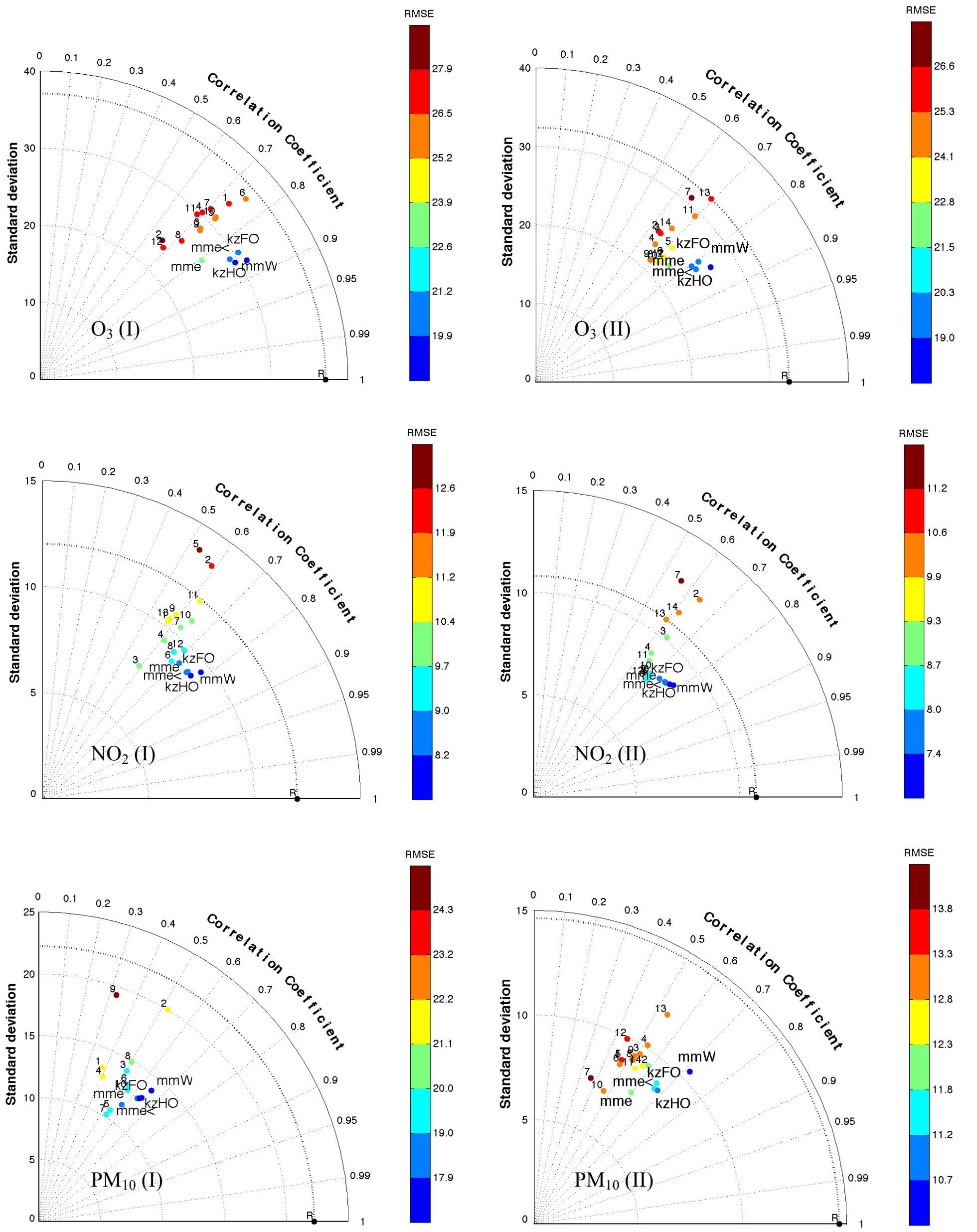


Figure 8. Composite skill of all deterministic models and ensemble estimators (mme, mme<, kzFO, kzHO, mmW) through Taylor plots. The point R represents the reference point (i.e. observations).

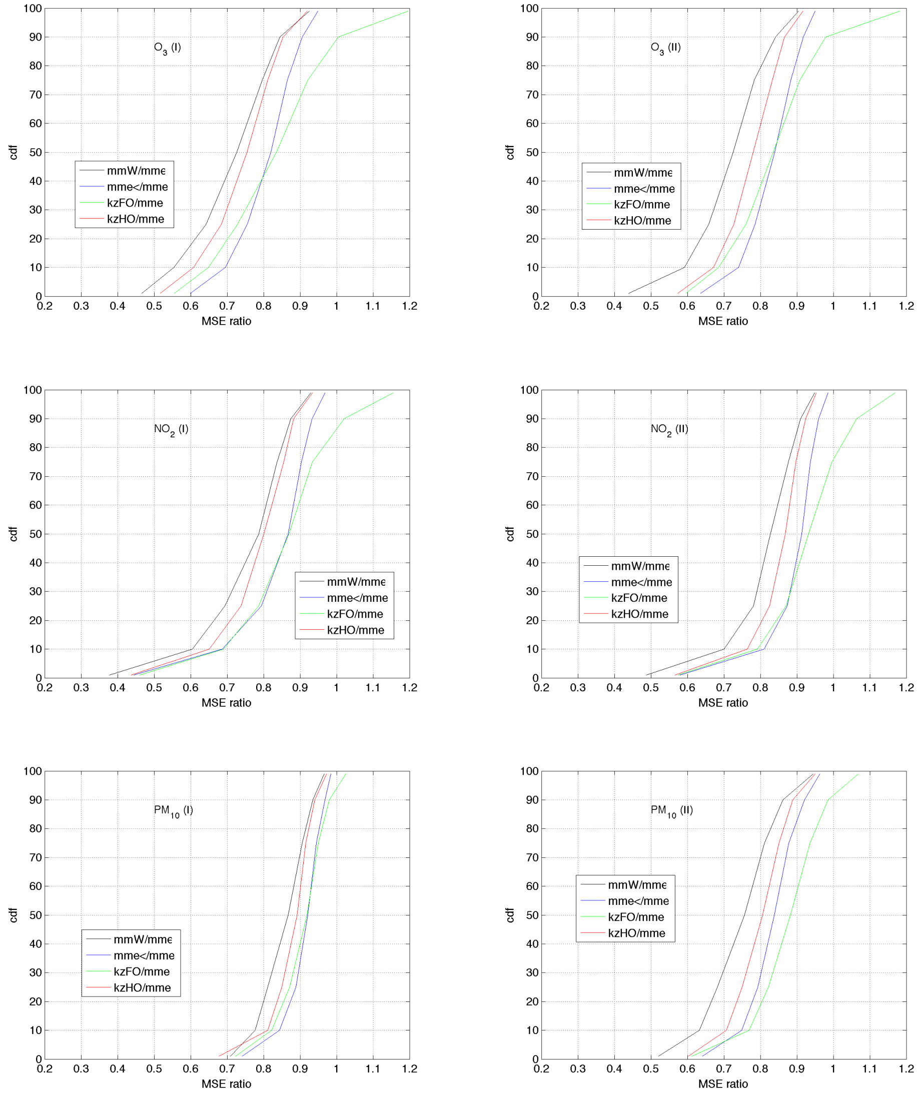


Figure 9. The cumulative density function of the indicator MSE_X/MSE_{MME} (X = mmW, mme<, kzFO, kzHO) evaluated at each monitoring site for the examined species of the two AQMEII phases.

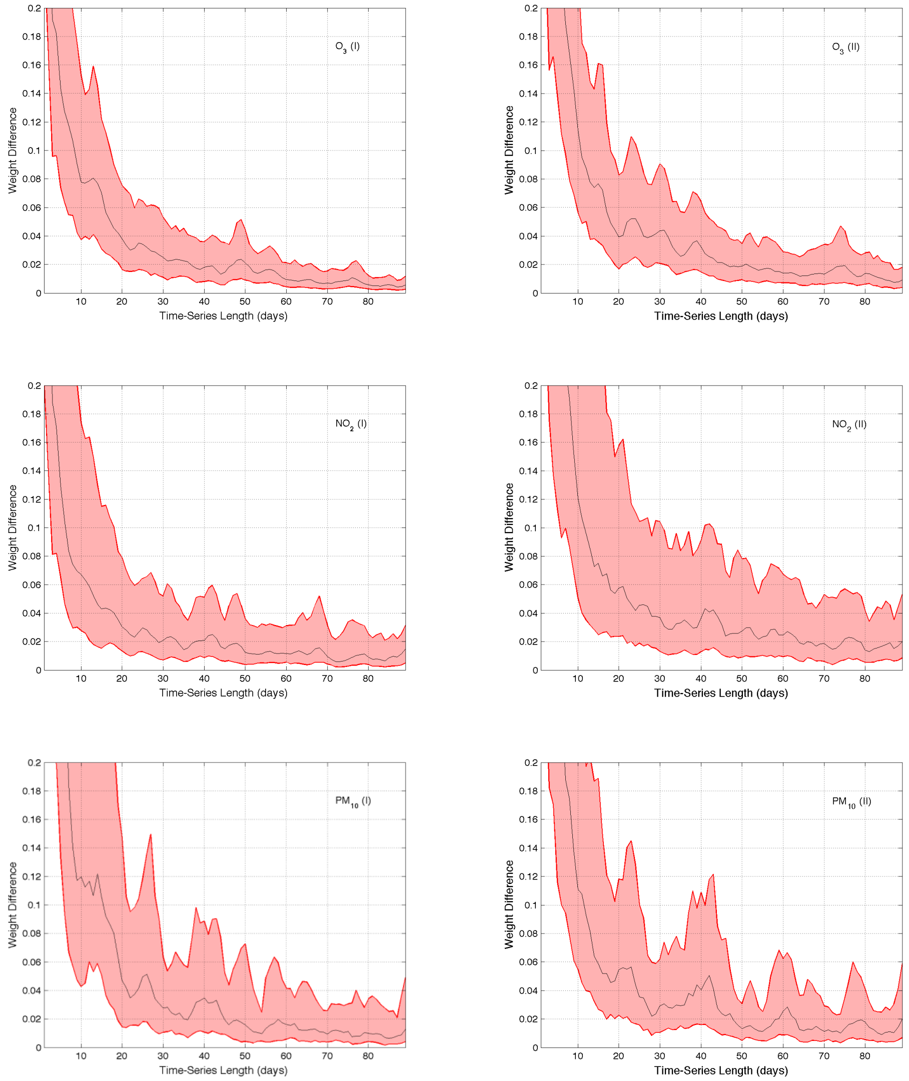


Figure 10. The interquartile range over all stations of the day-to-day difference in the weights arising from variable time-series length.