Atmos. Chem. Phys. Discuss., 14, 15803–15865, 2014
www.atmos-chem-phys-discuss.net/14/15803/2014/
doi:10.5194/acpd-14-15803-2014

Atmospheric Chemistry and Physics Discussions

De praeceptis ferendis: good practice in multi-model ensembles

I. Kioutsioukis[1,2] **and S. Galmarini**[1]

[1]European Commission, Joint Research Center, Institute for Environment and Sustainability, Ispra (VA), Italy
[2]Region of Central Macedonia, Thessaloniki, Greece

Received: 12 March 2014 – Accepted: 4 June 2014 – Published: 17 June 2014

Correspondence to: S. Galmarini (stefano.galmarini@jrc.ec.europa.eu)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Abstract

Ensembles of air quality models have been formally and empirically shown to outperform single models in many cases. Evidence suggests that ensemble error is reduced when the members form a diverse and accurate ensemble. Diversity and accuracy are hence two factors that should be taken care of while designing ensembles in order for them to provide better predictions. There exists a trade-off between diversity and accuracy for which one cannot be gained without expenses of the other. Theoretical aspects like the bias-variance-covariance decomposition and the accuracy-diversity decomposition are linked together and support the importance of creating ensemble that incorporates both the elements. Hence, the common practice of unconditional averaging of models without prior manipulation limits the advantages of ensemble averaging. We demonstrate the importance of ensemble accuracy and diversity through an inter-comparison of ensemble products for which a sound mathematical framework exists, and provide specific recommendations for model selection and weighting *for multi model ensembles*. To this end we have devised statistical tools that can be used for diagnostic evaluation of ensemble modelling products, complementing existing operational methods.

1 Introduction

A forecast is considered complete if it is accompanied by an estimate of its uncertainty (AMS, 2002). This generally requires the embedding of the modelling process into either a deterministic perturbation scheme (e.g., tangent linear, direct decoupled) or a probabilistic framework (e.g., Monte Carlo). Such approaches are used to quantify the effects of uncertainties arising from variations in model input (e.g., initial and boundary conditions, emissions) or model structure (e.g., parameterizations, numerical discretization).

ACPD
14, 15803–15865, 2014

Deterministic approaches are fast but they rely on the validity of the linearized approximation of error growth (Errico, 1997). The availability of computing means in recent years has boosted the application of the probabilistic approach (Leith, 1974) because it can sample the sources of uncertainty and their effect on the prediction error in a non-linear fashion without requiring model modifications. However, the sampling of the whole range of uncertainty could be quantified with the construction of very large sets of simulations that correspond to alternative configurations (data or model). This is unrealistic for 3-D models and leads to a hybrid scheme called *ensemble forecasting* (Molteni et al., 1996; Tracton et al., 1993). It is probabilistic in nature but it generally does not sample the input uncertainty in a formal mathematical way, limiting the extent of the available mathematical bibliography to interpret the results.

Single model ensembles (e.g. Mallet et al., 2006) assume the model is perfect and consist from a set of perturbed initial conditions and/or physics perturbations. It is traditionally used in weather forecasting, which is primarily driven by the initial conditions uncertainty. *Multi model ensembles* (e.g., Galmarini et al., 2004) (MME) quantify principally the model uncertainty as they are generally applied to the same exercise (i.e. input data). This approach is usually implemented in air pollution and climate modelling studies, where the uncertainty is predominantly process driven. The models in a MME should ideally have uncorrelated errors. Under such condition, the deterministic forecast generated from the MME mean is better than any single-model forecast due to the averaging out of the errors (Kalnay, 2003). Besides that, the MME spread quantifies the output uncertainty, providing an estimate of the forecast reliability.

The simulation error of the ensemble mean outperforms the one of the individual ensemble members only if the assumption that the models are i.i.d. (independent and identically distributed around the true state), is satisfied (Knutti et al., 2010). The i.i.d. assumption, however, is seldom object of verification and is rarely met in practice, with the net result that the simple ensemble mean does not guarantee the lowest error (higher accuracy) among all possible combinations. In such cases, the ensemble mean brings redundant information particularly for the upper and lower quartiles, making for

ACPD
14, 15803–15865, 2014

example the analysis of extremes risky. Extra effort is required in order to obtain an improved deterministic forecast such as the MME mean for i.i.d. members. The optimal solution requires some training phase, during which the models are manipulated towards the construction of an ensemble with a symmetric distribution around the truth. This can be achieved through either a weighting scheme that keeps all members (e.g., Potempski and Galmarini, 2009) or with a reduced ensemble (Galmarini et al., 2013; Solazzo et al., 2013) that makes use of only an *effective number of models*. Both approaches result in the optimum distribution of the models in the respective workspace.

Ensembles tend to yield better results when there is a significant diversity among the models. Many ensemble methods, therefore, seek to promote diversity among the models they combine. However, a definite connection between diversity and accuracy is still lacking. An accurate ensemble does not necessarily consist of independent models. There are conditions under which an ensemble with redundant members could be more accurate than one with independent members. Seen from another angle, similar to diversity, ensembles also tend to produce better results when they contain negatively correlated models. Ideally, the most accurate ensemble consists of members that are identically distributed around the observations. This property could not be parameterized as a monotonic function of characteristic properties for the selected members like independence, redundancy, etc.

In this work, we demonstrate the properties of a MME through the unprecedented database built from regional air quality models within the Air Quality Modelling Evaluation International Initiative (AQMEII). The idea is to exploit ways to promote the properties, through model selection or weighting, that guarantee a symmetric distribution of errors. This will require a training phase and will lead to a comparison between static and dynamic weights and their temporal scales predictability. Our motivation is to depict some best practices for air quality ensembles.

The paper is structured as follows: in Sect. 2, theoretical evidence on multi-model ensembles is presented. In Sect. 3, an example serves to show the contributing factors to the ensemble error. In Sect. 4 we analyse different properties of the ensemble

ACPD
14, 15803–15865, 2014

and their impact on the output error using the AQMEII data. In Sect. 5 we extent the results obtained in the previous section into spatial forecasting. Conclusions are drawn in Sect. 6.

2 Theoretical considerations

The aim of this section is to outline the documented mathematical evidence towards the reduction of the ensemble error. The following notation is used throughout the text:

Ensemble members (output of modelling systems) f_i

Ensemble $\bar{f} = \sum_{i=1}^{M} w_i f_i, \sum w_i = 1$

Desired value (measurement) μ

where M is the number of available members and w_i are the weights.

3 The bias-variance-covariance decomposition of the error

The bias-variance decomposition states that *the squared error of a model can be broken down into two components: bias and variance*.

$$\begin{aligned}\mathrm{MSE}\left(\bar{f}\right) &= E\left[\left(\bar{f}-\mu\right)^2\right] \\ &= E\left[\left(\bar{f}-\mu\right)^2\right] - \left[E\left(\bar{f}-\mu\right)\right]^2 + \left[E\left(\bar{f}-\mu\right)\right]^2 \\ &= \mathrm{Var}\left[\bar{f}-\mu\right] + \left[\mathrm{Bias}\left(\bar{f},\mu\right)\right]^2 \end{aligned} \quad (1)$$

ACPD
14, 15803–15865, 2014

The two components usually work in opposition: reducing the bias causes a variance enhancement, and vice versa. The *dilemma* is thus finding an optimal balance between bias and variance in order to make the error as small as possible (Geman et al., 1992; Bishop, 1995).

The error decomposition of a single model (case $M = 1$ in Eq. 1) can be extended to an ensemble of models, in which case the variance term becomes a matrix whose off-diagonal elements are the covariance among the models and the diagonal terms are the variance of each model:

$$\begin{aligned}
\mathrm{Var}\left[\bar{f}-\mu\right] &= \mathrm{Var}\left[\frac{1}{M}\sum f_i - \mu\right] = \frac{1}{M^2}\mathrm{Var}\left[\sum (f_i-\mu)\right] \\
&= \frac{1}{M^2}\left[\sum \mathrm{Var}(f_i-\mu) + 2\sum_{i<j}\mathrm{Cov}\left(f_i-\mu, f_j-\mu\right)\right] \\
&= \frac{1}{M}\left[\frac{1}{M}\sum \mathrm{Var}(f_i-\mu)\right] + \frac{M-1}{M}\left[\frac{1}{\frac{M(M-1)}{2}}\sum_{i<j}\mathrm{Cov}\left(f_i-\mu, f_j-\mu\right)\right] \\
&= \frac{1}{M}\overline{\mathrm{VarE}} + \left(1-\frac{1}{M}\right)\overline{\mathrm{CovE}}
\end{aligned}$$

$$\left[\mathrm{Bias}\left(\bar{f},\mu\right)\right]^2 = \left[\frac{1}{M}\sum f_i - \mu\right]^2 = \left[\frac{1}{M}\sum (f_i-\mu)\right]^2 = \overline{\mathrm{bias}}^2$$

Thus, *the squared error of ensemble can be broken into three terms, bias, variance and covariance*. Substituting the terms in Eq. (1), the *bias-variance-covariance* decomposition (Ueda and Nakano, 1996; Markowitz, 1952) is presented as follows:

$$\mathrm{MSE}\left(\bar{f}\right) = \overline{\mathrm{bias}}^2 + \frac{1}{M}\overline{\mathrm{varE}} + \left(1-\frac{1}{M}\right)\overline{\mathrm{covE}} \qquad (2)$$

Equation (2) is valid for uniform ensembles, i.e. $w_i = \frac{1}{M}$. The terms $\overline{\mathrm{bias}}$ and $\overline{\mathrm{varE}}$ are the average bias and variance of the ensemble members error (modelled time-series

ACPD
14, 15803–15865, 2014

minus observed time-series) respectively while the new term $\overline{\text{covE}}$ is the average covariance between pairs of distinct ensemble members error. From Eq. (2) follows:

- The more ensemble members we have, the closer is $\text{Var}\left[\bar{f} - \mu\right]$ to $\overline{\text{covE}}$;
- $\overline{\text{bias}}^2$ and $\overline{\text{varE}}$ are positive defined, but $\overline{\text{covE}}$ can be either positive or negative.

The error of an ensemble of models not only depends on the bias and variance of the ensemble members, but also depends critically on the amount of correlation among the model's errors, quantified in the covariance term. The *covariance* term indicates the *diversity* or *disparity* between the member networks as far as their error estimates are concerned. Hence, the more diverse the individual members an ensemble has, the less correlated they would be, which seems obvious. Given the positive nature of the other two terms and the trade-off between them, the quadratic error is minimized only in cases the covariance term is as little as possible. The lower the covariance term, the less the error correlation amongst the models, which implies reduced error of the ensemble. This is the main reason why *diversity* in ensembles is extremely important.

3.1 The accuracy-diversity decomposition of the error

Krogh and Vedelsby (1995) proved that at a single datapoint *the quadratic error of the ensemble estimator is guaranteed to be less than or equal to the average quadratic error of the component models*:

$$\left(\bar{f} - \mu\right)^2 = \sum_{i=1}^{M} w_i \left(f_i - \mu\right)^2 - \sum_{i=1}^{M} w_i \left(f_i - \bar{f}\right)^2 \tag{3}$$

Equation (3) shows that for any given set of models, the error of the ensemble will be less than or equal to the average error of the individual models. Of course, one of the individuals may in fact have lower error than the average, and lower than even the

ACPD
14, 15803–15865, 2014

ensemble, on a particular pattern. But, given that we have no criterion for identifying that best individual, all we could do is pick one at random. In other words, taking the combination of several models would be better on average over several patterns, than a method which selected one of the models at random.

The decomposition (3) is composed by two terms. The first is the weighted average error of the individuals (accuracy). The second is the diversity term, measuring the amount of variability among the ensemble member predictions. Since it is always positive, it is subtractive from the first term, meaning the ensemble is guaranteed lower error than the average individual error. The larger the diversity term, the larger is the ensemble error reduction. Here one may assume that the optimal error belongs to the combination that minimizes the weighted average error and maximizes the variability among the ensemble members. However, as the variability of the individual members rise, the value of the first term also increases. This therefore shows that diversity itself is not enough; we need to get the right balance between diversity and individual accuracy, in order to achieve lowest overall ensemble error (accuracy-diversity trade-off).

Unlike the bias-variance-covariance decomposition, the accuracy-diversity decomposition is a property of an ensemble trained on a single dataset. The exact link between the two decompositions is obtained by taking the expectation of the accuracy-diversity decomposition, assuming a uniform weighting. It can be proved that (Brown et al., 2005):

$$
\begin{aligned}
E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i-\mu)^2-\frac{1}{M}\sum_{i=1}^{M}\left(f_i-\bar{f}\right)^2\right) &= \overline{\text{bias}}^2+\frac{1}{M}\overline{\text{varE}}+\left(1-\frac{1}{M}\right)\overline{\text{covE}} \\
E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i-\mu)^2\right) &= \Omega+\overline{\text{bias}}^2 \qquad (4) \\
E\left(\frac{1}{M}\sum_{i=1}^{M}\left(f_i-\bar{f}\right)^2\right) &= \Omega-\frac{1}{M}\overline{\text{varE}}-\left(1-\frac{1}{M}\right)\overline{\text{covE}}
\end{aligned}
$$

ACPD
14, 15803–15865, 2014

The Ω term constitutes the interaction between the two parts of the ensemble error. This is the average variance of the models, plus a term measuring the average deviations of the individual expectations from the ensemble expectation. When we combine the two sides by subtracting the diversity term from the accuracy term from the average MSE, the interaction terms cancel out, and we get the original bias-variance-covariance decomposition back. The fact that the interaction exists illustrates why we cannot simply maximize diversity without affecting the other parts of the error – in effect, this interaction quantifies the accuracy-diversity trade-off for uniform ensembles.

3.2 The analytical optimization of the error

The two presented decompositions and their inter-connection are valid for uniform ensembles, i.e. $w_i = \frac{1}{M}$. Both indicate that error reduction in an ensemble can be achieved through selecting a subset of the members that have some *desired properties* and taking their arithmetic mean (equal weights). An alternative to this approach would be the use of non-uniform ensembles. Rather than selecting members, it keeps all models and the burden is passed to the assignment of the *correct weights*. A brief summary of the properties of non-uniform ensembles is presented in the following paragraphs.

The construction of the optimal ensemble has been exploited analytically by Potempski and Galmarini (2009). They provide different weighting schemes for the case of uncorrelated and correlated models by means of minimizing the MSE. Under the assumed condition of the models independence of observations and assuming also that the models are all unbiased (bias has been removed from the models through a statistical post-processing procedure), the formulas for the one-dimensional case (single-point optimization) are given in Table 1. Also, whether correlated or not, the models are assumed as random variables (i.e. their distribution is identical).

The optimal weights correspond to the linear combination of models with the minimum MSE. This can be considered as a transfer function that distributes identically the models around the truth. Using equal weights, the ensemble mean has lower MSE than the candidate models given specific conditions. For example, the arithmetic mean

ACPD
14, 15803–15865, 2014

of a 3-member ensemble, where models are uncorrelated, has lower MSE than the best candidate model only if the MSE ratio (worst/best) of the models is lower than 4. In other words, the RMSE ratio may not exceed 2, implying that the individual members should not be very different. The conditions for correlated models are more restrictive. Further, unlike the case of uncorrelated models, optimal weights for correlated models can be negative (Table 1).

3.3 Weakness of the traditional practice

The material presented in this section demonstrated clearly through a well-defined mathematical formulation that building ensembles on the basis of "including as more models as possible to the pool and taking their arithmetic mean" is generally far from optimal as it relies on conditions that are normally not fulfilled. At the same time, it provided the necessary ingredients for ensemble building, using either the entire members with weights assigned or a subset of them with equal weights. Specifically, the optimization of the ensemble error:

- through the bias-variance-covariance decomposition, points towards the bias correction of the models and the use of uncorrelated or negatively correlated ensemble members (equal weights, sub-ensemble);
- through the accuracy-diversity decomposition, relies on finding the trade-off point between accurate and diverse members (equal weights, sub-ensemble);
- through analytical formulas, provides weights for all ensemble members dependent on their error covariances (Potempski and Galmarini, 2009) (unequal weights, full ensemble);

Unlike the simple arithmetic mean of the entire ensemble, it is clear that all aforementioned cases require a learning process/algorithm. The aim of this work is to assess and compare the predictive skill of three ensemble products with well-defined mathematical properties, namely:

ACPD
14, 15803–15865, 2014

1. the arithmetic mean of the entire ensemble (mme)
2. the arithmetic mean of an ensemble subset (mme <), linked to the error decompositions (2.1, 2.2)
3. the weighted mean of the entire ensemble (mmW), linked to the analytical optimization (2.3).

4 Example

In this section, we present a theoretical example aimed at illustrating the basic ingredients of ensemble modeling discussed. Fourteen samples of 5000 records each have been generated; thirteen corresponding to output of model simulations and one acting as the observations. These synthetic time-series have been produced with Latin-hypercube sampling (McKay et al., 1979). The reason of selecting Latin-hypercube sampling over random sampling, besides the correct representation of variability across all percentiles (Helton and Davis; 2003), is its ability to generate random numbers with predefined correlation structure (Iman and Conover, 1982; Stein, 1987).

Figure 1 shows the RMSE distribution of the mean of all possible combinations of the ensemble members ($M = 13$) as a function of the ensemble size ($k = 1, \ldots, M$). The number of combinations of any k members is given by the factorial $\binom{M}{k}$, resulting in a total of 8191 combinations in this setting (e.g., 286 for $k = 3$, 1716 for $k = 6$, etc.). In the case of i.i.d. random variables (top row, left plot), increasing the number of members (k) moves the curves toward more skillful model combinations, as anticipated from the bias-variance-covariance decomposition. Further, the optimal weights do not deviate from the equal weighting scheme (with small random fluctuations though) traditionally used in the MMEs. Hence, the optimal combination (mme <) and the optimal weighted combination (mmW) coincide. However the i.i.d. situation is unrealistic for MME, therefore we will examine the ensemble skill by perturbing independently the three statistical measures of bias, variance and covariance.

ACPD
14, 15803–15865, 2014

Bias has been introduced into the ensemble by shifting the distribution of two-third of the models by a small amount, making one-third of the models unbiased, one-third biased positively and one-third biased negatively. The RMSE distribution of all possible combinations (top row, right plot) does not appear symmetric with respect to the mean RMSE, with particular distortions at the maximum RMSE for $k \leq 4$ (i.e., one-third of models). This upper bound is delineated from the ensemble combinations of all biased members of equal sign. Several combinations with multi-model error lower than the error of the full ensemble mean exist; at the same time, the whole RMSE distribution spans higher values compared to the i.i.d. case (note the change in scale). The optimal combination uses all unbiased models plus same amounts of biased equally members from both sides. As for the weighted ensemble, no clue can be inferred as its weights by definition assume unbiased models.

The effect of variance perturbations is displayed in the middle row. One third of the members (with ids 10–13 in particular) had deflated (left) or inflated (right) variance. Due to the bias-variance dilemma, the case with smaller variance achieves lower RMSE for low k (at the expense of PCC though) while the opposite is true for the cases exhibiting larger variance. The optimal weighted combination gives higher weight to the under-dispersed members and lower weight to the over-dispersed ones.

All examined cases so far were uncorrelated. Next, a positive correlation (bottom row, left plot) has been introduced among the first three members (ids 1–3) and separately, a negative correlation between two members (bottom row, right plot), with ids 5 and 8 namely. The upper (lower) bound of the error distribution of the combinations is distorted towards higher (lower) values by introducing positively (negatively) correlated members. Positively correlated members bring redundant information, where individual errors are added rather than cancelled out upon MME averaging. The optimal combination, for the case of positive correlations utilizes all i.i.d. members plus only one from each redundant cluster; for negative ones, it tends to use only anti-correlated members. The same is seen also for the optimal weighted scheme: positively correlated

ACPD
14, 15803–15865, 2014

members are treated as one, negatively correlated are significantly promoted over the i.i.d. members.

To summarize, ensemble averaging is a good practice when models are i.i.d. In reality, models depart from this idealized situation and MME brings together information from biased, under- and over-dispersed as well as correlated members. Under these circumstances, the equal weighting scheme or the use of all members well masks the benefits behind ensemble modelling. This example serves as a practical guideline to better understand the real issues faced when dealing with biased, inter-dependent members.

5 Empirical evidence

We now investigate the ensemble properties mentioned in the theoretical introduction using real-life spatially aggregated time-series from AQMEII (Rao et al., 2011). AQMEII was started in 2009 as a joint collaboration of the EU Joint Research Centre, the US-EPA and Environment Canada with the scope of bringing together the North American and European communities of regional scale air quality models. Within the initiative the two-continent model evaluation exercise was organized which consisted in having the two communities to simulate the air quality over north America and Europe for the year 2006 (full detail in Galmarini et al., 2012a). Data of several natures were collected and model evaluated (Galmarini et al., 2012b). The community of the participating models, which forms a multi-model set in terms of meteorological driver, air quality model, emission and chemical boundary conditions, is presented in detail in Galmarini et al. (2013). The model settings and input data are described in detail in Solazzo et al. (2012a, b), Schere et al. (2012), Pouliot et al. (2012), where references about model development and history are also provided.

The analysis considers *hourly* time-series for the JJA period. For European ozone, the ensemble constitutes from thirteen models that give rise to 8191 different combinations (ensemble products). All data used refer to Phase I of the initiative. The evaluation

ACPD
14, 15803–15865, 2014

of the examined ensemble products (mme, mmW, mme <) will rely on several indices of error statistics. We present them in Table 2. Those metrics can be used for the validation of each single ensemble configuration (f_i) as well as for the ensemble mean (f_{ens}).

Once the basic ingredients have been identified, a spatial analysis will follow in the next section.

5.1 Issue 1: properties of the MME error

According to the bias-variance-covariance decomposition, bias is an additive factor to the MSE and model outputs should be corrected for their bias before any ensemble treatment. The analytical optimization of the ensemble error and the defined weights (Table 1) also assume bias corrected simulations. Here we do not intend to review the available algorithms for the statistical bias correction (e.g., Dosio and Paurolo, 2011); the correction applied in this work refers to a simple shift of the whole distribution without any scaling or multiplicative transfer function.

The RMSE of each possible combination as a function of the ensemble order justifies the statement obtained theoretically (Fig. 2, top left), namely that the RMSE of the ensemble mean is lower than the mean error of the single models. This does not prevent individual model errors to be lower than the ensemble mean error. The curve, although it originates from real data (EU4r), shares the same properties with its synthetic counterpart (previous section). Specifically:

- the ensemble average reduces the *maximum* RMSE as the order is increased
- a plateau is reached at the *mean* RMSE for $k < M$, indicating that there is no advantage, on average, to combine more than k members ($k \sim 6$).
- a *minimum* RMSE, among all combinations, systematically emerges for ensembles with a number of members $k < M$ ($k \sim 3$–6).

ACPD
14, 15803–15865, 2014

The probability density function (pdf) of the RMSE plotted for $k = 6$ (similar for other values) demonstrates that there exist many combinations with lower error than the ensemble mean or the minimum of ensemble mean and best single model. Those skilled groupings represent roughly 40 % of the total combinations in the first case, quasi constant across different sub-regions and below 40 % with high variability across different sub-regions (due to the spatial variability of best model's skill) in the second case. This number is small (below 50 %) with non-random structure, implying that random draws from the pool of models is highly unlikely to produce better results than the ensemble mean; at the same time, it is high enough to leave space for significant improvements of the mme. The fractional contribution of individual models (for $k = 6$) to those best sub-groups is given with the red numbers. The normalization has been done with the number of combinations that includes each model id (for $k = 6$ it equals 792). For example, among all combinations, at $k = 6$, that may contain the model with id 12, two-thirds of them (67 %) are skilful. The percentages indicate preference to combinations including more frequently some models (e.g., 4, 6, 9, 12) but at the same time they do not isolate any single model. Further, the optimal weights of the full ensemble given with the bar plot (multiplied by a factor of 10) have a complex pattern as a result of different model variances and covariances. Definitely, they depart from homogeneity (equal weighting scheme shown with the red straight line).

The error, variance and covariance (with observations) of the thirteen ensemble members are presented in a Taylor plot (Fig. 2, top right). They visually form three clusters. Low skill cluster includes models 1, 2 and 10 that have the highest error, minimum correlation with observed data and appear under-dispersed. Model 5 also belongs to that group but has improved variance. The intermediate skill cluster contains models 3, 6, 7, 11 and 13 with average (11, 13) to low (3, 7, 6) error, correlation ranging from 0.8 (11, 13) to 0.9 (6) but all models are under-dispersed. The highest skill cluster (4, 8, 9, 12) includes members with low error, high correlation and the right variance ratio (with a light over-dispersion though). Compared with the participation statistics of the previous graph, we see that the models contributing more frequently to skilled combinations

ACPD
14, 15803–15865, 2014

belong to highest skill cluster; the contrary is true for the low skill cluster. Good models have at least twice as much probability to form good ensemble groups compared to low skill models. On the other hand, even low skill models can yield good results under proper combination scheme. Overall, the multi-model average (mme) is a robust estimate with lower error than the candidate models but with reduced variance.

According to the presented decompositions, the combinations with reduced error should have a better balance between accuracy and diversity than the ensemble mean. A 2-dimensional plot of accuracy vs. diversity, with RMSE displayed as a third dimension (in color) is shown in Fig. 2 (middle row, left). The black lines define the convex hull in the (accuracy, diversity) space of specific ensemble order, ranging from 2 in the outer polygon to 12 (i.e. M-1) in the innermost one. As expected theoretically, the separate optimization of accuracy and diversity will not produce the best ensemble output. For all ensemble orders, the optimal combination consists of accurate averaged representations of sufficient diversity between members, i.e. with an ideal trade-off between accuracy and diversity. In particular, all skilled combinations are clearly seen in this stratified chart; they form a well-defined area, traceable according to the ensemble order, that contains combinations with accuracy better than the average accuracy and *ideal* diversity (within a wide-range though) for the specific accuracy. For example, combinations of average accuracy form skilled ensemble products only if their diversity is very high. Analogously, combinations with very good accuracy form skilled ensemble products only if their diversity is not relatively high. Diversity with respect to the ensemble mean can be derived independently to the observations. This however is not true for the accuracy part implying that a minimum training is required. Last, we observe that as ensemble order increases, accuracy and diversity become more and more bounded (with accuracy being more disperse than diversity), limiting any improvement. The ensemble product with the right trade-off between accuracy and diversity (mme <), improves the error and also replicates better the observed variance (Fig. 2, top right). The same conclusion is also reached for the ensemble product generated through an analytical optimization (mmW).

ACPD
14, 15803–15865, 2014

Similar results are obtained in terms of the variance-covariance decomposition in Fig. 2 (middle row, right). Here the convex hull areas, ranging from 3 to 12, move towards lower mean variance and higher mean covariance with increasing ensemble order. Higher spread is evidenced for the covariance term. As we include more members in the ensemble, the variance term in the decomposed error formula is becoming lower but the covariance term is deteriorated. Skilful combinations have relatively low covariance. Ensembles consisting of highly correlated members bring redundant errors in the ensemble that does not cancel out upon averaging, producing overall bigger errors.

The conditions granting an ensemble superior to the best single model is also attempted in Fig. 2 (last row). We have seen that the mme error is a function of accuracy and diversity (or variance and covariance). An analytical optimization of this error (Potempski and Galmarini, 2009) yields the necessary conditions for being lower than the one of the best model (Table 1). For uncorrelated models, the only constrain is the skill difference (MSE ratio) of the worst over the best single model; for correlated models, it is more complex and it also related to the amount of redundancy in the ensemble, i.e. the error dependence. The explained variation by the highest eigenvalue reflects the degrees of freedom in the ensemble (and hence the redundancy). The pairwise plot as a function of the RMSE ratio of mme over the best single model (for ensemble order = 6, left) shows that mme can outscore any single model provided the model error ratio and redundancy follows a specific pattern. For example, the benefits of ensemble averaging are devalued if we combine members that have big differences in skill and dependent errors.

The distribution of the models around the truth should possess higher symmetry for the case of the skilled combinations. This thought is illustrated in Fig. 3 by means of the cumulative density function of the full ensemble (left plot) and the reduced optimal one consisting of only 3 models (right plot). The ensemble mean is a good candidate for the interquartile range but its ability to capture extremes (tails of the distribution) is weaker since the majority of models systematically over- or under-estimate extreme

ACPD
14, 15803–15865, 2014

concentrations (over for min, under for max). The optimal combination (right plot) corresponds to the models that are generally distributed randomly around the truth, i.e. when put together their errors cancel out upon averaging. This is evidenced across all percentiles. Such combinations are capable for studies of *extreme events*. At the same time, their previsions have lower spread (i.e. *uncertainty*). As seen before, for the selected ensemble order ($k = 3$), the skilful combinations have small error covariance (they are both accurate and diverse).

To summarize, the unconditional averaging of ensemble members is highly unlikely to systematically generate a forecast with higher skill than its members across all percentiles as models generally depart significantly from behaving as an i.i.d. sample. Further, the ensemble mean is superior to the best single model given conditions that relate to the individual skill of the members and the ensemble redundancy. Good practice includes bias correction of the models and the construction of a "balanced" ensemble through either a weighting approach (straightforward) or through the clustering of members with the desired properties (in the form of accuracy/diversity, variance/covariance, redundancy, etc). In the next section we explore some measures to achieve such clusters.

5.2 Issue 2: clustering measures

Given a dataset of N instances $\mathbf{X} = \{X_1, X_2, \ldots, X_N\}$, a clustering algorithm generates r disjoint clusters based on a distance metric. Each clustering solution is a partition of the dataset $\mathbf{X}$ into K_i ($1 \leq i \leq r$) disjoint clusters of instances. A typical output of a clustering algorithm is a dendrogram, where redundant models are grouped together and the level of similarity among groups is based on the distance between the elements of the input matrix. Clustering algorithms are sensitive to the controlling options (the *agglomerative method*, the *distance metric*, the *number of clusters* and the *cut-off distance*) that need to be determined case by case (Fern and Brodley, 2004). Here, we use the the unweighted pair-group average as the *agglomeration method* and the standard Euclidean distance as the *distance metric.* The clustering algorithm has been

ACPD
14, 15803–15865, 2014

utilized against two different input matrices, namely the corr (e_i, e_j) and the corr (d_i, d_j) (for details see Solazzo et al., 2013). Common practice suggests cutting the dendrogram at the height where the distance from the next clustered groups is relatively large, and the retained number of clusters is small compared to the original number of models (Riccio et al., 2012). For this reason, the *cut-off value* (the threshold similarity above which clusters are to be considered disjointed) is set to 0.10 for corr (d_i, d_j) and 0.4 for corr (e_i, e_j).

The application of the above produced five disjointed clusters (Fig. 4, top row). Looking at the corr (e_i, e_j) dendrogram (left plot) or the corr (d_i, d_j) dendrogram (right plot), for example, the two main branches at the top further split into two more at a relatively low similarity level, suggesting a plausible way to proceed. A parallel inspection at the Taylor plot (Fig. 2) reveals the similarities of each cluster in terms of error, correlation and variance. Clustering according to corr (d_i, d_j) generates the visual clusters of the Taylor plor while corr (e_i, e_j) clustering is coarser. Many ensemble combinations with non-redundant members can be inferred from those plots; in addition, combinations that should be avoided are also marked.

A decomposition of each deterministic model's error into spectral components provides another roadmap for clustering models. Using four components (ID, DU, SY, LT: for details see Galmarini et al., 2013) with the Kolmogorov–Zurbenco filter (Zurbenko, 1986), it is evident that the models with particularly high total error have all deficiencies with specific spectral component (Fig. 4, bottom row). The diurnal component in models 1, 2, 5 and 10 has error similar to the total error of the other models (from all components). If we repeat the analysis with two-component decomposition (ID+DU, SY+LT) that has limited energy leak between them, the conclusion still remains the same. Models with particular high systematic errors, as evidenced through the spectral decomposition that reflects *process-based* performance, should be treated with caution within the ensemble. We do not argue that there is no benefit from using them but that their unconditional use is not fundamentally correct.

ACPD
14, 15803–15865, 2014

The integrated skill of the selected clusters through dendrograms or spectral analysis is compared through a Taylor plot (Fig. 4, bottom row). For all combinations, their trace is found in an area of high competence. At the same plot it is also displayed the skill of two products based on spectral optimization, namely the kzFO (1st order combination of the four optimal spectral components; see Galmarini et al., 2013) and kzHO (higher order combination of two quasi-independent spectral components: ID+DU, SY+LT). The kzFO provides a clear improvement over mme while the kzHO boosts further the mme < skill. The mean of those five independently generated products shows an improvement over the mme (Fig. 2). Further, the spread of the formed ensemble products is lower compared to the deterministic model's scatter, resulting in lower uncertainty. Averaging ensemble products produced through an elegant mathematical approach that constrains their properties is a potential pathway to improved *forecasts* with lower uncertainty. Last, we should point that no model was eventually excluded by the combinations but a*ll deterministic models have been utilized in at least one ensemble product.*

To summarize, good practice includes the clustering of members through multiple different algorithms that operate on dissimilar properties (like redundancy, diversity, negative correlation, spectral decomposition, etc). Averaging those combinations generated independently, hence having in principle uncorrelated errors, form a ground for skilled forecasts of lower uncertainty compared to the ensemble mean, increasing further the forecast reliability.

5.3 Issue 3: ensemble training

In this section we will test the temporal and spatial robustness of our ensemble products. We will work on the concepts of the least error combinations (mme < and mmW) using time series from different AQMEII sub-regions. Throughout this exploratory analysis (hindcast), we will derive the spatiotemporal properties of the weights and the ensemble constitution.

ACPD
14, 15803–15865, 2014

The bias-variance-covariance decomposition requires negative correlation learning algorithms (e.g., Liu and Yao, 1999; Lin et al., 2008; Zanda et al., 2007); the accuracy-diversity decomposition relies on learning diversity algorithms (e.g. Kuncheva, L. and Whitaker, 2003; Brown et al., 2005). The use of uncorrelated or diverse members alone, which are easily calculated through various metrics, does not imply an accurate ensemble. For this reason, a global handy approach to optimal ensemble forecasting and member selection, based on proven mathematical statements, still does not exist (ensemble output can be optimized through analytical formulas only for diagnostic problems). Therefore, the optimal approach under the current mathematical state is the ensemble training prior to forecasting, utilizing various approaches for model weighting (e.g., Gneiting et al., 2005; Potempski and Galmarini, 2009) or sub-selecting (see Solazzo et al., 2013 for a presentation of reducing dimensionality approaches linked to redundancy). Some key elements of this process explored hereafter include the *learning period*, the *algorithms* and their controlling properties, the *effective number of models* and the *weight stability*.

Learning period and scheme. The selection of the necessary training period should take into account the *memory capacity* of the atmosphere. Using complexity theory (e.g., Malamud and Turcotte, 1999), the ozone time-series demonstrates non-stationarity and strong persistence (e.g., Varotsos et al., 2012). This encourages the use of a scheme derived from an accurate recent representation of ozone to medium-range forecasts (e.g. Galmarini et al., 2013).

Following the evidences presented in Sect. 4.1, bias-reduction should be always applied prior to any ensemble manipulation. For this purpose, all simulations have been de-biased at each examined window size (e.g., 1 day, 3 months, etc). Results are shown for JJA 2006 at four selected sub-regions (AQMEII database), using variable window size (1 day, 2 days, 4 days, 23 days, 46 days, 92 days) and weighting/sub-selecting scheme (mme, mmW, mme <). The dependence of the $\langle \mathrm{RMSE} \rangle$ on the averaging window is shown in Fig. 5. The following inferences can be drawn:

ACPD
14, 15803–15865, 2014

1. *Error:* The skill of the deterministic models, even after bias-correction, varies with location. A very good model at one site may perform averagely in another. As for the ensemble products:

   – *mme.* The error of the ensemble mean is superior to the mean of the individual model errors (proved analytically) but is not necessarily better than the skill of the "locally" best model.

   – *mmW.* The error of the weighted ensemble mean (mmW) is always superior since it has been analytically derived to minimize the MSE. For small window sizes (less than 4 days), the mmW error is superior to the theoretically derived ensemble error if models were uncorrelated (= ⟨var⟩/nm).

   – *mme<.* The optimal error derived from a reduced-size ensemble mean (mme <) with the optimal accuracy-diversity trade-off is always lower than the error utilizing the full ensemble since models are not i.i.d. It is also, by construction, always lower than the best model's error and higher than the mmW's error.

2. *Temporal sensitivity of the error:* As the window size decreases, the ⟨RMSE⟩ decrease due to the lowering of the error variance (bias correction is applied once in the 92 days case and 92 times in the daily cases). The relative amount of decrease for mme is inversely proportional to the diurnal variability because bias correction has a more pronounced impact in cases with lower variance (Fig. 6). For example, the largest (smallest) change is seen in EU2r (EU3r) that demonstrates the least (highest) variability. The cases with high variability, where the majority of models fail to simulate well, have a prominent improvement if treated with more sophisticated ensemble products such as the mmW and mme <. On the other hand, in cases where mme is better compared to the individual models, as in the EU2r case (narrow distribution of intermediate levels), the possibility for ensemble improvements is suppressed.

ACPD
14, 15803–15865, 2014

3. *mme vs. best model:* In terms of the ensemble error gain, it is variable as it depends significantly on the individual model distributions around the truth. Without loss of generality, if we consider the 92 day case, we see that for all models the MSE ratio (worst/best) is lower than 4.37 (EU1r: 4.37, EU2r: 2.96, EU3r: 1.99, EU4r: 2.60). If models were uncorrelated, their mme would always be superior to any single model since all ratios are smaller than 14 ($= M + 1$). Figure 5 shows that only in EU4r mme is better than the individual models. This occurs because for correlated models, the condition is also restricted by the redundancy (eigenvalues spectrum). The RMSE ratio of mme over the best single model for the joint restrictions in the case of M (= 13) correlated models (Fig. 2) shows that only in EU4r the explained variation by the highest eigenvalue has the correct value for the specified model MSE ratio [EU1r (67, 2.5), EU2r (64, 1.8), EU3r (76, 1.5), EU4r (59, 1.7)]. The isolines with RMSE ratio lower than one reflect the cases with a balanced distribution of members. Indeed, in EU4r (and EU2r), the distribution of the models around the observations is more symmetric, as can be seen in Fig. 6. On the other hand, a significant departure from symmetry can be seen for EU1r and EU3r, resulting in a sub-optimal ranking of the mme. The distribution around the truth in the weighted ensemble (mmW) and the sub-ensemble (mme <) has always higher symmetry compared to mme, as can be seen in the same Figure.

4. mme vs. mme <: The estimation of the optimal weights is straightforward (Table 1), but the sub-selection of members in mme < is not. Since mme < uses equal weights, we can apply the concepts deployed by the two error decompositions and compare those properties with the ones of mme. We can then examine whether they can provide guidance towards members selection. Figure 7 displays the accuracy ratio (mme </mme) vs. the corresponding diversity ratio for all (92 in total) 1 day segments. At the same figure, variance ratio vs. its corresponding covariance ratio is also shown. The color scale indicates the mse ratio between the two ensemble means. Error minimization through mme <, for all examined sub-regions, demonstrates that the optimal combination:

ACPD
14, 15803–15865, 2014

- Improves mme < accuracy over mme and by a smaller portion lowers its diversity (i.e. in big ensembles diversity is distorted less than accuracy). In other words, between accuracy and diversity, the controlling factor in those experiments in terms of error minimization is accuracy more than diversity. This is partially explained by the fact the accuracy values have higher spread over diversity.
- Lowers mme < variance over mme (fewer members) and by a higher portion lowers its covariance. In other words, between variance and covariance, the controlling factor for error minimization is covariance more than variance.

Those findings indicate that, for example, focusing on learning diversity algorithms (maximize diversity) does not guarantee an improvement over the mme whilst the minimization of the model's error covariance is more promising.

Effective number of models. Next we discuss the concept of the effective number (N_{eff}) of models. In principle, N_{eff} reflects the degrees of freedom in the system (i.e. number of non-redundant members that cover the output space ideally and hence, can be used to generalize). It is not a property of the physical system (e.g. its principal modes of oscillation). An analytical way to calculate N_{eff} is through the formula proposed by Bretherton et al. (1999). Using eigen-analysis, it estimates the number of models needed to reproduce the variability of the full ensemble (Fig. 8). Depending on the matrix whose skeleton is investigated (i.e. error, diversity, etc), different numbers arise for N_{eff}. For example, applying the eigen-analysis on the error matrix, regardless if normalized or not, it yields $N_{\text{eff}} = 3$. If N_{eff} is calculated using the diversity, it equals 5. The largest value (7) is obtained when N_{eff} is calculated using the d_m matrix. Alternatively, one may calculate all possible combinations of models and plot them as a function of sub-ensemble size (as seen in Fig. 2). Comparing the range of N_{eff} found through eigen-analysis (3–7) with the one found through error optimization (plateau of the minimum RMSE curve in Fig. 2), we observe that they coincide. In other words, the N_{eff} calculated using the d_m (or diversity) matrix should provide the upper boundary of the ensemble size.

ACPD
14, 15803–15865, 2014

Last, the fact that N_{eff} is generally less than the full ensemble size should not be conceived that some models are useless. In fact, all models are likely to participate in the optimal N_{eff} combination, with different frequencies though (Fig. 8). Unlike models, as it has been demonstrated, many model combinations are useless.

Weight stability. We now explore the temporal robustness of the weighting schemes in order to identify the predictive skill of the products. The spread of the weights (Fig. 9) is presented for two window sizes, 1 day (92 cases, left) and 92 days (1 case, right).

- The mmW weights arise from an analytical optimization approach and they are real numbers (i.e. can be negative). The significant error minimization seen earlier (Fig. 5) for the daily simulations originates from a highly variable weighting scheme, lacking any autocorrelation pattern (not shown). The weights calculated over the whole JJA period (bar plot, right) do not generally have the same magnitude with the mean weights of the daily blocks (red circle).
- The mme < weights arise from an exploratory optimization approach and they are binary numbers (0/1). The contribution (frequency) of each model in the daily scheme (left panel), besides the peaks that vary by sub-region (a model is not optimal at all locations: for example, model 10 is frequently used in EU2r and never in EU4r), contain non-zero contribution from all ensemble members. Daily and seasonal contributions have more similarities than in the mmW case.
- Although calculated with different approaches, the weight peaks at seasonal scale (in absolute values) of the mmW and mme < are coherent.

Besides the day to day variability of the weights, we also explore another aspect of their temporal variability. The weights have been re-calculated for variable time-series length that is progressively increasing from 1 to 92 days, for the four European sub-regions (Fig. 10). Although no convergence actually occur, the mmW weights tend to stabilize after 40–60 days. The same is approximately also true for the effective number of models. Linked to the previous discussion, it provides a lower bound for the training window length that generates robust weight estimates.

ACPD
14, 15803–15865, 2014

To summarize, the relative skill of the deterministic models radically varies with location. The error of the ensemble mean is not necessarily better than the skill of the "locally" best model, but its expectation over multiple locations is, making the ensemble mean a skilled product on average. A continuous spatial superiority over all single models is feasible in ensemble products such as mmW and mme <. As those products require some training phase, good practice includes first, the identification of the temporal window length that allows robust (i.e. almost stationary) estimates for the weights and the effective number of models (memory scale of the system) and then, the training of the ensemble at those temporal scales.

5.4 Issue 4: ensemble predictability

In the previous section we estimated weights for the full ensemble (mmW) or a subset of it (mme <) in a diagnostic mode. Following the explored temporal sensitivity of the weights and N_{eff}, in this section we examine the robustness of those estimates into future cases. Are they capable of making accurate predictions or they just overfit the data over the historic epoch?

Two different sets of weights will be examined for each model, namely *static* (weights calculated over a 60 day window and applied on the remaining 30 daily forecasts) and *dynamic* (weights calculated over the most recent temporal window –day0– and applied on its successive –day0+1–). The reasoning behind the dynamic weighting testing is that, although weights (mmW) lack any autocorrelation pattern (i.e., what is optimal yesterday is not optimal today), this does not imply that this quasi-optimal weighting for tomorrow is not still a good ensemble product (mmW weights are real, hence there are infinite weighting vectors where only one is optimal but there should exist many combinations without major skill difference from the optimal).

In view of the sensitivity of the dynamic weights vs. the static ones, we investigate the skill of the ensemble products (mmW, mme <) and compare it with mme. Two additional products are also considered: the 1‰ of the total members that maximize diversity

ACPD
14, 15803–15865, 2014

(divMX1) or minimize error covariance (covMN1). The following conclusions can be inferred from Fig. 11 for the daily forecasts:

- Diversity alone generally does not outscore mme, neither with static nor with dynamic weights. It gives similar results but can also produce worse forecasts when mme is well balanced between accuracy and diversity (EU2r). This experiment shows that there may exist many diverse combinations of low accuracy. On the other hand, covariance (covMN1) is a more powerful indicator for ensemble optimization than diversity (divMX1).

- The weights derived through analytical optimization (mmW) do not correspond to products with similar properties between consecutive days. Dynamic weighting can result at high MSE values for the prediction day. On the other hand, static weights outscore all other products.

- Mme < is always superior to the mme, in all examined modes (historic, prognostic with static/dynamic weights).

Weighting is a risky process (Weigel et al., 2010) and its robustness should be thoroughly explored prior to operational forecasting. In diagnostic mode (*H*), mmW minimizes the error achieving an order of magnitude lower MSE compared to the other ensemble products (Table 3). In prognostic mode, the minimum error is obtained with mmW utilizing static weights, followed by mme < with static weights also. It is particularly noticeable the significant reduction of the peak MSE cases in those two schemes. An improvement similar to the one obtained through the mmW scheme (bias correction, model weighting) has been documented in weather forecasting (Krishnamurti et al., 1999); the phenomenological different approaches in model weighting are however equivalent. Dynamic weights could be also used for the reduced ensembles, based either on diversity (divMX1), covariance (covMN1) or error (mme <) measures, but they lead to erroneous forecasts for the analytically optimized ensemble output (mmW). We should point that those are the best expected results as they rely on ideally bias-corrected simulations.

ACPD
14, 15803–15865, 2014

To summarize, the predictability skill intensely depends on the temporal autocorrelation of the selected cost function (e.g. combinations of peak diversity, covariance, error, etc). Good practice suggests the use of static over dynamic weighting scheme for prediction purposes with the mmW or mme < ensemble products. The use of dynamic scheme was also found competitive for mme < but it was prone to false forecasts using the mmW.

6 Spatial application

In the previous section we have seen that in prognostic mode, mmW with static weights results in the least error previsions. In view of the operational evaluation, we now explore the spatial extension of the method. Specifically, using observed and modelled time-series at the station level rather than at the regional level, we test the spatial forecast skill of mme, mmW and mme < on a *blind* time-series. The approach is the following: using two-thirds of the JJA time series (training dataset) we train the mmW and mme < models and then we apply the weights learned into the remaining one-third of the records (test dataset). All the presented skill hereafter refers to the test dataset.

Skill. The composite skill of the selected ensemble products, originating from all blind forecasts at the 451 stations (aggregated) is presented in a Taylor plot (Fig. 12) together with the single determinist models. The benefits of ensemble treatment, either in the form of simple averaging models (mme) as well as using more sophisticated techniques (mme <, mmW) are clearly evident. Besides the error (RMSE), mme < and mmW also improve the correlation and the variance of the output with respect to mme. As seen in the cumulative density function plot, the improvement is reflecting the better capture of the 50 % of values outside the interquartile range, i.e. the lower than 25th and the higher than 75th percentile values.

The results are now spatially disaggregated and the latitudinal and longitudinal forecast skill of mme, mmW and mme < is shown in Fig. 13 for the gross error (RMSE) and the ability to capture the upper tail of the distribution, i.e. extremes (hit rate). The

ACPD
14, 15803–15865, 2014

weighted ensemble, in the form of mmW or mme <, significantly improves both indices over the ensemble mean. The advancement is happening at all single locations, as the cdf plot of the RMSE ratios with mme displays. The error is lowered by up to 35 % for mmW and 25 % for mme <. Half of the stations experience RMSE lowering in the mmW (mme <) case by up to 13 % (10 %) and the other half in the range 13 % to 35 % (10 % to 25 %). The tendency for larger improvement is found at the sites with the higher RMSE using the mme, except the Po valley area where the small improvement through the mmW and mme < indicate possibly inconsistencies of the deterministic models. The histogram of the errors from all stations for (mme, mme <, mmW) has a mean of (21.7, 19.6, 18.6) and a standard deviation of (5.8, 5.2, 4.6) implying that besides skill, forecast uncertainty also benefits from a similar improvement.

In view of the extremes, the correct identification of concentrations over the 120 $\mu g\, m^{-3}$ threshold value (right plot), has a clear latitude dependence in *mme* (the southern the better for ozone) that is considerably corrected in both mmW and mme <, with a more homogeneous pattern in mmW. The median hit rate of mme is 28 % and becomes 44 % in mme < and nearly doubles (52 %) in mmW. One quarter of the total stations laying at middle to high latitudes experience the highest improvement; a hit rate of less than 10 % in mme becomes up to 40 % in mmW and 30 % in mme <.

Weights. Figure 14 displays the mmW weight for each participating model at the observed sites (one figure per candidate model) for the training dataset. Although the optimization has been applied at each monitoring station individually, it can be inferred that the weighting pattern (per model) shows more of a coherent image across the continent, rather than a random design, reflecting a robust error covariance across the continent. On the opposite case, it can provide a mean for discriminating the performance of individual models. This spatial robustness of the weights is particularly important for the re-gridding of the results at locations not used in the training. The discontinuities observed at the spatial pattern of the weights at some sites are also present in the RMSE plot obtained with those weights (Fig. 13), indicating a probable error at the measurement site rather than the models. Last, the similarities seen earlier

ACPD
14, 15803–15865, 2014

in the zero-dimensional case between the mmW and mme< weights (frequency of model use) are also present in the spatial case.

Effective number of models. The calculation of M_{eff} (Bretherton et al., 1999) using the corr (e_i, e_j) or corr (d_i, d_j) matrices frames the bounds of the effective number of models. We find them to vary between 2 and 8, through a homogeneous spatial pattern (Fig. 15). Indeed, using analytical error minimization over all combinations (i.e. the one with the right trade-off between accuracy and diversity), M_{eff} covers all bins between 2 and 8, peaking at 3 to 4 members. The spatial variability is due to the absence of any filtering in the latter case. At half of the stations, evenly distributed across the domain, mme< uses only either 3 or 4 models, while over 80 % of the sites need 2–5 models from the pool.

We investigate now the statistical properties of the three ensemble products as a function of the M_{eff} calculated from the minimum error. The mean is well captured by all products (Fig. 16). It is decreasing for small M_{eff} (≤ 4) and remains roughly constant for higher values. This indicates that ensembles tend to be more symmetric at lower concentrations, pointing again that one of the areas where mme fails is extreme values, since only few models actually capture them. The latter statement is augmented from the Coefficient of Variation plot. It unfolds the differences in the statistical distribution of the three ensemble products. *The spread (range) of concentrations is monotonically decreasing as M_{eff} increases.* For $M_{\text{eff}} \leq 4$, this is due to equal reductions in mean and standard deviation, for $M_{\text{eff}} > 4$ it is due to decrease in standard deviation only (as CoV is decreasing but mean is stable). *The statistical distributions of three ensemble products start to converge for $M_{\text{eff}} > 6$, i.e. when the range of concentration is well bounded below* $120\,\mu g\,m^{-3}$. Finally, skewness and kurtosis do not demonstrate any significant dependence from M_{eff} (not shown).

The findings of the previous paragraph for the statistical distribution are explored hereafter for the skill with respect to M_{eff} (Fig. 16). The dissimilarities among the three ensemble products are clearly unfolded in all examined skill scores. The correlation (PCC) with observations is nearly independent of the M_{eff} for mme< and mmW. On

ACPD
14, 15803–15865, 2014

the other hand, mme has notably lower PCC for $M_{\text{eff}} \leq 4$, pointing again to the discrepancies in capturing the whole range of variability when there is a significant amount of extreme records (over 120 $\mu g\, m^{-3}$). Similar result is found for the standard deviation ratio (STDR). In terms of error (RMSE), it is a decreasing function of M_{eff} and the three ensemble products start to converge for $M_{\text{eff}} > 6$. As M_{eff} increases, the distribution of the models around the observations is gradually becoming more symmetric, hence the gain from mmW or mme< is minimized as the mme sample has already a quite symmetric distribution. Taken together with the distribution convergence seen in the previous paragraph, the results demonstrated that *the MME sample resembles the properties of an i.i.d. sample only for cases without extreme percentiles, since only few models are able to forecast them*. In turn, this points that *as long as the variance of some models departs significantly from the observed variance, the benefits from improvements in the ensemble skill in the form of mme< or mmW over mme become substantial*. Last, the improved hit rate (hitR) in mmW and mme< over mme seen in Fig. 13, has a coherent pattern across all M_{eff} values, as also seen in Fig. 16.

The contribution of different members as a function of M_{eff} is investigated (Fig. 17). For $M_{\text{eff}} \geq 6$, when the concentration range is more or less captured by all models, an even distribution of the selected models is found. Contrary, for the wide-most range ($M_{\text{eff}} < 4$), the contributions from specific models are particularly evident. Compared to the Taylor plot (Fig. 12), the models with the least frequency of selection had either lower STDR or higher RMSE. On the other hand, the most frequently used models belonged to a cluster of skilled contributors (low RMSE, high PCC, high STDR), however the selection among them took into account their joint contribution (error correlation) hence not all were selected.

Variance Inflation. According to the error decomposition presented in the introduction, bias correction has a net effect on the ensemble error. Further, the combination of models with variance close to the observed one, as well as the mixture of negatively correlated models are two other properties towards realistic ensemble representations. So far we dealt only with bias corrected simulations; here we present a plot of

ACPD
14, 15803–15865, 2014

the model's skill if we also correct the model's variance. As the purpose of this work is not the evaluation of the different correction strategies, we apply a simple multiplicative correction factor to the whole bias-corrected time series. The results are presented in Fig. 18 through a comparison of the composite skill (PCC, RMSE, STDR) in Taylor plots as well as through binned bias plots.

The skill of the numerical models in simulating ozone (1st column) is enhanced with the inclusion of the variance correction, which is also reflected in the ensemble products and in particular in mme and mme <. As expected, the second correction is also accompanied with an increase in the effective number of models as it yields more symmetric fields. The binned mean bias plot demonstrates that the ensemble products retain the same ability sequence in the two schemes across all ranges (i.e. 1st mmW, 2nd mme <, 3rd mme) with the known overestimation tendency for concentrations below 75 $\mu g\,m^{-3}$ and underestimation above that threshold. The differences between the schemes and products become substantial for the limited records exceeding the 180 $\mu g\,m^{-3}$ value. *In general, the mmW provides noteworthy better forecasts over mme < and mme even with fewer corrections (for example mmW with only bias correction scores better than mme < with bias and variance correction); this also applies for mme < over mme.*

For the other two pollutants (NO_2 and PM_{10}), some of the results seen in ozone are also valid like the improvement in the model's skill and the increase of the effective number of models. Compared to ozone simulations, the distance between the three ensemble products is lower in the Taylor plot indicating a mild improvement over mme. This is also confirmed through the analysis of the binned mean bias. In addition, the seasonality expressed through the PCC is lower in the case of NO_2 and PM_{10}. *This points out that mme < and mmW improve the skill of mme up to a point, further improvement requires an advancement of the core uncertainty factors inside the deterministic models like the emissions, the boundary conditions and the parameterization of physical processes*.

ACPD
14, 15803–15865, 2014

The gross improvement in the RMSE of the multi-model ensemble mean achieved through the bias and variance correction compared to only bias correction was 0.6 % for O_3, 2.1 % for NO_2 and 11.8 % for PM_{10}. On the other hand, the improvement in the RMSE achieved through the exploitation of the ensemble mean in the form of mmW or mme < was 8.6 % for O_3, 14.9 % for NO_2 and 13.5 % for PM_{10}. *Hence, even with bias and variance corrections on the models of an ensemble, superior improvements can be achieved through the optimization of an error decomposition approach*.

7 Summary and conclusions

Ensemble forecasting with multi model ensembles improves the forecast skill by reducing the non-linear error growth and averaging out individual models' error components. The *mme* (equal weights) is a spatiotemporal robust estimate of the actual state with increased accuracy (single errors cancel out) but with variance lower than the observations. Its skill degrades outside the interquartile range due to the inefficiency of the majority of the models to simulate extreme percentiles, where hence averaging brings mainly redundant information. The last property limits the usefulness of the ensemble mean, particularly for the study of extreme events, unless a mechanism that account for ensemble redundancy is taken into account. Possible pathways to eliminate this distortion and yield ensemble output with symmetric residuals across all distribution bins are:

- *mmW* (optimize error through model weighting; keep all models): for short-range forecasting (horizon < 4 days), it achieves lower error than the theoretical one for uncorrelated equally weighted ensemble. However, the variability of those weights at that scale is beyond any predictability. Nevertheless, learning over long time-periods (~ 2 months) and using those weights even at small time scales proved robust and accurate. Its skill outperformed all other ensemble products as well as individual models.

ACPD
14, 15803–15865, 2014

– *mme* < (optimize error through trade-off between accuracy and diversity; average on selected subset of models): for the 13 member ensemble, the effective number of models was in the range 2–8, with the peak between 3 and 4. Hundreds of ideal combinations exist within this range that can be detected through different algorithms applied on specific properties (e.g. error correlation, redundancy, diversity, negative correlation etc.) The learning algorithms work better with a direct optimization of the error rather than a dependent function of it (e.g., diversity). Its skill was significantly better over mme and individual models and it demonstrated the highest robustness with respect to the length of the training period.

An adaptive learning algorithm (weights or members) applied recursively can be seen as an application of a *flow-dependent error covariance*. Its usefulness was demonstrated for the case of mmW and mme <, where risky practices were also identified (e.g., mmW with dynamic weights). Compared to the traditional ensemble mean, the use of non-redundant sub-ensembles results in lower forecast uncertainty and increased skill for studies of the extremes. However, as the skill of even the best model is limited for very high values (e.g. $> 150\,\mu g\,m^{-3}\,O_3$), so does the skill of the ensemble products. Hence, besides any statistical post-treatment of the ensemble to coherently improve forecast skill, there is a need for continuous model improvement, especially for cases that depart from intermediate levels.

The skill of the three candidate ensemble products was also cross-validated as a function of the symmetry of the model distribution around the truth. With increasing symmetry, the three products were converging. Those cases were occurring for intermediate concentration ranges, that all models are somehow tuned to replicate. As the ensemble was becoming more assymetric, the error gain from mme < or mmW over mme was gradually increasing, reaching on average 15 % and 30 % respectively. All the extreme records were found in the assymetric era of the ensemble.

Hence, an ensemble may contain infinite number of models but the ideal ensemble should be constructed from this pool based on some criteria that reflect a symmetrical error distribution. The multi model mean defines the threshold point from where the

ACPD
14, 15803–15865, 2014

non-trivial problem of weighting or sub-selecting should progress further. A general roadmap of good practices is attempted hereafter:

1. Generate a raw ensemble.
2. Apply *bias correction* techniques to remove systematic errors.
3. Optimize distribution symmetry over a *training set* of *proper size* using either all members or a subset of them. The first approach concludes with a *weighting scheme,* the second with the identification of the *effective number of models* and the allowed/forbidden *combinations of members* that can be sampled to constitute effective ensembles. The extent of the training dataset is confined from physical concepts as well as the statistical properties of the specific ensemble.
4. Average the weighted or reduced ensemble.

The above procedure does not imply any spatial or cross-variate dependence. It aims at optimizing ensemble averaging at single locations for single variables. A framework for the optimization of the ensemble skill for multivariate spatial dependence, like the multi-dimensional optimization (Potempski and Galmarini, 2009) or the ensemble-copula coupling (Schefzik et al., 2013), will be assessed in a future study.

Acknowledgements. The authors wish to thank Efisio Solazzo for his constructive comments on the manuscript.

References

AMS (American Meteorological Society): Enhancing weather information with probability forecasts, B. Am. Meteorol. Soc., 83, 450–452, 2002.

Bishop, C. M.: Neural Networks for Pattern Recognition, Oxford University Press, New York, NY, USA, 1995.

ACPD
14, 15803–15865, 2014

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

Brown, G., Wyatt, J., Harris, R., and Yao, X.: Diversity creation methods: a survey and categorisation, Journal of Information Fusion, 6, 5–20, 2005.

Dosio, A. and Paruolo, P.: Bias correction of the ENSEMBLES high-resolution climate change projections for use by impact models: evaluation on the present climate, J. Geophys. Res.-Atmos., 116, D16106, doi:10.1029/2011jd015934, 2011.

Errico, R.: What is an adjoint model?, B. Am. Meteorol. Soc., 78, 2577–2591, 1997.

Fern, X. Z. and Brodley, C. E.: Solving cluster ensemble problems by bipartite graph partitioning, in: Proceedings of 21th International Conference on Machine Learning (ICML2004), TBanff, Alberta, Canada, 4–8 July 2004, ACM Press, 281–288, doi:10.1145/1015330.1015414, 2004

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J. C., Bellasio, R., Bompay, F., Buckley, R., Bouzom, M., Champion, H., D'Amours, R., Davakis, E., Eleveld, H., Geertsema, G. T., Glaab, H., Kollax, M., Ilvonen, M., Manning, A., Pechinger, U., Persson, C., Polreich, E., Potemski, S., Prodanova, M., Saltbones, J., Slaper, H., Sofiev, M. A., Syrakov, D., Sørensen, J. H., L.Van der Auwera, Valkama, I., and Zelazny, R.: Ensemble dispersion forecasting – Part I: concept, approach and indicators, Atmos. Environ., 38, 4607–4617, 2004.

Galmarini, S., Rao, S. T., and Steyn, D. G.: Preface, Atmos. Environ., 53, 1–3, 2012a.

Galmarini, S., Bianconi, R., Appel, W., Solazzo, E., Mosca, S., Grossi, P., Moran, M., Schere, K., and Rao, S. T.: ENSEMBLE and AMET: two systems and approaches to a harmonized, simplified and efficient facility for air quality models development and evaluation, Atmos. Environ., 53, 51–59, 2012b.

Galmarini, S., Kioutsioukis, I., and Solazzo, E.: *E pluribus unum**: ensemble air quality predictions, Atmos. Chem. Phys., 13, 7153–7182, doi:10.5194/acp-13-7153-2013, 2013.

Geman, S., Bienenstock, E., and Doursat., R.: Neural networks and the bias/variance dilemma, Neural Comput., 4, 1–58, 1992.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T.: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, Mon. Weather Rev., 133, 1098–1118, 2005.

ACPD
14, 15803–15865, 2014

Helton, J. C. and Davis, F. J.: Latin hypercube sampling and the propagation of uncertainty in analyses of complex systems, Reliab. Eng. Syst. Safe., 81, 23–69, 2003.
Iman, R. L. and Conover, W. J.: A distribution-free approach to inducing rank correlation among input variables, Commun. Stat. B-Simul., 11, 311–334, 1982.
Kalnay, E.: Atmospheric Modelling, Data Assimilation and Predictability, Cambridge University Press, New York, 341 pp., 2003.
Knutti, R., Abramowitz, G., Collins, M., Eyring, V., Gleckler, P. J., Hewitson, B., and Mearns, L.: Good practice guidance paper on assessing and combining multi model climate projections, in: Meeting Report of the Intergovernmental Panel on Climate Change Expert Meeting on Assessing and Combining Multi Model Climate Projections, Boulder, Colorado, USA, 25–27 January 2010, edited by: Stocker, T. F., Qin, D., Plattner, G.-K., Tignor, M., and Midgley, P. M., IPCC Working Group I Technical Support Unit, University of Bern, Bern, Switzerland, 13 pp., 2010.
Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., and Surendran, S.: Improved weather and seasonal climate forecasts from multi-model superensemble, Science, 285, 1548–1550, 1999.
Krogh, A. and Vedelsby, J.: Neural network ensembles, cross validation, and active learning, Adv. Neur. In., 7, 231–238, 1995.
Kuncheva, L. and Whitaker, C.: Measures of diversity in classifier ensembles, Mach. Learn., 51, 181–207, 2003.
Leith, C. E.: Theoretical skill of Monte Carlo forecasts, Mon. Weather Rev., 102, 409–418, 1974.
Lin, M., Tang, K., and Yao, X.: Selective negative correlation learning algorithm for incremental learning, in: Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN'08), Hongkong, China, 2526–2531, 1–6 June 2008.
Liu, Y. and Yao, X.: Ensemble learning via negative correlation, Neural Networks, 12, 1399–1404, 1999.
Malamud, B. D. and Turcotte, D. L.: Self-affine time series: measures of weak and strong persistence, J. Stat. Plan. Infer., 80, 173–196, 1999.
Mallet, V. and Sportisse, B.: Uncertainty in a chemistry-transport model due to physical parameterizations and numerical approximations: an ensemble approach applied to ozone modelling, J. Geophys. Res., 111, D01302, doi:10.1029/2005JD006149, 2006.
Markowitz, H.: Portfolio selection, J. Financ., 7, 77–91, 1952.

ACPD
14, 15803–15865, 2014

McKay, M. D., Beckman, R. J., and Conover, W. J.: A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics, 21, 239–245, 1979.
Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The new ECMWF ensemble prediction system: methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, 1996.
Potempski, S. and Galmarini, S.: *Est modus in rebus*: analytical properties of multi-model ensembles, Atmos. Chem. Phys., 9, 9471–9489, doi:10.5194/acp-9-9471-2009, 2009.
Pouliot, G., Pierce, T., Denier van der Gon, H., Schaap, M., Moran, M., and Nopmongcol, U.: Comparing emissions inventories and model-ready emissions datasets between Europe and North America for the AQMEII Project, Atmos. Environ., 53, 4–14, 2012.
Rao, S. T., Galmarini, S., and Puckett, K.: Air quality model evaluation international initiative (AQMEII): advancing the state of the science in regional photochemical modelling and its applications, B. Am. Meteorol. Soc., 92, 23–30, 2011.
Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., and Potempski, S.: On the systematic reduction of data complexity in multi-model ensemble atmospheric dispersion modelling, J. Geophys. Res., 117, D05314, doi:10.1029/2011JD016503, 2012.
Schefzik, R., Thorarinsdottir, T., and Gneiting., T.: Uncertainty quantification in complex simulation models using ensemble copula coupling, Stat. Sci., 28, 616–640, 2013.
Schere, K., Flemming, J., Vautard, R., Chemel, C., Colette, A., Hogrefe, C., Bessagnet, B., Meleux, F., Mathur, R., Roselle, S., Hu, R.-M., Sokhi, R. S., Rao, S. T., and Galmarini, S.: Trace gas/aerosol boundary concentrations and their impacts on continental-scale AQMEII modeling domains, Atmos. Environ., 53, 38–50, 2012.
Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Denier van der Gon, H., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jeričević, A., Kraljević, L., Miranda, A. I., Nopmongcol, U., Pirovano, G., Prank, M., Riccio, A., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R., Yarwood, G., Zhang, J., Rao, S., and Galmarini, S.: Model evaluation and ensemble modelling of surface-level ozone in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 60–74, 2012a.
Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Wyat Appel, K., Bessagnet, B., Brandt, J., Christensen, J. H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Miranda, A. I., Nopmongcol, U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., Wolke, R.,

ACPD
14, 15803–15865, 2014

Yarwood, G., Zhang, J., Rao, S. T., and Galmarini, S.: Operational model evaluation for particulate matter in Europe and North America in the context of AQMEII, Atmos. Environ., 53, 75–92, 2012b.
Solazzo, E., Riccio, A., Kioutsioukis, I., and Galmarini, S.: Pauci ex tanto numero: reduce redundancy in multi-model ensembles, Atmos. Chem. Phys., 13, 8315–8333, doi:10.5194/acp-13-8315-2013, 2013.
Stein, M.: Large sample properties of simulations using latin hypercube sampling, Technometrics, 29, 143–151, 1987.
Tracton, M. S. and Kalnay, E.: Operational ensemble prediction at the National Meteorological Center: practical aspects, Weather Forecast., 8, 379–398, 1993.
Ueda, N. and Nakano., R.: Generalization error of ensemble estimators, in: Proceedings of International Conference on Neural, Networks, 90–95, Washington, DC, USA, 3–6 June 1996.
Varotsos, C., Efstathiou, M., Tzanis, C., and Deligiorgi, D.: On the limits of the air pollution predictability; the case of the surface ozone at Athens, Greece, Environ. Sci. Pollut. R., 19, 295–300, 2012.
Weigel, A., Knutti, R., Liniger, M., and Appenzeller, C.: Risks of model weighting in multimodel climate projections, J. Climate, 23, 4175–4191, 2010.
Zanda, M., Brown, G., Fumera, G., and Roli, F.: Ensemble learning in linearly combined classifiers via negative correlation, Lect. Notes Comput. Sci., 4472, 440–449, 2007.
Zurbenko, I. G.: The Spectral Analysis of Time Series, North-Holland, Amsterdam, 236 pp., 1986.

ACPD
14, 15803–15865, 2014

Table 1. Analytical formulas for the 1-dimensional case (single-point optimization).

	Uncorrelated models	Correlated models
Optimal Weights	$a_k = \frac{\frac{1}{\sigma_k^2}}{\sum_j \frac{1}{\sigma_j^2}}$	$\overline{a} = \frac{\mathbf{K}^{-1}I}{(\mathbf{K}^{-1}I,I)}$
Limits for mme (ensemble mean)	$\mathrm{MSE}\left(\overline{x}\right) \leq \mathrm{MSE}(x_1) \leq \ldots \leq \mathrm{MSE}(x_m)$ if $\frac{\mathrm{MSE}(x_m)}{\mathrm{MSE}(x_1)} \leq m+1$	$\mathrm{MSE}\left(\overline{x}\right) \leq s_1 \leq s_2 \leq \ldots \leq s_m$ if $\frac{s_m}{s_1} \leq m$
Definitions	σ_j^2 = variance of model's j error	$\mathbf{K}$ = error covariance matrix $I = [1, 1, \ldots, 1]^T$ s_j = eigenvalues of $\mathbf{K}$

ACPD
14, 15803–15865, 2014

Table 2. Indices of skill and redundancy.

Pearson Correlation Coefficient (PCC)	$\mathrm{PCC} = \frac{\frac{1}{N}\sum_i \left(f_i - \overline{f_i}\right)\left(\mu_i - \overline{\mu}\right)}{\sigma_{f_i}\sigma_{\mu}}$
Mean Bias (MB)	$\mathrm{MB} = \frac{\sum(f_i - \mu_i)}{N}$
Root Mean Square Error (RMSE)	$\mathrm{RMSE} = \sqrt{\frac{\sum(f_i - \mu_i)^2}{N}}$
e_m and d_m indicators	$d_m = e_m^* - R \cdot \mathrm{MME}^*$ $e_{\mathrm{i,\,m}} = \frac{f_{\mathrm{i,\,m}} - \mu_i}{\sigma_i}$
Threshold indices based on a contingency table for events (Forecasted/Observed)	Hit rate $(H) = \frac{Y/Y}{Y/Y + N/Y}$
Accuracy term (acc)	$\mathrm{acc} = E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i - \mu)^2\right)$
Diversity term (div)	$\mathrm{div} = E\left(\frac{1}{M}\sum_{i=1}^{M}\left(f_i - \overline{f}\right)^2\right)$
Squared Bias (bias)	$\overline{\mathrm{bias}}^2 = E\left(\frac{\sum(f_i - \mu_i)}{N}\right)^2$
Variance of errors (varE)	$\overline{\mathrm{varE}} = E(\mathrm{var}(f - \mu))$
Covariance of errors (covE)	$\overline{\mathrm{covE}} = E(\mathrm{cov}(f - \mu))$
Correlation of errors (pccE)	$\overline{\mathrm{pccE}} = E(\mathrm{pcc}(f - \mu))$

ACPD
14, 15803–15865, 2014

Table 3. The mean MSE of the 92 daily cases, in hindcast (H) and forecast (F) mode, with either static (s) or dynamic (d) weights for five selected ensemble products (mme, dvMX1,covMN1, mmW, mme <). The comparison has been applied to four European sub-regions. The cases with MSE lower than mme are given in bold.

	EU1r			EU2r			EU3r			EU4r		
	H	F(d)	F(s)	H	F(d)	F(s)	H	F(d)	F(s)	H	F(d)	F(s)
mme	48.3	48.7	48.7	23.6	23.8	23.8	276.2	277.6	277.6	44.8	44.8	44.8
divMX1	**43.**	**42.9**	**44.6**	30.4	31.	30.4	**238.9**	**238.4**	**246.4**	**43.7**	**43.6**	**43.4**
covMN1	**23.9**	**32.2**	**35.4**	**23.0**	27.6	26.9	**161.5**	**198.6**	**206.6**	**27.6**	**35.7**	**34.5**
mmW	**0.6**	48.8	**15.0**	**0.5**	33.1	**12.3**	**3.4**	358.6	**103.5**	**1.1**	129.6	**22.9**
mme <	**9.4**	**21.8**	**18.5**	**7.6**	**17.6**	**21.4**	**115.3**	**171.**	**146.7**	**11.0**	**29.6**	**24.8**

ACPD
14, 15803–15865, 2014

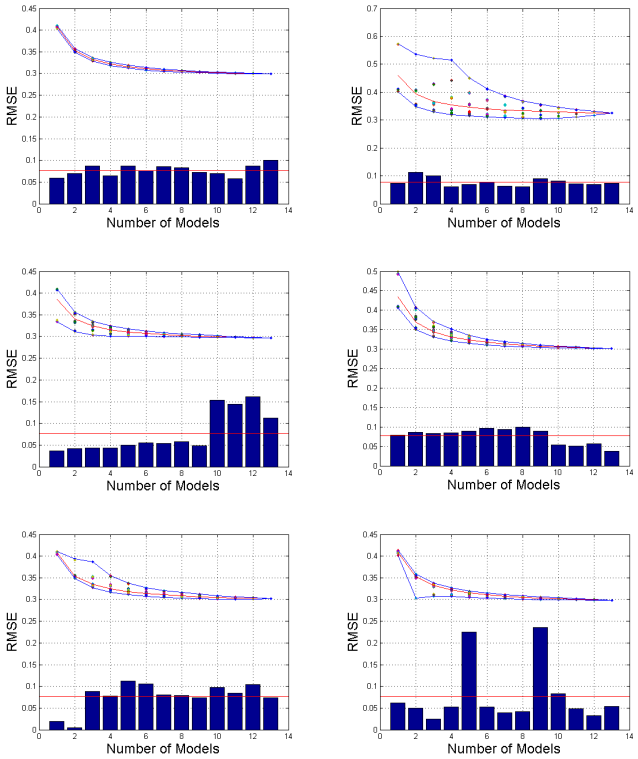
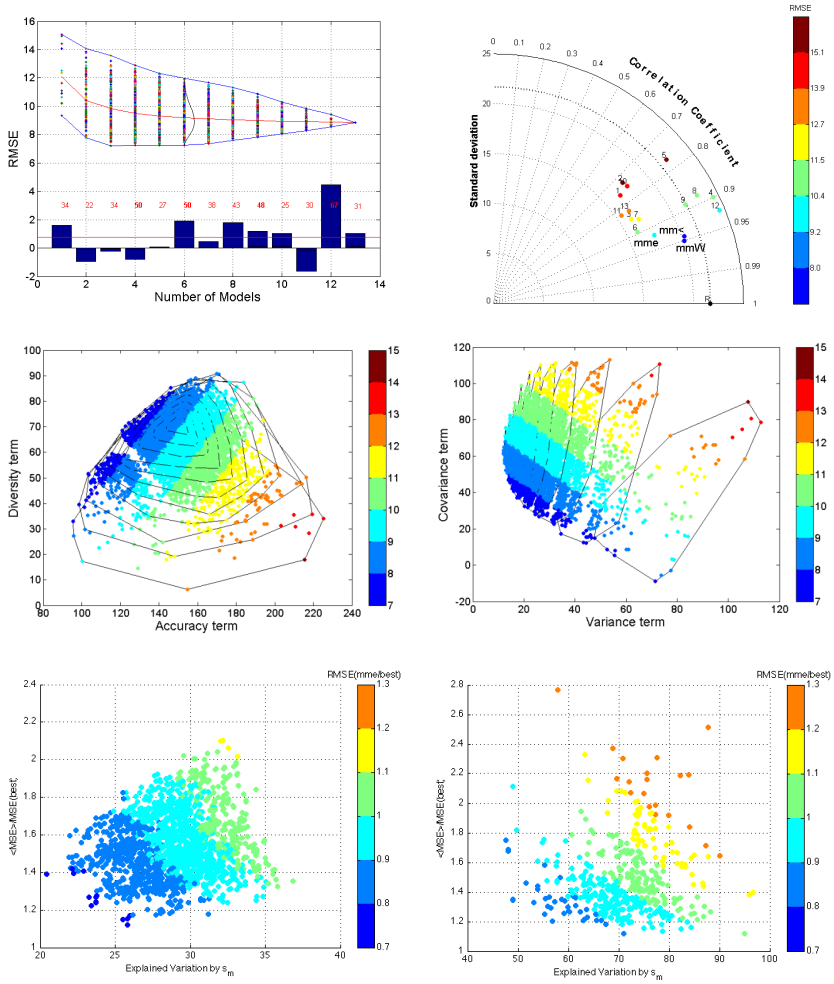


Figure 1. Ensemble error (RMSE) from all possible combinations of candidate models. The red curve on each plot represents the mean of the distribution of any k-model combinations while the blue curves form the min and max of the each respective distribution. [top left] i.i.d., [top right] bias perturbation, [middle] variance perturbation, [bottom] covariance perturbation. Please read text for explanations. At the same plot, the bar chart expresses the optimal weight of each model in the full ensemble and the straight red line symbolizes the equal weight value. In this case, the horizontal axis represents the id of the model.

ACPD
14, 15803–15865, 2014



ACPD
14, 15803–15865, 2014

De praeceptis ferendis

I. Kioutsioukis and S. Galmarini

Figure 2. [top left] Ensemble error (RMSE) from all possible combinations of candidate models (EU4r). The notation is similar to Fig. 1. The numbers in red express the fractional contribution of each model to skilled combinations. [top right] Multi aspects of individual model skill through Taylor plot. [middle] Four dimensional representation of accuracy – diversity (left) and variance – covariance (right), with respect to RMSE (color scale) and ensemble order (isolines). The isolines represent the multi-dimensional convex hull as a function of ensemble order. Isolines shrink with increasing ensemble order [bottom] The RMSE ratio of mme over the best single model as a function of redundancy (explained variation by sm) and model skill difference ($\langle \mathrm{MSE} \rangle / \mathrm{MSE}$ (best), evaluated from all combinations of 6th order (left) and 13th order (right). The diagram on the right has been evaluated at all observation sites.

ACPD
14, 15803–15865, 2014

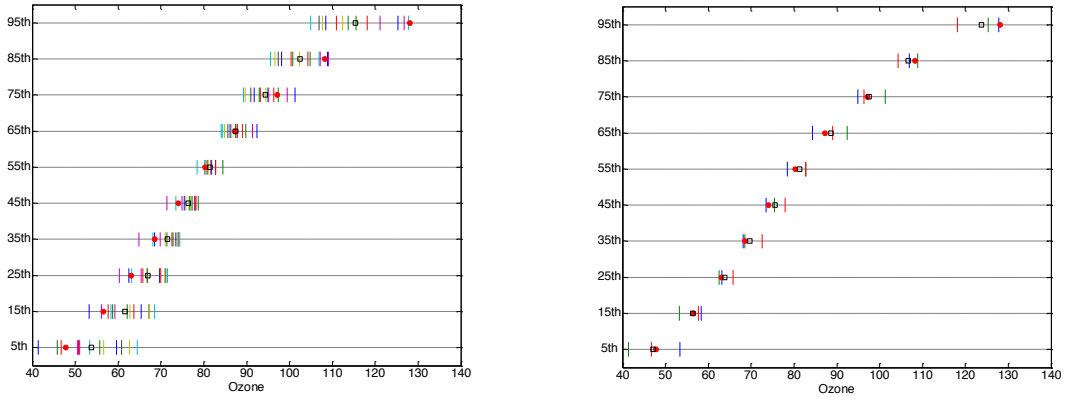
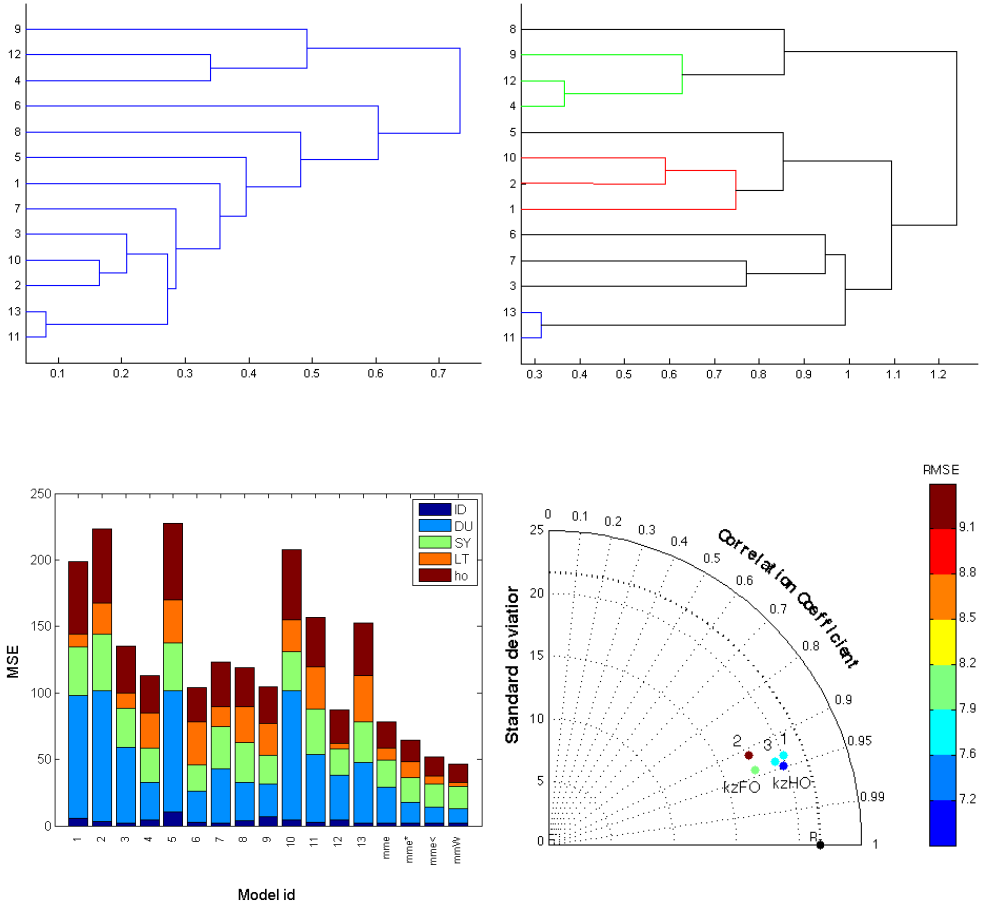


Figure 3. Cumulative density function of observations (red circle) and models (coloured lines). The ensemble mean is displayed with a square. Full ensemble (left) vs. optimal combination (right).

ACPD
14, 15803–15865, 2014



ACPD
14, 15803–15865, 2014

Figure 4. [Top] The potential use of various indices as proxies for clustering low error combinations, based on corr (e_i, e_j) dendogram (left) and corr (d_i, d_j) dendrogram (right). [Middle] Process contribution to model's error: the MSE of the deterministic models (1–13) and the ensemble products as a function of physical processes, decomposed with two variants of the Kolmogorov-Zurbenko filter. The left plot utilizes four spectral components (ID, DU, SY, LT) and the right plot only two (ID+DU, SY+LT). The latter case has less energy leak between the components, as quantified by all the higher-order interactions (ho term). The mme* is the mme averaged over the subset without significant systematic errors. [Bottom] Skill of selected statistical ensemble products for $M_{\text{eff}} = 5$ by means of a Taylor plot: (1) em dendrogram cluster [5, 6, 8, 9, 12], (2) dm dendrogram cluster [5, 6, 7, 8, 13] (3) kz filter removal (mme*) [3, 4, 6, 7, 8, 9, 12]. In addition, kzFO is the first order kz model [ID: 7, DU: 6, SY: 12, LT: 12] and kzHO is the higher order kz model [ID+DU: 4, 6, 9, 11, 12; SY+LT: 1, 3, 6, 10, 12].

ACPD
14, 15803–15865, 2014

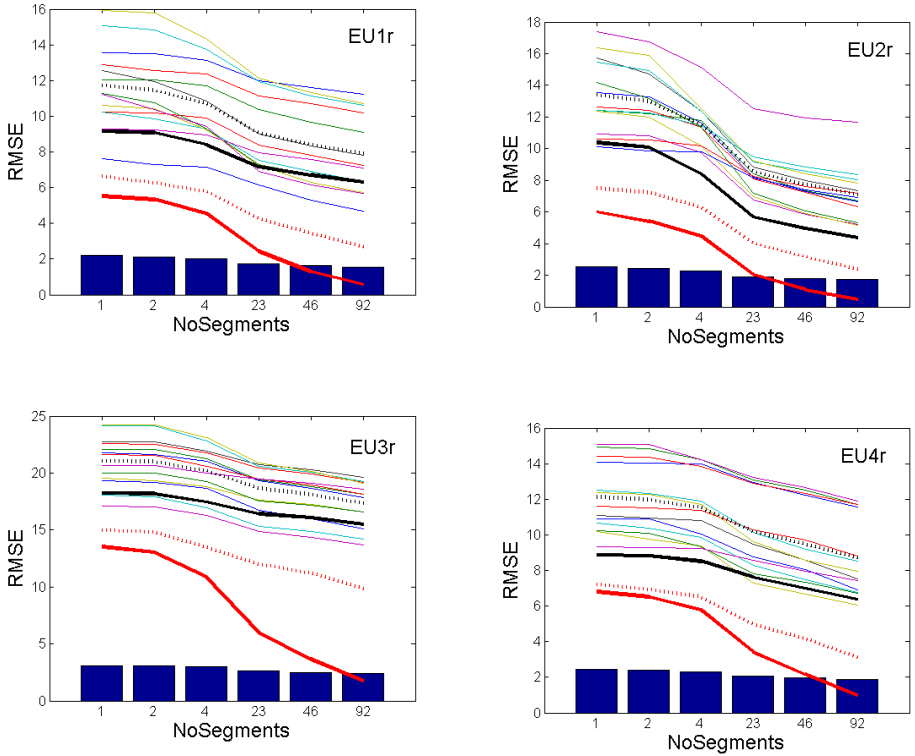


Figure 5. The mean RMSE of the models (colored lines) as a function of window size (1–92 days). In addition, selected ensemble products are also displayed: mme (thick black), $\langle mm_i \rangle$ (thick dotted-black), mmW (thick red), mme < (thick dotted red). The bars show the theoretical minimum value (< var >/nm) for uncorrelated models. [Bottom] The link between the two decompositions for ensembles of order M (= 13). The colorbar reflects the RMSE ratio (mme/best).

ACPD
14, 15803–15865, 2014

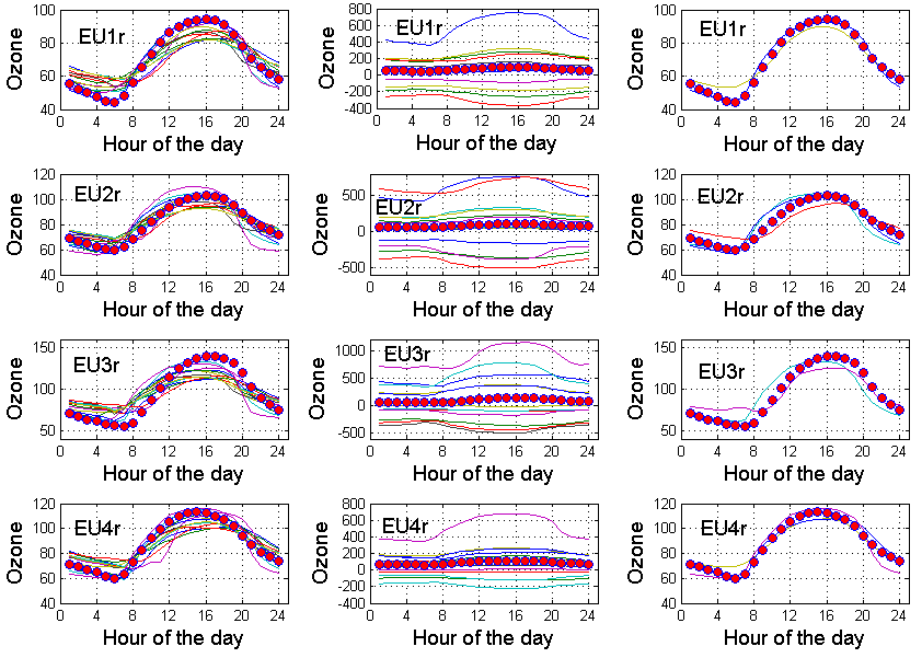


Figure 6. [left column] Mean diurnal ozone cycles for JJA of the ensemble members (lines). Observations are presented with filled circles. [middle column] Distribution of models around the observations using optimal weights (mmW). [right column] Distribution of models around the observations using the combination that optimizes the accuracy-diversity trade-off (mme <).

ACPD
14, 15803–15865, 2014

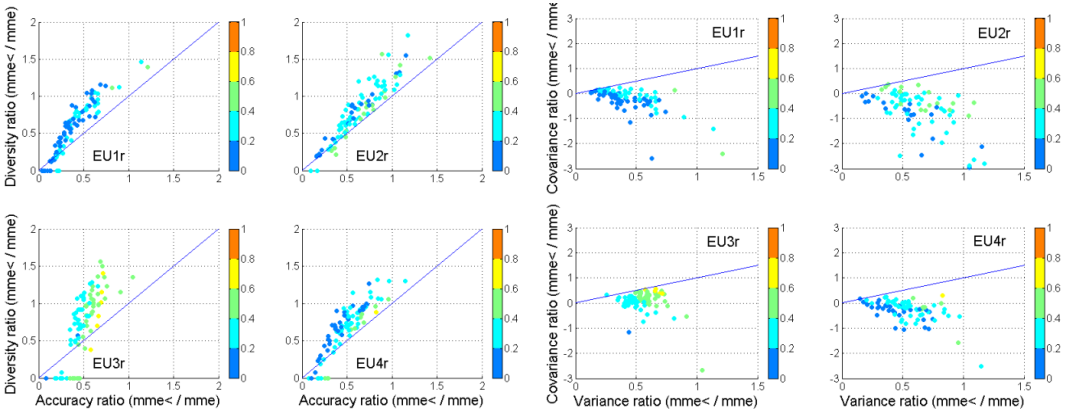


Figure 7. Comparison between mme < and mme with respect to the error decomposition. Each of the 92 dots corresponds to an individual 1 day simulation. The color scale represents the RMSE ratio. All ratios have been calculated as *property*(mme <)/*property*(mme). [left] Ratios of accuracy vs. ratios of diversity. [right] Ratios of variance vs. ratios of covariance.

ACPD
14, 15803–15865, 2014

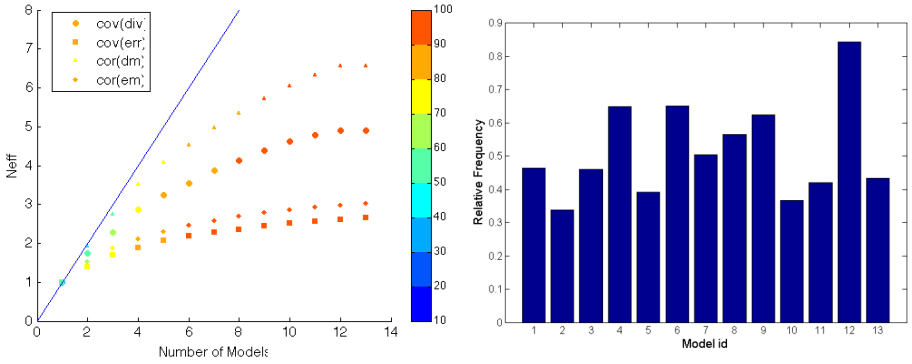


Figure 8. Effective number of models calculated through an eigenvalue formula. The color scale corresponds to the explained variance (left). Relative frequency of each model's participation in combinations with error lower than the mme (right).

ACPD
14, 15803–15865, 2014

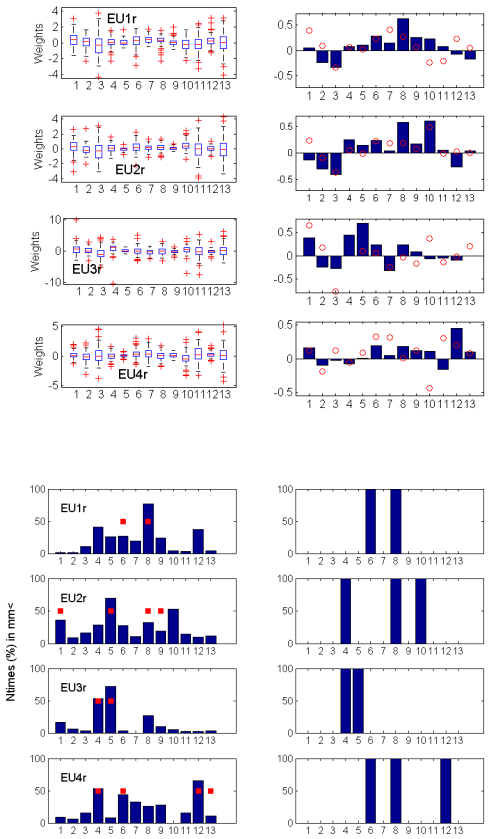


Figure 9. Weight Sensitivity for mmW (top) and mme < (bottom), for the 92 daily segments (left) and the seasonal 92-days segment (right). For mmW, the boxplots present the weights over all examined daily cases (day-to-day variability) while the barplots show the weight over the one seasonal case (red circles indicate the mean value of the weights derived from all daily cases). For mme <, the barplots show each model's frequency of participation in the optimal sub-ensemble.

ACPD
14, 15803–15865, 2014

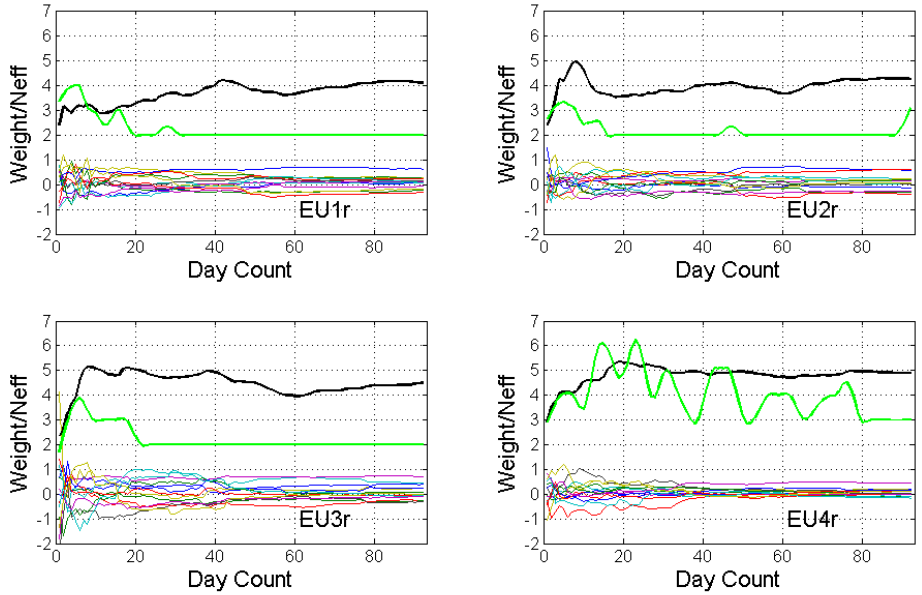


Figure 10. Variability of weights (thin lines) and effective number of models (thick lines) as a function of time-series length. Each thin-line represents a different model. The effective number of models is calculated through eigenanalysis (black) or error minimization (green).

ACPD
14, 15803–15865, 2014

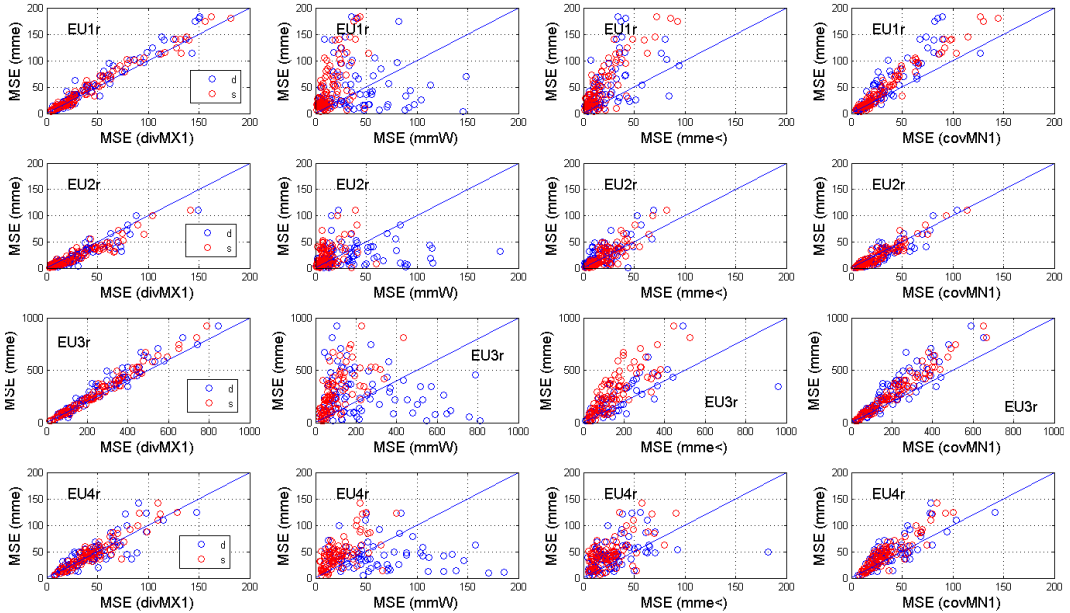


Figure 11. The MSE of the examined ensemble products (divMX1, mmW, mme <, covMN1) vs. the mme, for the 92 cases of 1 day blocks. Blue for the dynamic weights, red for the static. From top to bottom, EU1r to EU4r.

ACPD
14, 15803–15865, 2014

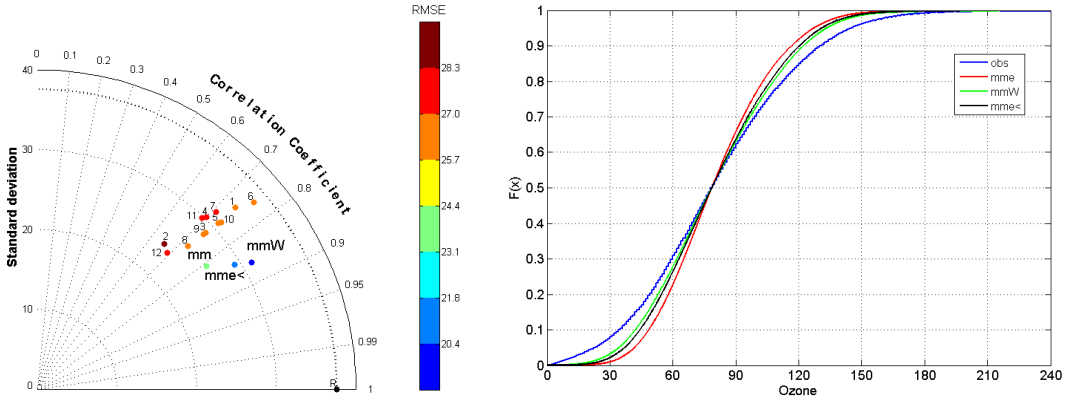


Figure 12. Aggregation of the spatial results from the 451 stations for the test dataset. [top left] The skill of mme, mme < and mmW in a Taylot plot together with the deterministic models. [top right] The cdfs of mme, mme < and mmW alongside the obs.

ACPD
14, 15803–15865, 2014

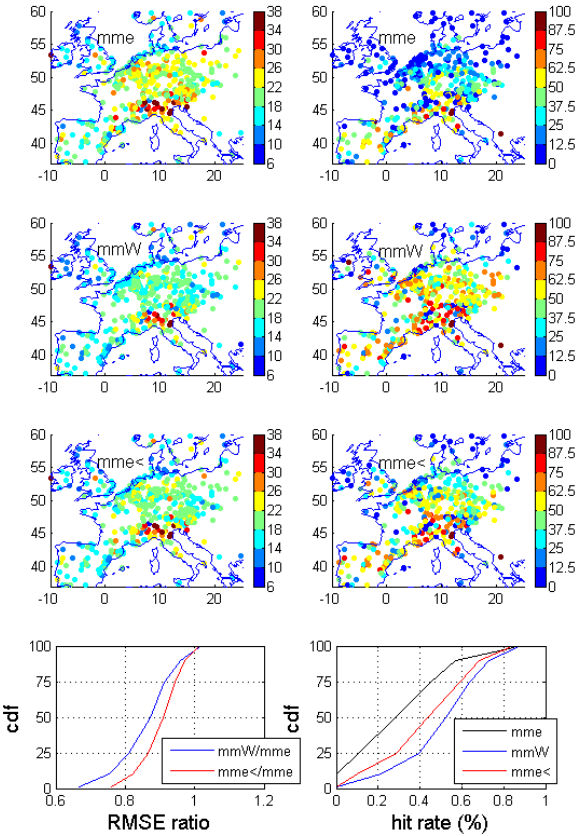


Figure 13. [Top row] The RMSE of ozone at each observed site for mme (left). The behavior at the upper tail of the distribution; percentage of correct hits for events > 120 µg m^{-3} for mme (right) [2nd, 3rd row] Like top row but for mmW and mme <. [Bottom row] The cdf of each spatial plot.

ACPD
14, 15803–15865, 2014

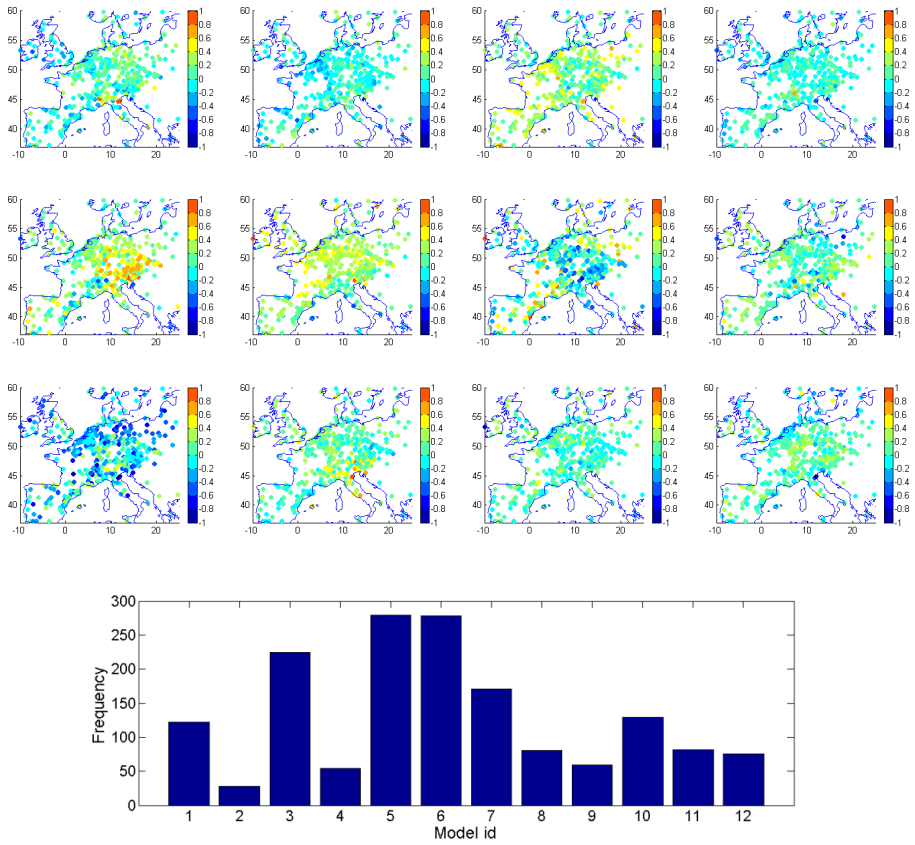


Figure 14. Ozone spatial weights (mmW) calculated for each model (1–12) for JJA (1 segment) at the observed rural sites (segment de-bias) and aggregated frequency of model use in mme <, from all the 451 stations for the test dataset.

ACPD
14, 15803–15865, 2014

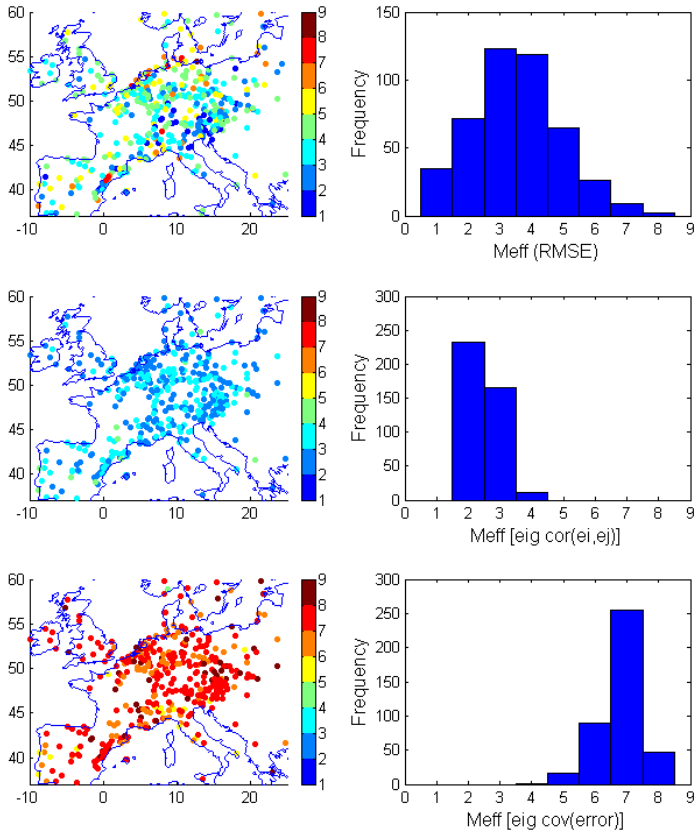


Figure 15. [Top] Spatial distribution of M_{eff} based on minimum error combination (left) and its histogram (right). [Middle] like top but for M_{eff} based on the eigenvalues of the corr (e_i, e_j) matrix. [Bottom] like top but for M_{eff} based on the eigenvalues of the covariance of the diversity matrix.

ACPD
14, 15803–15865, 2014

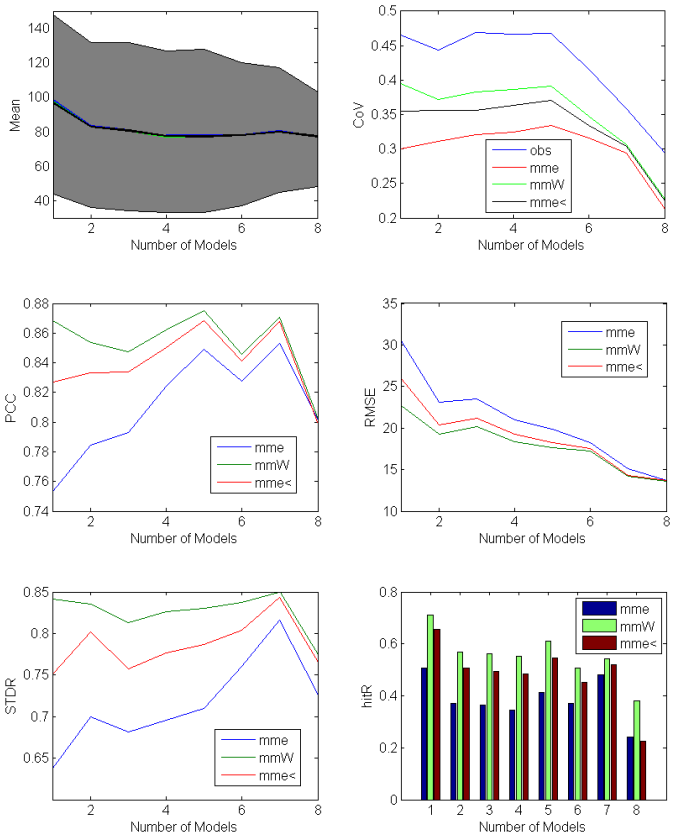


Figure 16. [Top] Statistical properties of mme, mme < and mmW forecasts vs. obs from the 451 stations for the test dataset as a function of M_{eff}: mean and coefficient of variation (StandardDeviation/Mean). The shadow area in the mean plot shows the 10th and 90th percentile of the observed concentrations. [Middle, Bottom] Forecast Skill of mme, mme < and mmW from the 451 stations for the test dataset as a function of M_{eff}: PCC, RMSE, STDR and hitR.

ACPD
14, 15803–15865, 2014

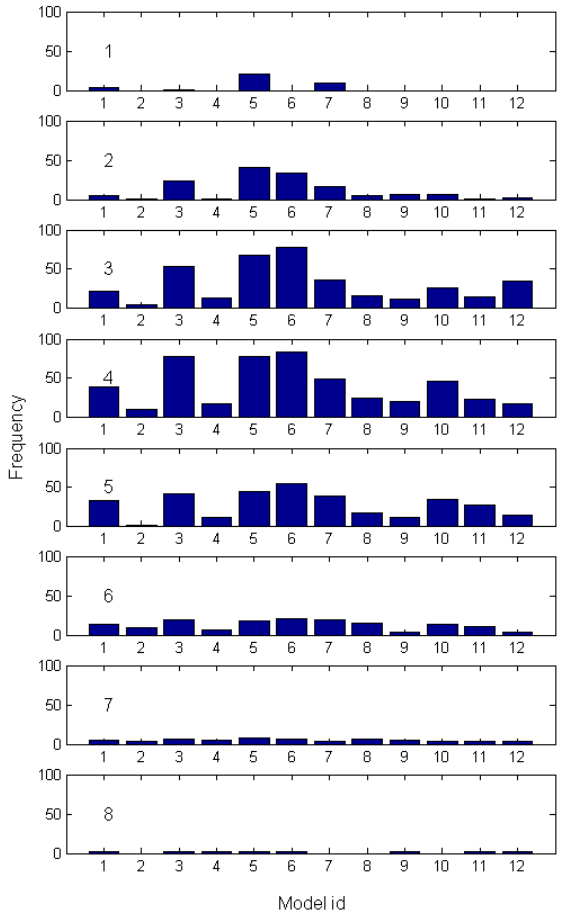
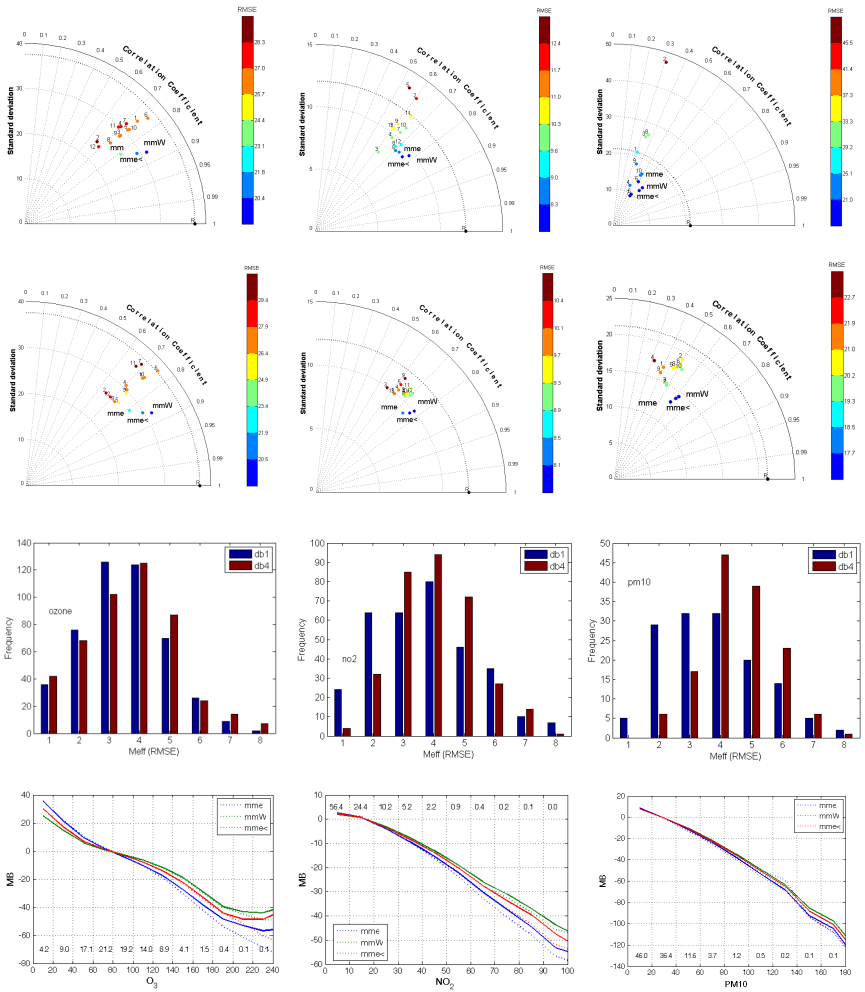


Figure 17. Frequency of model use in mme <, from the 451 stations for the test dataset as a function of M_{eff}.

ACPD
14, 15803–15865, 2014



ACPD
14, 15803–15865, 2014

I. Kioutsioukis and S. Galmarini

Figure 18. The Taylor diagrams on the 1st row refer to only bias correction (db1) while on the 2nd row refer to bias plus variance correction (db4). The bar plot on the 3rd row show the distribution of the effective number of models in the two schemes. The line plots at the last row compare the binned bias of the two correction schemes (db1: dotted, db4: line); the percentage of values within each bin is also given. Each column shows a different pollutant (O_3, NO_2, PM_{10}). The plots have been produced from the aggregated time series incorporating all the stations of the test dataset.

ACPD
14, 15803–15865, 2014