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

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

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. 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 these 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*. The sophisticated ensemble averaging techniques, following proper training, were shown to have higher skill across all distribution bins compared to solely ensemble averaging forecasts.

Keywords: *AQMEII, multi-model ensembles, error decomposition, model weighting, forecast skill, air quality*

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).

Deterministic approaches are fast but they rely on the validity of the linearized approximation of error growth (Errico, 1997). The availability of increasingly powerful computing in recent years has boosted the feasibility and use 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 3D 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 statistical methods 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 uncertainty in the initial conditions. *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 conditions, the deterministic forecast generated from the MME mean is better than any single-model forecast due to the averaging out of the errors as well as the better sampling of the input uncertainty (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 error 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 subject to 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 example the analysis of extremes less reliable. 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., Gneiting et al., 2005; 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 only. Seen from another angle, similar to diversity, ensembles also tend to produce better results when they contain negatively correlated models[1]. Ideally, the most accurate ensemble consists of members that are distributed randomly around the observations (i.e. unbiased and uncorrelated). This ‘randomness’ in the model outputs of an ensemble is not a pragmatic condition. Nevertheless, an optimal ensemble can be constructed *a posteriori* by inducing this property in the members.

In this work, we attempt to give an overview of the critical elements in deterministic forecasting with ensembles, with particular focus on the ensemble built from regional air quality models within the Air Quality Modelling Evaluation International Initiative (AQMEII). The overall goal of the study is to highlight the properties, through model selection or weighting, that guarantee a symmetric distribution of errors and eventually produce a single improved forecast out of an ensemble. Starting from a presentation of the available mathematical framework, many important aspects of ensemble forecasting are

[1] It will be demonstrated later in the article

demonstrated using synthetic and real time-series. Our motivation is to depict some best practices for deterministic forecasting with air quality ensembles.

The paper is structured as follows: in section 2, theoretical evidence on multi-model ensembles is presented together with an example that serves to show the contributing factors to the ensemble error. In section 3, we present the data and the methodology. In section 4 we decompose and analyse the ensemble error and its properties using spatially-aggregated AQMEII data. In section 5 we apply the results obtained in the previous section into forecasting at all monitoring stations (continental scale). Conclusions are drawn in section 6.

2. Theoretical Considerations

The aim of this section is to outline the documented mathematical evidence towards the reduction of the ensemble error. The notation used throughout the text is summarized in Table **1**.

2.1. The bias-variance-covariance decomposition of the ensemble error

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

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

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 Equation 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:

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

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

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

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

Equation (2) is valid for uniform ensembles, i.e. $w_i=\frac{1}{M}$. The terms $\overline{bias}$ and $\overline{varE}$ are the average bias and variance of the ensemble members error (modelled time-series minus observed time-series) respectively while the new term $\overline{covE}$ is the average covariance between pairs of distinct ensemble members error. From Equations (2) follows:

- The more ensemble members we have, the closer is $Var[\bar{f}-\mu]$ to $\overline{covE}$;
- $\overline{bias}^2$ and $\overline{varE}$ are positive defined, but $\overline{covE}$ 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. Given the positive nature of the bias and variance terms and the decreasing importance of the variance term as we include more members, the minimization of the quadratic ensemble error ideally suggests unbiased (or bias-corrected) members with low error correlation amongst them (to lower the covariance term).

2.2. The accuracy-diversity decomposition of the ensemble 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*:

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

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 ensemble, on a particular pattern. But, given that we have no criterion for identifying *a priori* that best individual (i.e. which ensemble member will best match the observations at future time-steps), 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 last statement is not self-evident for non-random sampling of the best member (e.g. conditioned to past errors from the models).

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; it is necessary 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):

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

The Ω term (Brown et al., 2005) 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.

2.3. The analytical optimization of the ensemble error

The two decompositions presented 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 some 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 2. Also, whether correlated or not, the models are assumed as random variables. The optimal ensemble corresponds 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 (Table 2). For uncorrelated models, the only constraint is the skill difference (MSE ratio) of the worst over the best single model. For example, the arithmetic mean of a 3-member ensemble 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 (Potempski and Galmarini, 2009; Weigel et al., 2010) and besides skill difference, they also depend on error correlation measures. Further, unlike the case of uncorrelated models, optimal weights for correlated models can be negative (Table 2). There is no physical interpretation for the negative weights; if they arise for some models, it is simply a result of the optimization of the cancelling out of the individual errors. For example, models with highly correlated errors may be given weights of opposite sign.

2.4. Example

We now present a theoretical example aimed at illustrating the basic ingredients of ensemble modelling 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,…,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 (**Figure 1a**), 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 show little deviation from the equal weighting scheme (with small random fluctuations though) traditionally used in the MMEs. Hence, the optimal combination (mmeS) and the optimal weighted combination (mmeW) 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.

Bias has been introduced into the ensemble by shifting the distribution of two-thirds 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 (**Figure 1b**) does not appear symmetric with respect to the mean RMSE, with notable distortions at the maximum RMSE for k≤4 (i.e., one-third of models). The upper bound of the RMSE values is defined from the ensemble combinations consisting of 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 (i.e. lowest RMSE) uses all unbiased models plus same amounts of biased equally members from both sides. As for the weighted ensemble, no conclusion 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 (**Figure 1c**) or inflated (**Figure 1d**) variance. Due to the bias-variance dilemma, the case with smaller variance (left) achieves lower RMSE for low k (compared to the i.i.d. case) 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 (**Figure 1e**) has been introduced among the first three members (ids 1-3) and separately, a negative correlation between two members (**Figure 1f**), 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 (i.e., the sub-ensemble has only non-correlated members); for negative ones, it tends to use only anti-correlated members. The same is seen also for the optimal weighted scheme: positively correlated members are treated as one, negatively correlated are significantly promoted over the i.i.d. members.

Using one-at-a-time perturbations in bias, variance and covariance, we investigated the skill of the three examined ensemble products through synthetic timeseries. The outcome of the exersize shows the following:

a. mme: its RMSE is reduced, compared to the i.i.d. case, if within the sample exist few members with lower variance or negative correlation. Contrary, its error is augmented from the presence of biased members. All the above can be directly explained by the bias-variance-covariance decomposition.
b. mmeS: for bias and variance perturbations, the optimal combination tends to use subsets built from i.i.d. members and members with balanced properties, i.e. biased from both signs, under- and over-dispersed. For the correlation perturbations, the optimal combination uses:
   - the subset built from i.i.d. members and only one member from the positively correlated cluster
   - the subset built from the negatively correlated members
c. mmeW: compared to the i.i.d. members, the weighted scheme:
   - reinforces members with lower variance and weakens members with higher variance
   - treats all redundant members as one and reinforces negatively correlated 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 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.

3. Data and Methodology

The material presented in the previous section demonstrated clearly through a well-defined mathematical formulation that building ensembles on the basis of “including as many 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. The necessary ingredients for ensemble building, using either the entire members with weights assigned or a subset of them with equal weights, and specifically, the optimization of the ensemble error:

- points, through the bias-variance-covariance decomposition, towards the bias correction of the models and the use of uncorrelated or negatively correlated ensemble members (equal weights, sub-ensemble);
- relies, through the accuracy-diversity decomposition, on finding the trade-off point between accurate and diverse members (equal weights, sub-ensemble);
- provides, through analytical formulas, weights for all ensemble members dependent on their error covariances (Potempski & 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, (a) the arithmetic mean of the entire ensemble (mme), (b) the arithmetic mean of an ensemble subset (mmeS), linked to the error decompositions (2.1, 2.2) and (c) the weighted mean of the entire ensemble (mmeW), linked to the analytical optimization (2.3). Note that mmeS is a general case of mme and a special case of mmeW (if weights can only take two discrete values). The principal objective addressed is the emergence of ways to produce a single improved forecast out of an ensemble that potentially outscores the traditional arithmetic mean as well as the best numerical model.

The critical model parameters for the techniques investigated in this work for ensemble member weighting or selecting, are bias and weights (straightforward) for mmeW and effective number of models and cluster selection for mmeS. They are briefly explained now.

- *Bias correction*. 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 2) 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; Delle Monache et al., 2008; Kang et al., 2008; McKeen et al., 2005; Galmarini et al, 2013); the correction applied in this work refers to a simple shift of the whole distribution within the examined temporal window, without any scaling or multiplicative transfer function.

- *Effective number of models*. The optimal ensemble estimator generally uses a subset of the available models, characterized as effective number (M_{EFF}) of models. In principle, M_{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). An analytical way to calculate M_{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.
- *Clustering Measures*. Given a dataset of N instances $\mathbf{X} = \{X_1, X_2, . . . , X_N\}$, a clustering algorithm generates r disjoint clusters based on a distance metric. Each clustering solution is a partition of the dataset **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 for each particular data-set (Fern and Brodley, 2004). Here, we use the unweighted pair-group average as the *agglomeration method* and the standard Euclidean distance as the *distance metric.* The clustering algorithm has been utilized against the d_m matrix defined in Table **1**, namely the corr(d_i,d_j), which generates more dissimilar errors compared to the e_m metric (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).

All time-series utilised originate 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, involving 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 drivers, air quality models, emissions 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 direct comparison of the simulated fields with the air quality measurements available from monitoring stations across the continent, at large temporal and spatial scales, is considered essential to assess model performance and identify model deficiencies (Dennis et al., 2010). This analysis falls within the context of operational evaluation of regional-scale chemical weather systems where most of the peaks in the energy spectrum are in the high-frequency era (hour, day, week). Together with the fact that the monitoring network extends over the whole continent, it emerges that the AQMEII database is suitable to capture the core temporal and spatial dependencies of the examined pollutants.

The analysis considers *hourly* time-series for the JJA (June-July-August) period. For European ozone, the ensemble constitutes from thirteen models, which give rise to 8191 different combinations (ensemble products). In section 4, we make use of spatially aggregated time-series (EU1 to EU4, illustrated in Figure 9a) while section 5 utilises time-series at point locations (451 stations). All data used refer to Phase I of the initiative. The evaluation of the examined ensemble products (mme, mmeW, mmeS) will rely on several indices of error statistics calculated at rural receptors. We present them in Table **1**. Those metrics can be used for the validation of each single ensemble configuration (f_i) as well as for the ensemble mean (f_{ens}).

4. Interpretation of the ensemble error in light of its terms

The goal of the section is to assess the properties of the ensemble error for the examined ensemble products and in particular, the characteristics that show robustness and allow the position of skilled predictions. Once those basic ingredients have been identified over a few regionally-averaged time-series, the potential predictability of the ensemble schemes at all available stations will be assessed in the next section.

The cumulative density functions (cdf) of the models and the observations at the four sub-regions are presented in **Figure 2**. The distribution of the models around the observations, across all percentiles, demonstrates the highest symmetry in EU4. On the opposite side we find EU3, where the ensemble is reliable only around the median. For the other two domains, EU1 and EU2, the ensemble replicates well the interquartile range but the averaging out of errors does not work properly at the extremes. The comparison of the cdfs demonstrates that the ensemble mean (mme) at the extreme percentiles should be treated with caution.

In an ideal ensemble, the rank histogram distribution should, on average, be flat. But, a flat rank histogram does not necessarily indicate a good forecast (Hamill, 2001); it only measures whether the observed probability distribution is well represented by the ensemble. In fact, the analyzed dataset (EU4r) has a relatively flat Talagrand diagram (Figure **3a**) but this accurate representation of the observational variability is not reflected symmetrically across all distribution bins as already seen in Figure **2**. If we would plot four rank histograms, one for each distribution quartile, we would face significant departures from flatness, especially outside the interquartile range.

Focusing on the ensemble error, the RMSE of the mean of all possible combinations as a function of the ensemble size (Figure **3b**) justifies the statement obtained theoretically, 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~6).
- a *minimum RMSE*, among all combinations, systematically emerges for ensembles with a number of members k < M (k~3-6). Applying the eigen-analysis on the error matrix, it also yields M_{EFF}=3.

The probability density function 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 are well below 50% of the total combinations, implying that random draws from the pool of models is highly unlikely to produce better results than the ensemble mean; at the same time, those fractions are not negligible, leaving space for significant improvements of the mme. For k=6, the 13 models give rise to 1716 combinations; each model participates at 792 of them. The fractional contribution of individual models (for k=6) to skilled sub-groups (portion of skilled combinations per model) is given with the red numbers. 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 complicated pattern as a result of different model variances and covariances. Clearly, they depart from homogeneity (equal weighting scheme shown with the red straight line).

The error, variance and correlation (with observations) of the thirteen ensemble members are presented in a Taylor plot (**Figure 3c**). They visually form three clusters. A low skill cluster includes models 1, 2 and 10, which have the highest error, minimum correlation with observed data and appear under-dispersed. Model 5 also belongs to that group but its variance is closer to the variance of the observations. The intermediate skill cluster contains models 3, 6, 7, 11 and 13 with average (11, 13) to low (3, 7, 6) error, and 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 slight over-dispersion though). Considering the participation statistics of the previous graph (given by the red numbers), we see that the models contributing more frequently to skilled combinations 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 part of skilful ensemble groups compared to low skill models. On the other hand, even low skill models can yield good results in the right combination. Overall, the multi-model average (mme) is a robust estimate with lower error than the candidate models but with reduced variance.

The application of the clustering procedure yielded five disjointed clusters (**Figure 3d**). Looking at the dendrogram, 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 of the Taylor plot reveals the similarities of each cluster in terms of error, correlation and variance. Clustering according to d_m generates the clusters visible in the Taylor plot. Many ensemble combinations with non-redundant members can be inferred from those plots; in addition, combinations that should be avoided are also noted. The d_m dendrogram also explains the reasoning behind the negative weights calculated analytically. The model pairs identified with highly correlated errors [like 4 and 12 or 11 and 13] are given weights of opposite sign, as seen in Figure **3b**.

Error statistics (<RMSE>) of the ensemble members and products (mme, mmeW, mmeS) for JJA 2006 at all selected sub-regions, using variable window size (1 day, 2 days, 4 days, 23 days, 46 days, 92 days) are shown in Figure **4**. The x-axis is the number of chunks in which the JJA time-series is sliced; hence it is inversely proportional to the window size. The skill of the deterministic models varies with location. A very good model at one site may perform averagely in another. As for the ensemble products, the following inferences can be drawn:

a. mme vs best model

The conditions leading to an ensemble superior to the best single model are illustrated in Figure **5** (without loss of generality, we consider the EU4r case). For correlated models, they depend on the skill difference among members and the amount of redundancy in the ensemble (i.e. the error dependence). The variation explained by the highest eigenvalue reflects the degrees of freedom in the ensemble (and hence the redundancy). The pairwise plot (Figure **5a**) of the skill difference (measured by <MSE>/MSE(best)) versus the ensemble redundancy (measured by the explained variation by the maximum eigenvalue) 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 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. The ensemble error gain (i.e. the difference between the ensemble error and the average error of the models) 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 (see Table 2), the mme error would always be lower than any single model's error since the MSE ratios (worst/best) are smaller than 14 (=M+1). Figure **4** shows that only in EU4r mme error is better than the individual models. This occurs because for correlated models, the condition is also restricted by the redundancy (eigenvalues spectrum). The joint conditions for the skill difference and the redundancy, for correlated models, granting an ensemble with mme error lower than the best model are presented in Figure **5b**. The RMSE ratio of mme over the best single model for the case of M (=13) correlated models shows that only in EU4r the explained variation by the highest eigenvalue has the correct value for the specified model MSE ratio [(explained variation by the highest eigenvalue, skill difference): 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 more profound balanced distribution of members. Indeed, in EU4r, the distribution of the models around the observations, across all percentiles, demonstrates high symmetry (**Figure 2**).

b. mme vs mmeS

The error derived from a reduced-size ensemble mean (mmeS) 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 mmeW's error. The estimation of the optimal weights is straightforward (Table 2), but the sub-selection of members in mmeS is not. Since mmeS uses equal weights, we can apply the concepts deployed by the two error decompositions and compare those properties with the ones of mme.

(i) <u>*Accuracy-Diversity*</u>. A 2-dimensional plot of accuracy versus diversity, with RMSE displayed as a third dimension (in color) is shown in **Figure 5c**. 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 (i.e. minimum MSE) 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 skilful ensemble products only if their diversity is very high. Analogously, combinations with good accuracy (better than average) but low diversity result in combinations with skill lower than the mme. Diversity with respect to the

ensemble mean can be derived independently of 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.

(ii) <u>*Variance-Covariance*</u>. Similar results are obtained in terms of the variance-covariance decomposition in **Figure 5d**. 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 falls while the covariance term deteriorates. Skilful combinations have relatively low covariance. Ensembles consisting of strongly positively correlated members bring redundant errors in the ensemble that does not cancel out upon averaging, producing overall larger errors.

Following the discussion of the previous section, we examine if the direction of move in the 2D space of the error terms, from mme to mmeS, has any systematic regularities. **Figure 6** displays the fractional change in accuracy [-1 + accuracy ratio (mmeS/mme)] versus the corresponding fractional change in diversity for all (92 in total) 1-day segments. At the same figure, we plot the corresponding changes of variance/covariance and skill difference/explained variability. The color scale indicates the RMSE ratio between the two ensemble means. Using dissimilar time-series from the four examined sub-regions, we observe that the optimal sub-ensemble combination (mmeS) compared to the full ensemble (mme) generally:

- Improves accuracy and by a smaller portion lowers its diversity. In other words, between accuracy and diversity, the controlling factor in those experiments in terms of error minimization is accuracy more than diversity.
- Lowers variance (term in Eq. 2) and by a higher portion lowers the covariance (term), implying that, between variance and covariance, the controlling factor for error minimization is covariance more than variance.
- Reduces the redundancy (as measured by the explained variability by the maximum eigenvalue) and by a higher rate reduces the skill difference among members, indicating that skill difference is more pronounced in error minimization than error correlation.

The converged findings from four dissimilar ozone time-series indicate that, for example, training mmeS through learning diversity algorithms (e.g. Kuncheva, L. and Whitaker 2003;

Brown et al., 2005) is not as effective as algorithms applied on the model's error covariance (e.g., Liu and Yao, 1999; Lin et al., 2008, Zanda et al., 2007).

c. mme vs mmeW

The error of the weighted ensemble mean (mmeW) is always superior since it has been analytically derived to minimize the MSE. For small window sizes (less than 4 days), the mmeW error is superior to the theoretically derived lower bound for the mme error (2nd term in the bias-variance-covariance decomposition) if models were uncorrelated. An insight for the sign of the weights can be inferred from the clustering according to d_m.

Like mmeS, mmeW improves the error and also replicates better the observed variance (**Figure 3c**) (similar results apply also to the ensemble product generated from spectral optimization demonstrated in Galmarini et al., 2013). The distribution around the truth in all those ensemble products has always higher symmetry compared to mme, as can be seen in **Figure 2**. In addition, they all perform much better at the extremes compared to the mean of the full ensemble.

4.1. Sensitivity of the ensemble error to the length of the training data

The temporal robustness of the two weighting schemes is now explored in order to identify the predictive skill of those products. 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 forecasts at daily to weekly time-scales (e.g. Galmarini et al, 2013).

The weights, the mean bias and the effective number of models have been re-calculated for variable time-series length that is progressively increasing from 1 to 92 days, for the four European sub-regions (**Figure 7**). The differences in the parameters weights and MB, calculated from consecutive blocks, show that both tend to stabilize after 40-60 days. The same is approximately also true for the effective number of models. Linked to the previous discussion, we hence conclude from the use of different time-series that a lower bound for the training window length that generates robust weight estimates is roughly 2 months.

Following the explored temporal sensitivity of the weights and M$_{EFF}$, we now examine the robustness of those estimates into future cases and in particular their capability in making accurate predictions. All ensemble products have been evaluated against the same test set, consisting of 30 equally spaced days from JJA (3rd June, 6th June, 9th June, etc). Eight different sets of weights are examined for each ensemble model, originating from four different lengths for the training period (namely, 1 day, 11 days, 31 days and 62 days) and two bias-correction schemes (namely, the ideal for the test set and the one calculated from the training set). We denote the weights trained over a sufficiently long training period as *static* (e.g. weights calculated over a sample of 62 days), to distinguish them from the *dynamic* weights (i.e. 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 (mmeW) 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 (mmeW weights are real numbers, hence there are infinite weighting vectors where only one is optimal but there should exist many combinations without major skill difference from the optimal).

The sensitivity of the ensemble products skill as a function of the training period length and the bias correction scheme is presented in Table 3. The following conclusions can be inferred for the daily forecasts:

- The weights derived through analytical optimization (mmeW) do not correspond to products with similar properties between consecutive days in cases of limited-length training datasets. On the other hand, static weights trained over a period longer than 30 days outscore all other products.
- MmeS is always superior to the mme, in all examined modes (historic, prognostic with static/dynamic weights). It also achieves lower error than mmeW with dynamic weights.
- In view of the predictability limits of each scheme, the achieved forecast MSE of mmeW is roughly 25 times higher than its hindcast MSE if bias correction is ideal and 50 times its hindcast MSE if bias correction is non-optimal. For mmeS, the forecast MSE is roughly double its hindcast MSE if bias correction is ideal and quadruple its hindcast MSE if bias correction is non-optimal.

- In many cases, the forecast MSE of mmeS and mmeW outscores the hindcast MSE of the mme. It systematically emerges in cases with ideal bias-correction.

Weighting is a risky process (Weigel et al., 2010) and its robustness should be thoroughly explored prior to operational forecasting. In diagnostic mode (training phase), mmeW minimizes the error achieving at least an order of magnitude lower MSE compared to the other ensemble products (Table 3). In prognostic mode (testing phase), if the training data have sufficient extent (at least 30 days), the minimum error is obtained with mmeW while for the case of limited training data, the minimum error is obtained with mmeS. An improvement similar to the one obtained through the mmeW scheme (bias correction, model weighting) has been documented in weather forecasting with MME (Krishnamurti et al., 1999), where weights were estimated from multiple regression. Similarly, improvement based on recent representation of an ensemble subset is documented in Galmarini et al., 2013. Other ensemble products based on learning diversity or covariance did not systematically outscored mme (not shown).

5. Predictability assessment at the monitoring stations

In the previous section, using four dissimilar regionally-averaged time-series, we have seen that in prognostic mode, mmeW with static weights (i.e. calculated over a 60 day interval) 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, mmeW and mmeS on *blind* time-series. We split records into a test dataset (30 equally spaced days from JJA: 3rd June, 6th June, 9th June, etc) and a train dataset (remaining two-third of the records). Using the train dataset, we first bias correct the time-series and then we estimate the mmeW weights and mmeS subset. Last, we apply the estimated parameters from the training dataset (weights, bias, M_{EFF}, clusters) into the test dataset.

Training Phase. **Figure 8** displays the mmeW weights 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 spatially robust error covariance. 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. Last, the highest frequency of model use in mmeS is observed for the models having the higher mmeW weights. Hence, although calculated with different approaches, the weight peaks at seasonal scale of the mmeW and mmeS have similarities (i.e. models 3, 5 and 6 that receive on average the highest weights are also the ones used most frequently in mmeS).

Using various input matrices, we find the effective number of models to vary between 2 and 8, through a homogeneous spatial pattern (**Figure 9**). 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, mmeS uses only either 3 or 4 models, while over 80% of the sites need 2-5 models from the pool.

Testing Phase. The presented results hereafter assess the predictability of the examined schemes trained over a finite time-series. Besides the summary statistics, the skill is also evaluated geographically as well as a function of the effective number of models. In addition, the effect of a 2nd order correction in bias is investigated. We conclude with the presentation of results for NO_2 and PM10, following the same methodological framework.

Forecast 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 (**Figure 13**) together with the single deterministic models. The benefits of ensemble treatment, either in the form of simple averaging models (mme) as well as using more sophisticated techniques (mmeS, mmeW) are clearly evident. Besides the error (RMSE), mmeS and mmeW also improve the correlation and the variance of the output with respect to mme. 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, mmeW and mmeS is shown in **Figure 10** for the gross error (RMSE) and the ability to capture the extreme upper tail of the distribution via the hit rate indicator. The weighted ensemble, in the form of mmeW or mmeS, 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 mmeW and 25% for mmeS. Half of the stations experience RMSE lowering in the mmeW (mmeS) case by up to 13% (10%) and the other half in the range 13% to 35% (10% to 25%). There exists a weak tendency for larger improvement at the sites with the higher RMSE. The histogram of the errors from all stations for (mme, mmeS, mmeW) 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 μg/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 mmeW and mmeS, with a more homogeneous pattern in mmeW. The median hit rate of mme is 28% and becomes 44% in mmeS and nearly doubles (52%) in mmeW. 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 mmeW and 30% in mmeS.

Effect of M_{EFF}

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 (**Figure 11a**). 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 (**Figure 11b**). It unfolds the differences in the statistical distribution of the three ensemble products. Overall, the spread (range) of concentrations is monotonically decreasing as $\mathrm{M_{eff}}$ increases. For M_{eff} ≤4, this is due to equal reductions in mean and standard deviation, for M_{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 $\mathrm{M_{eff}}$ >6, i.e. when the range of concentration is well bounded below 120 μg/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}. 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 mmeS and mmeW (**Figure 11c**). On the other hand, mme has notably lower PCC for $M_{eff} \leq 4$, pointing again to the discrepancies in capturing the whole range of variability when there is a significant amount of extreme records (over 120 μg/m^3). Similar result is found for the standard deviation ratio (STDR) (**Figure 11e**). In terms of error (RMSE) (**Figure 11d**), it is a decreasing function of M_{eff} and the three ensemble products start to converge for $M_{eff} > 6$. As M_{eff} increases, the distribution of the models around the observations is gradually becoming more symmetric, hence the gain from mmeW or mmeS is minimized as the mme sample has already a quite symmetric distribution. This can be seen in **Figure 12**, where Talagrand diagrams have been plotted according to the station's M_{EFF}. 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 mmeS or mmeW over mme become substantial. Last, the improved hit rate (hitR) in mmeW and mmeS over mme seen in **Figure 10**, has a coherent pattern across all M_{eff} values, as also seen in **Figure 11f**.

Effect of the bias-correction scheme

So far, the model outputs were separately adjusted for systematic errors by a 1st order bias correction. Here we test the effect of an additional adjustment applied on their spread through a 2nd order bias correction. 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 **Figure 13** through a comparison of their composite skill 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 2nd order correction, which is also reflected in the ensemble products and in particular in mme and mmeS. 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 mmeW, 2nd mmeS, 3rd mme) with the known overestimation tendency for concentrations below 75 μg/m^3 and underestimation above that threshold. The differences between the schemes and products become substantial for the limited records exceeding the 180 μg/m^3 value. In general, the mmeW provides noteworthy better forecasts over mmeS and mme even with fewer corrections (for example mmeW trained with 1st order corrected models scores better than mmeS from 2nd order corrected models); this also applies for mmeS over mme.

Results for other pollutants (NO_2, PM10)

For the other two pollutants (NO_2 and PM10), 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 PM10. Hence, between different species, the statistical improvements are proportional to the MME skill in forecasting the specific species. In othet words, mmeS and mmeW 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.

The gross improvement in the RMSE of the multi-model ensemble mean achieved through a 2nd order bias correction, compared to 1st order, was 0.6% for O_3, 2.1% for NO_2 and 11.8% for PM10. On the other hand, the improvement in the RMSE achieved through the exploitation of the ensemble mean in the form of mmeW or mmeS was 8.6% for O_3, 14.9% for NO_2 and 13.5% for PM10. Hence, the improvement in the error of the ensemble mean achieved through spread adjustment, on top of the correction of the systematic errors, does not outscore the improvements that can be achieved through proper weighting or sub-selecting.

6. Summary & 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 investigated to eliminate this distortion and yield ensemble output with symmetric residuals across all distribution bins are model weighting and model sub-selecting, both supported by mathematical evidence. The analysis makes use of continental-scale simulations and observations from AQMEII.

The goal of this work is to evaluate potential schemes to produce a single improved forecast out of an ensemble. The key results, obtained from the application of two general-purpose ensemble models to a representative air-quality dataset, can be summarized as follows (in order of decreasing generality):

1. 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 a random sample (i.e. under the i.i.d. assumption). Further, the ensemble mean is superior to the best single model given conditions that relate to the skill difference of the members and the ensemble redundancy.
2. 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 *mmeW* (error optimization through model weighting; keep all models) and *mmeS* (error optimization through trade-off between accuracy and diversity or variance and covariance; average on selected subset of models).
3. Unlike mme, mmeW and mmeS require some training phase to find robust weights or clusters. The mmeW skill was more sensitive to its controlling factors than mmeS. A 2-month period was found necessary for the stabilization of the mmeW weights. On the other hand, mmeS was robust using both static/dynamic modes. In prognostic mode, if the training data have sufficient extent (at least 30 days), the minimum error

is obtained with mmeW while for the case of limited training data, the minimum error is obtained with mmeS. Specifically:

- *mmeW*: the weights were rather sensitive to the length of the training period, requiring at least 30 days to approach an asymptotic consensus. Nevertheless, learning over long time-periods (~2 months) and using those weights in predictive mode proved robust and accurate. Under proper training, its forecast skill outperformed all other ensemble products as well as individual models. The improvement across all stations over the mme was up to 35% for the RMSE and around 85% for the median hit rate.
- *mmeS*: for the 13 member ensemble, the effective number of models was in the range 2-8, with the peak between 3 and 4. 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. For training data of limited length (< 1 month), its skill was also better than mmeW. For ozone, switching from mme to mmeS, the properties that were relatively corrected more were accuracy (over diversity), error covariance (over error variance) and skill difference (over error correlation). The learning algorithms for subset selection, based on a sole dependent function of the error (e.g., diversity) rather than the error, did not achieve higher skill than mme. The improvement across all stations over the mme was up to 25% for the RMSE and 57% for the median hit rate.

4. The gross improvement in the RMSE of the multi-model ensemble mean achieved through the first and second moment correction of the modelled time-series, compared to only first moment correction was 0.6% for O_3, 2.1% for NO_2 and 11.8% for PM10. On the other hand, the improvement in the RMSE achieved through the exploitation of the ensemble mean in the form of mmeW or mmeS was 8.6% for O_3, 14.9% for NO_2 and 13.5% for PM10. Hence, even with adjustments in the systematic error and the spread in the models of an ensemble, a portion of its potential predictability is lost by using solely full ensemble averaging; superior improvements can be achieved through the optimization of an error decomposition approach.
5. For i.i.d. samples, the effective number of models equals the ensemble size (members). The mmeS and mmeW improve the skill of mme by constraining the ensemble into another where participating models replicate better the properties of an i.i.d. sample. Using M_{EFF} as indicator of i.i.d. sample, the decomposition of the skill as

a function of the effective number of models demonstrated that for ozone, the three products were converging with increasing M_{EFF}. Those cases were occurring for intermediate concentration ranges, that all models are somehow tuned to replicate. On the other end, as M_{EFF} was decreasing and the ensemble was departing from behaving as an i.i.d. sample, the error gain from mmeS or mmeW over mme was gradually increasing, reaching on average 15% and 30% respectively. The extreme records were generally found in the assymetric range of the ensemble.

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 μg/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.

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 benchmark against which all other weighting schemes should be evaluated. A general roadmap for the non-trivial problem of weighting (mmeW) or sub-selecting (mmeS) from an ensemble is attempted hereafter:

I. Generate a raw ensemble and apply *bias correction* techniques to remove systematic errors (prerequisite for mmeW)

II. Evaluate indices of skill difference and redundancy to assess the superiority of the ensemble mean against the best single model

III. 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 length of the training dataset is determined from physical concepts as well as the statistical properties of the specific ensemble.

IV. 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, Bulletin of the American Meteorological Society 83: 450-452, 2002.

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

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

Brown, G., Wyatt, J., Harris, R. and Yao, X.: Diversity creation methods: a survey and categorisation, Journal of Information Fusion, 6(1): 5-20, 2005.

Delle Monache L., J. Wilczak, S. McKeen, G. Grell, M. Pagowski, S. Peckham, R. Stull, J. McHenry, J. McQueen: A Kalman-filter bias correction method applied to deterministic, ensemble averaged, and probabilistic forecast of surface ozone, Tellus Ser. B, 60: 238-249, 2008.

Dosio, A., Paruolo, P.: Bias correction of the ENSEMBLES high-resolution climate change projections for use by impact models: Evaluation on the present climate, Journal of Geophysical Research D: Atmospheres, Volume 116, Issue 16, Article number D16106, doi:10.1029/2011jd015934, 2011.

Errico R.: What Is an Adjoint Model?, Bulletin of the American Meteorological Society, 78, 2577- 2591, 1997.

Fern, X. Z. and Brodley, C. E.: Solving cluster ensemble problems by bipartite graph partitioning, in Proceedings of 21st International Conference on Machine Learning

(ICML2004), Banff, Alberta, Canada, 4-8 July 2004, ACM Press, pp. 281-288, doi:10.1145/1015330.1015414, 2004.

Galmarini S., R. Bianconi, W. Klug, T. Mikkelsen, R. Addis, S. Andronopoulos, P. Astrup, A. Baklanov, J. Bartniki, J.C. Bartzis, R. Bellasio, F. Bompay, R. Buckley, M. Bouzom, H. Champion, R. D'Amours, E. Davakis, H. Eleveld, G.T. Geertsema, H. Glaab, M. Kollax, M. Ilvonen, A. Manning, U. Pechinger, C. Persson, E. Polreich, S. Potemski, M. Prodanova, J. Saltbones, H. Slaper, M.A. Sofiev, D. Syrakov, J.H. Sørensen, L.Van der Auwera, I. Valkama, R. Zelazny: Ensemble dispersion forecasting—Part I: concept, approach and indicators, Atmospheric Environment 38(28): 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, 2013.

Geman S., E. Bienenstock, and R. Doursat.: Neural networks and the bias/variance dilemma, Neural Computation 4(1): 1-58, 1992.

Gneiting T., A.E. Raftery, A.H. Westveld, and T. Goldman: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation, Monthly Weather Review 133: 1098-1118, 2005.

Hamill T.: Interpretation of Rank Histograms for Verifying Ensemble Forecasts, Monthly Weather Review. 129: 550-560, 2001.

Helton J. C. and F. J. Davis: Latin Hypercube Sampling and the Propagation of Uncertainty in Analyses of Complex Systems, Reliability Engineering and System Safety, 81: 23-69, 2003.

Iman R. L. and W. J. Conover: A Distribution-Free Approach to Inducing Rank Correlation Among Input Variables, Communications in Statistics: Simulation and Computation, B11: 311-334, 1982.

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

Kang D., R. Mathur, S.T. Rao, S. Yu: Bias adjustment techniques for improving ozone air quality forecasts, J. Geophys. Res., 113, D23308, doi:10.1029/2008JD010151, 2008.

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., C.M. Kishtawal, T.E. LaRow, D.R. Bachiochi, Z. Zhang, C.E. Williford, S. Gadgil, S. Surendran: Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble, Science 285(5433): 1548-1550, 1999.

Krogh A. and J. Vedelsby: Neural network ensembles, cross validation, and active learning, In Advances in Neural Information Processing Systems 7, pages 231-238, 1995.

Kuncheva, L. and Whitaker, C.: Measures of diversity in classifier ensembles, Machine Learning 51: 181-207, 2003.

Leith C.E.: Theoretical skill of Monte Carlo forecasts, Monthly Weather Review 102: 409-418, 1974.

Lin M., K. Tang, X. Yao: Selective negative correlation learning algorithm for incremental learning, in: Proceedings of the 2008 International Joint Conference on Neural Networks (IJCNN'08), Hongkong, China, pp. 2526-2531, 1-6 June 2008.

Liu, Y., Yao, X.: Ensemble learning via negative correlation. Neural Networks 12: 1399–1404, 1999.

Malamud, B.D., Turcotte, D.L.: Self-affine time series: measures of weak and strong persistence, Journal of statistical planning and inference 80(1-2): 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, Journal of Finance, 7, 77-91, 1952.

McKay M. D., R. J. Beckman, and W. J. Conover: 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.

McKeen S., J. Wilczak, G. Grell, I. Djalalova, S. Peckham, E.-Y. Hsie, W. Gong, V. Bouchet, S. Menard, R. Moffet, J. McHenry, J. McQueen, Y. Tang, G.R. Carmichael, M. Pagowski, A. Chan, T. Dye, G. Frost, P. Lee, R. Mathur: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004, J. Geophys. Res., 110, D21307, doi:10.1029/2005JD005858, 2005.

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, 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., T. Thorarinsdottir and T. Gneiting.: Uncertainty Quantification in Complex Simulation Models Using Ensemble Copula Coupling, Statistical Science 28 (4): 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., A. Riccio, I. Kioutsioukis, and S. Galmarini: Pauci ex tanto numero: reduce redundancy in multi-model ensembles, Atmos. Chem. Phys. 13: 8315–8333, 2013.

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., Jeriˇcevi´c, A., Kraljevi´c, 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., 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.

Stein M.: Large Sample Properties of Simulations Using Latin Hypercube Sampling, Technometrics 29(2): 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 R. Nakano.: Generalization error of ensemble estimators, In Proceedings of International Conference on Neural Networks, pages 90–95, Washington, DC, USA, 3-6 June 1996.

Varotsos C., M. Efstathiou, C. Tzanis and D. Deligiorgi: On the limits of the air pollution predictability; the case of the surface ozone at Athens, Greece, Environmental Science and Pollution Research 19(1): 295-300, 2012.

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

Zanda M., G. Brown, G. Fumera, and F. Roli: Ensemble Learning in Linearly Combined Classifiers Via Negative Correlation, Lecture Notes in Computer Science 4472, 524 pp, 440 - 449, 2007.

Table 1. Notation and Indices of skill and redundancy. A ‘*’ indicates standardized vectors.

Ensemble members (output of modelling systems) Ensemble Desired value (measurement)	$\boldsymbol{f_i}$, i=1,...,M $\bar{f} = \sum_{i=1}^{M} w_i f_i \, , \sum w_i = 1$ μ
Pearson Correlation Coefficient	$PCC \ = \frac{\frac{1}{N}\sum_i (f_i - \overline{f_i})(\mu_i - \bar{\mu})}{\sigma_{f_i}\sigma_\mu}$
Mean Bias	$MB \ = \frac{\sum(f_i - \mu_i)}{N}$
Root Mean Square Error	$RMSE \ = \sqrt{\frac{\sum(f_i - \mu_i)^2}{N}}$
Normalised deviation of models from observations	$e_m^* = \frac{e_i - \overline{e_\imath}}{\sigma_{e_i}} \ where \ e_i = \frac{f_i - \mu_i}{\sigma_i}$
Difference between the model error and the weighted multi-model error pattern	$d_m = e_m^* - R_{m,MM} \cdot MM^* \quad MM = \frac{1}{M}\sum_i e_i$
Threshold indices based on a contingency table for events (Forecasted/Observed)	Hit rate $= \frac{Y/Y}{Y/Y + N/Y}$
Accuracy term	$acc = E\left(\frac{1}{M}\sum_{i=1}^{M}(f_i - \mu)^2\right)$
Diversity term	$div = E\left(\frac{1}{M}\sum_{i=1}^{M}\left(f_i - \bar{f}\right)^2\right)$
Squared Bias	$\overline{b\imath as}^2 \ = \mathrm{E}\left(\frac{\sum(f_i - \mu_i)}{N}\right)^2$
Variance of errors	$\overline{varE} \ = E(var(\boldsymbol{f} - \mu))$
Covariance of errors	$\overline{covE} \ = E(cov(\boldsymbol{f} - \mu))$

Table 2. Analytical formulas for the 1-dimensional (single-point optimization) case (from Potempski and Galmarini, 2009).

	Uncorrelated models	**Correlated models**
Optimal Weights	$a_k = \frac{\frac{1}{\sigma_k^2}}{\sum_j \frac{1}{\sigma_j^2}}$	$\bar{\boldsymbol{a}} = \frac{\boldsymbol{K}^{-1}\boldsymbol{l}}{(\boldsymbol{K}^{-1}\boldsymbol{l},\boldsymbol{l})}$
Limits for mme (ensemble mean)	$MSE(\bar{f}) \leq MSE(f_1) \leq \cdots \leq MSE(f_m)$ if $\frac{MSE(f_m)}{MSE(f_1)} \leq M + 1$	$MSE(\bar{f}) \leq s_1 \leq s_2 \leq \cdots \leq s_m$ if $\frac{s_m}{s_1} \leq M$
Definitions	$\sigma_j^2 = variance\ of\ model's\ j\ error$	$s_j = eigenvalues\ of\ \boldsymbol{K}$ $\boldsymbol{K} = error\ covariance\ matrix$ $\boldsymbol{l} = [1,1,\ldots,1]^T$

Table 3. The mean MSE of the 30 daily cases, in training mode (H) and testing mode as a function of the training period length (1 day, 11 days, 31 days, 62 days) and ideal/non-ideal bias correction. The comparison has been applied to four European sub-regions and three selected ensemble products (mme, mmeW, mmeS). The cases with MSE lower than mme are given in bold and the best member is displayed with green color.

	Ideal bias correction					**Predicted bias correction**				
EU1r	H	1d	11d	31d	62d	H	1d	11d	31d	62d
mme	49.0	49.0	49.0	49.0	49.0	49.0	86.7	87.6	86.7	86.8
mmeS	9.9	**20.6**	**23.0**	**23.3**	**18.6**	9.9	**82.9**	**45.6**	**42.7**	**42.7**
mmeW	0.6	**41.8**	**18.2**	**14.3**	**13.7**	0.6	544.1	**39.3**	**28.8**	**27.8**
EU2r	H	1d	11d	31d	62d	H	1d	11d	31d	62d
mme	28.4	28.4	28.4	28.4	28.4	28.4	153.1	117.4	109.6	110.1
mmeS	10.2	**22.1**	**19.6**	**24.8**	**24.5**	10.2	**140.6**	**64.2**	**54.2**	**57.6**
mmeW	0.5	37.1	**24.3**	**15.0**	**13.7**	0.5	1021.3	**60.8**	**34.7**	**34.1**
EU3r	H	1d	11d	31d	62d	H	1d	11d	31d	62d
mme	285.5	285.5	285.5	285.5	285.5	285.5	371.0	342.9	342.8	342.6
mmeS	113.3	**176.4**	**190.7**	**140.3**	**140.3**	113.3	**299.8**	**246.1**	**207.0**	**206.9**
mmeW	1.7	507.4	**195.0**	**127.6**	**116.4**	1.7	4208.2	**323.1**	**203.7**	**185.3**
EU4r	H	1d	11d	31d	62d	H	1d	11d	31d	62d
mme	37.7	37.7	37.7	37.7	37.7	37.7	134.9	83.9	72.9	72.9
mmeS	9.7	**27.3**	**23.3**	**22.8**	**23.5**	9.7	138.4	**63.1**	**53.5**	**52.5**
mmeW	0.9	146.8	**29.2**	**25.2**	**22.6**	0.9	578.7	**83.5**	**53.1**	**48.3**

Figure Captions

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. (a) i.i.d. [top left], (b) bias perturbation [top right], (c,d) variance perturbations [middle], (e,f) covariance perturbations [bottom]. Please read text for explanations and note the different range of the y-axis between the different panels. 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.

Figure 2: Cumulative density function of observations (red circle) and models (coloured lines). At the same plot, the three ensemble estimators are also displayed, namely the multi-model ensemble mean (mme: square), the optimal weighted ensemble estimator (mmeW: green circle) and the optimal accuracy-diversity ensemble estimator (mmeS: blue circle). Please note the different range of the x-axis between the different panels.

Figure 3 (a) Talagrand diagram of the full ensemble (top left). (b) Ensemble error (RMSE) from all possible combinations of candidate models (EU4r). The notation is similar to Figure 1. The numbers in red express the fractional contribution of each model to skilled combinations (top right). (c) Multiple aspects of individual model skill through Taylor plot. The point R on the x-axis represents the reference field (i.e. observations) [bottom left]. (d) Clustering members with the corr(d_i,d_j) matrix (bottom right).

Figure 4: The mean RMSE of the models (colored lines) as a function of window size (1 day – 92 days). In addition, selected ensemble products are also displayed: mme (thick black), <mm_i> (thick dotted-black), mmeW (thick red), mmeS (thick dotted red). The bars show the theoretical minimum value (<var>/nm) for uncorrelated models. Please note the different range of the y-axis between the different panels.

Figure 5: (a,b) The RMSE ratio of mme over the best single model as a function of redundancy (explained variation by the maximum eigenvalue s_m) and model skill difference (<MSE>/MSE(best), evaluated from all combinations of 6th order (top left) and 13th order (top right). The diagram on the right has been evaluated at all observation sites. (c,d) Four dimensional representation of accuracy - diversity (bottom left) and variance – covariance (bottom 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.

Figure 6: Comparison between mmeS 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 calculated as *property*(mmeS)/*property*(mme). [top] Fractional change in accuracy versus fractional change in diversity. [middle] Fractional change of variance versus fractional change of covariance. [bottom] Fractional change in skill difference versus fractional change in error correlation. Please note the different range of the y-axis between the different panels.

Figure 7: Variability of weights (left column), bias (middle column) and effective number of models (right column) as a function of time-series length. Each thin-line represents a different model. The effective number of models is calculated through eigen-analysis and error minimization.

Figure 8: Ozone spatial weights (mmeW) 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 mmeS, from all the 451 stations for the test dataset.

Figure 9: [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 covariance of the diversity

matrix. [Bottom] like top but for M_{eff} based on the eigenvalues of the cor(e_i,e_j) matrix. Please note the different range of the y-axis between the different histograms.

Figure 10: [Top row] The RMSE of ozone at each observed site for mme (left). The behaviour at the upper tail of the distribution; percentage of correct hits for events > 120 µg/m3 for mme (right) [2nd, 3rd row] Like top row but for mmeW and mmeS. [Bottom row] The cdf of each spatial plot.

Figure 11: [Top] Statistical properties of mme, mmeS and mmeW forecasts versus observations from the 451 stations for the test dataset as a function of *Meff:* (a) mean and (b) 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, mmeS and mmeW from the 451 stations for the test dataset as a function of *Meff*: (c) PCC, (d) RMSE, (e) STDR (standard deviation ratio) and (f) Hit Rate.

Figure 12: [Top] Talagrand diagram of the full ensemble aggregated at the stations as a function of M_{EFF}. [Bottom] Cumulative density function of observations (red circle) and models (coloured lines), aggregated at the stations as a function of M_{EFF}. The ensemble mean is displayed with a square.

Figure 13: 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. Please note the different range of the x and y-axis between the different panels.

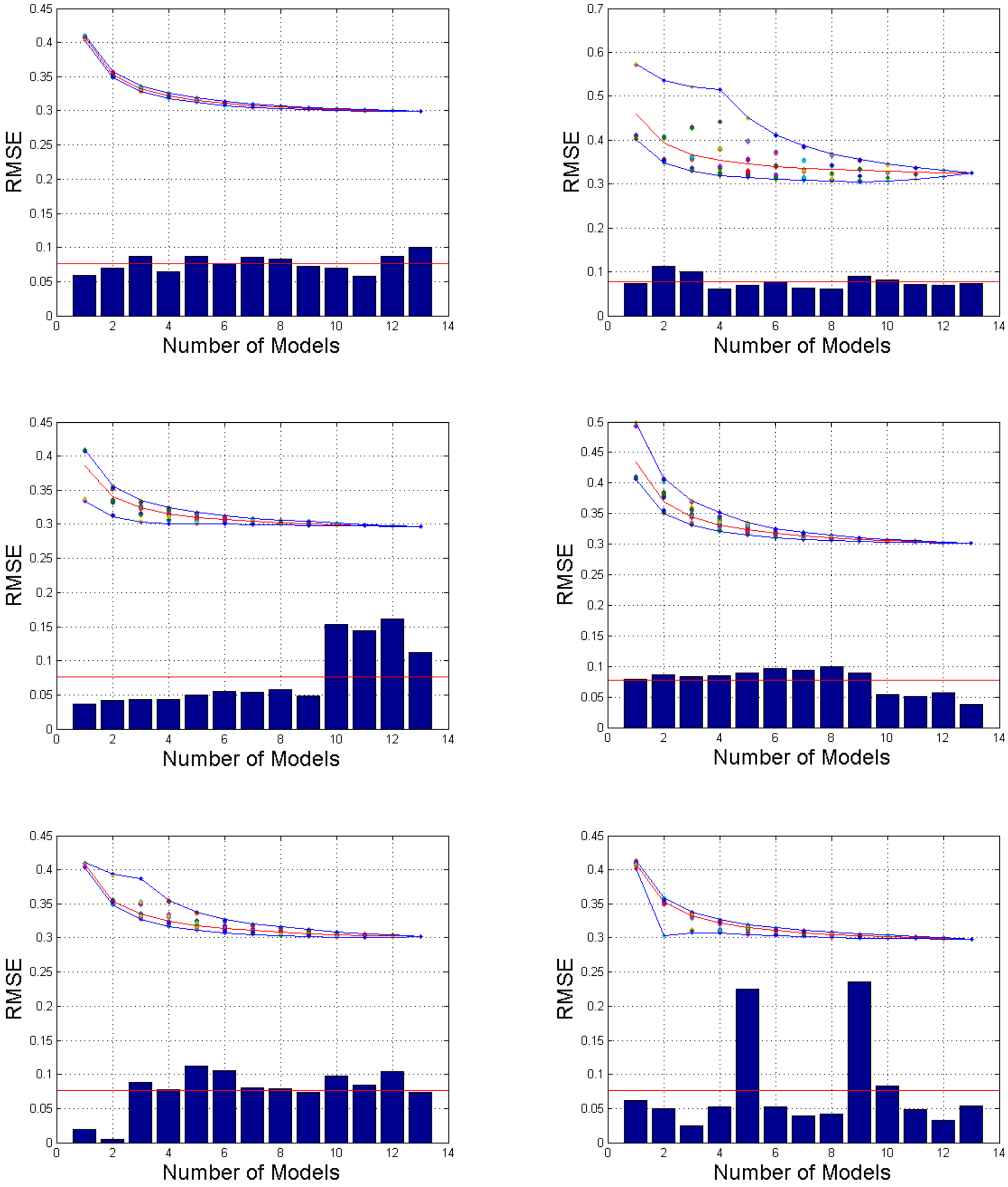

Figure 1: Ensemble error (RMSE) from all possible combinations of candidate models. The red curve on each plot represents the mean of the distribution of any k-model combinations while the blue curves form the min and max of the each respective distribution. (a) i.i.d. [top left], (b) bias perturbation [top right], (c,d) variance perturbations [middle], (e,f) covariance perturbations [bottom]. Please read text for explanations and note the different range of the y-axis between the different panels.

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.

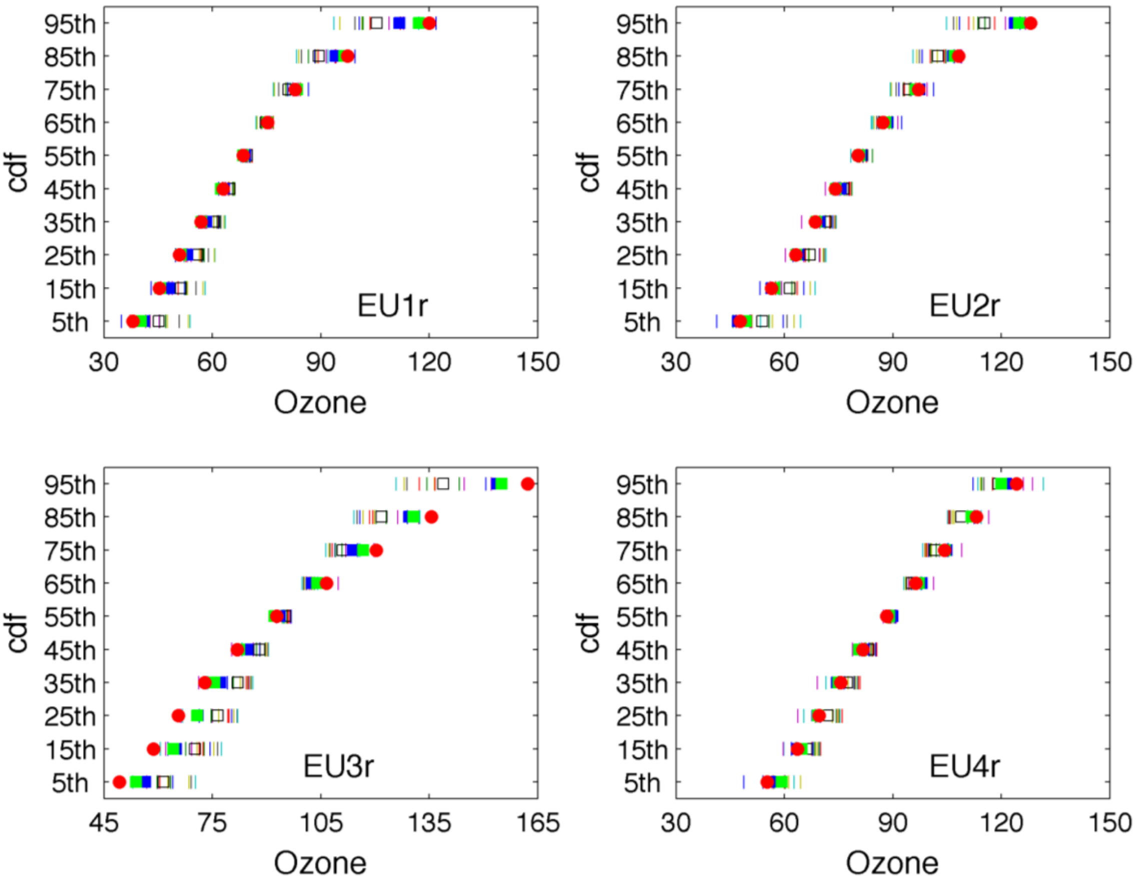

Figure 2: Cumulative density function of observations (red circle) and models (coloured lines). At the same plot, the three ensemble estimators are also displayed, namely the multi-model ensemble mean (mme: square), the optimal weighted ensemble estimator (mmeW: green circle) and the optimal accuracy-diversity ensemble estimator (mmeS: blue circle). Please note the different range of the x-axis between the different panels.

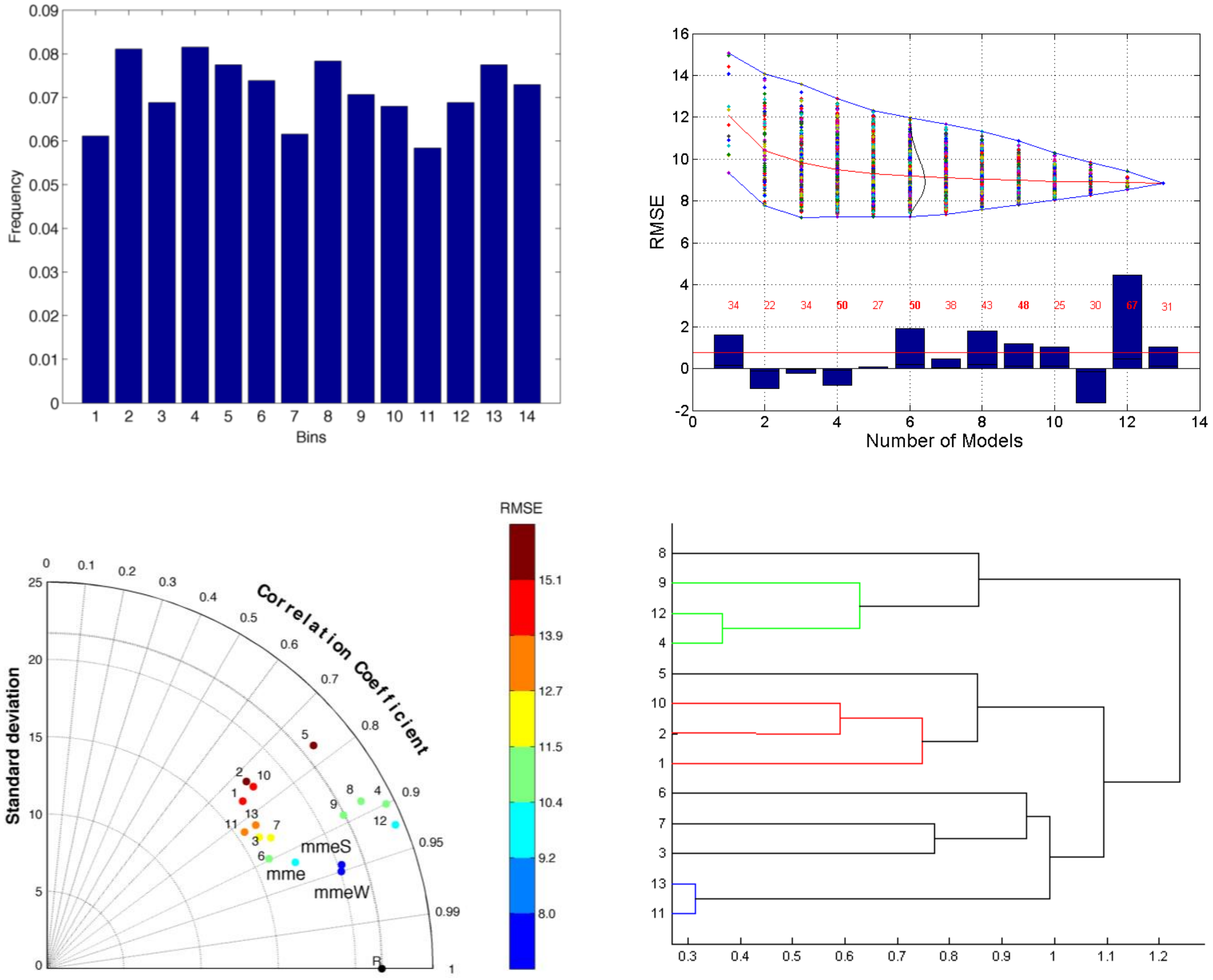

Figure 3: (a) Talagrand diagram of the full ensemble (top left). (b) Ensemble error (RMSE) from all possible combinations of candidate models (EU4r). The notation is similar to Figure 1. The numbers in red express the fractional contribution of each model to skilled combinations (top right). (c) Multiple aspects of individual model skill through Taylor plot. The point R on the x-axis represents the reference field (i.e. observations) [bottom left]. (d) Clustering members with the corr(d_i,d_j) matrix (bottom right).

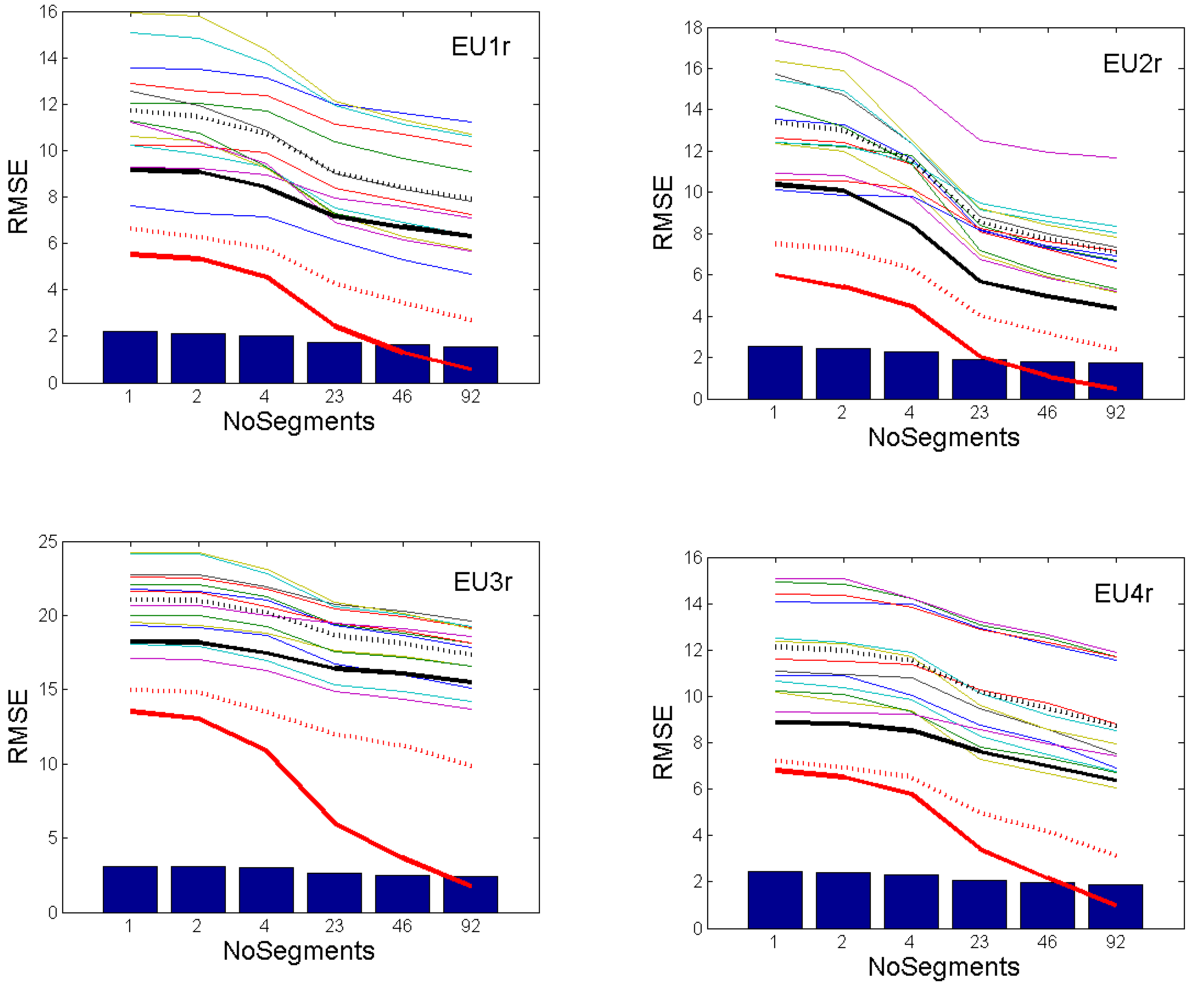

Figure 4: The mean RMSE of the models (colored lines) as a function of window size (1 day – 92 days). In addition, selected ensemble products are also displayed: mme (thick black), <mm$_i$> (thick dotted-black), mmeW (thick red), mmeS (thick dotted red). The bars show the theoretical minimum value (<var>/nm) for uncorrelated models. Please note the different range of the y-axis between the different panels.

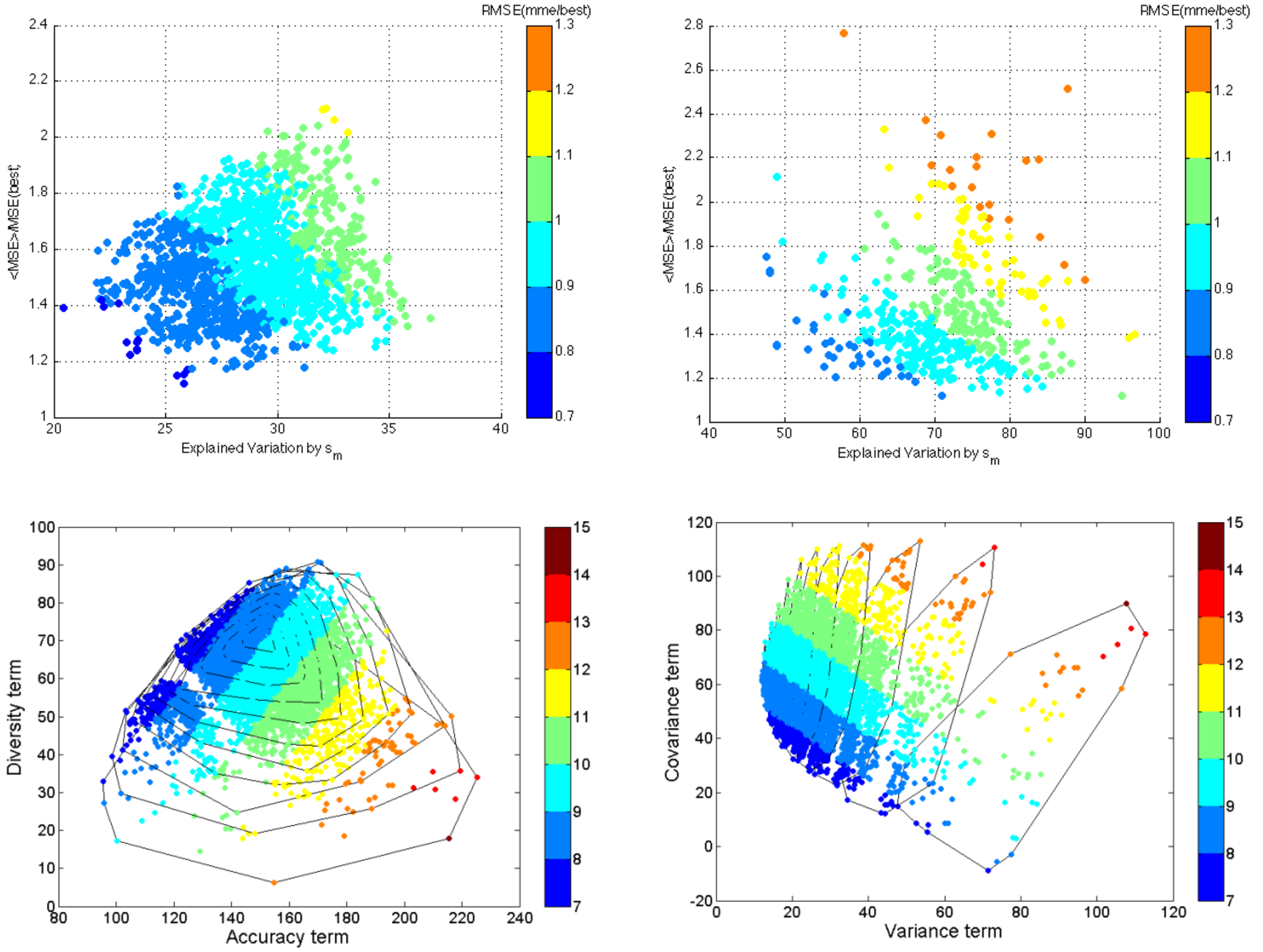

Figure 5: (a,b) The RMSE ratio of mme over the best single model as a function of redundancy (explained variation by the maximum eigenvalue s_m) and model skill difference (<MSE>/MSE(best), evaluated from all combinations of 6th order (top left) and 13th order (top right). The diagram on the right has been evaluated at all observation sites. (c,d) Four dimensional representation of accuracy - diversity (bottom left) and variance – covariance (bottom 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

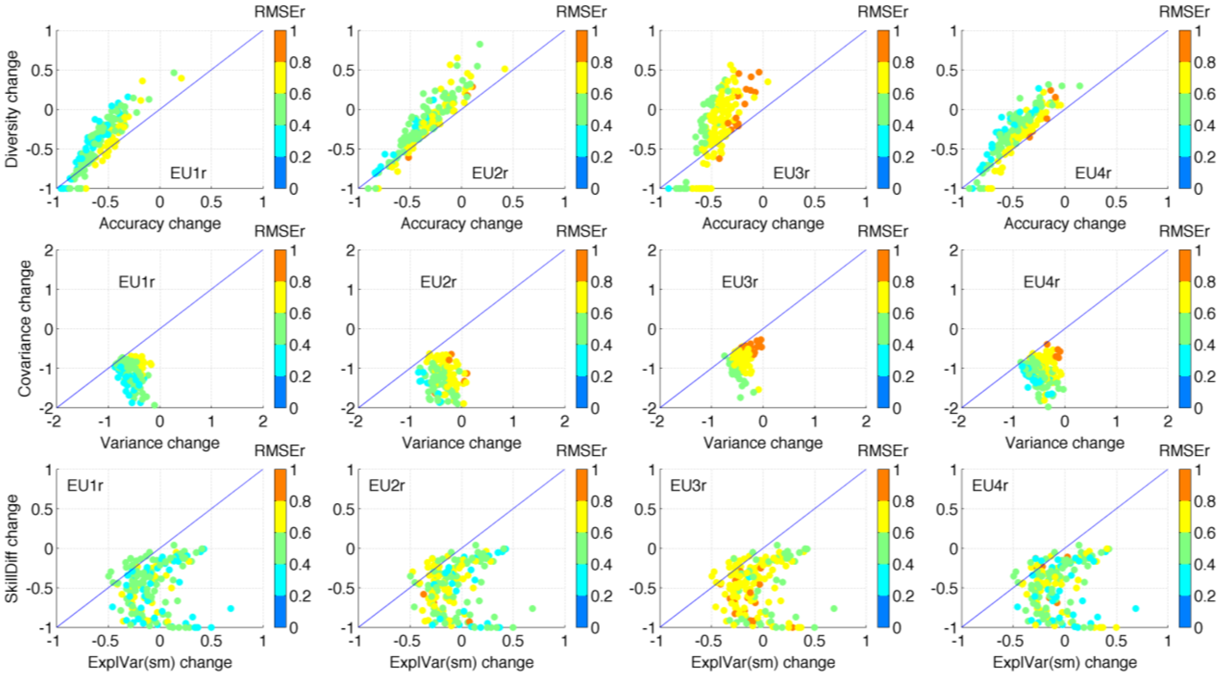

Figure 6: Comparison between mmeS and mme with respect to the error decomposition. Each of the 92 dots corresponds to an individual 1-day simulation. The color scale represents the RMSE ratio calculated as *property*(mmeS)/*property*(mme). [top] Fractional change in accuracy versus fractional change in diversity. [middle] Fractional change of variance versus fractional change of covariance. [bottom] Fractional change in skill difference versus fractional change in error correlation. Please note the different range of the y-axis between the different panels.

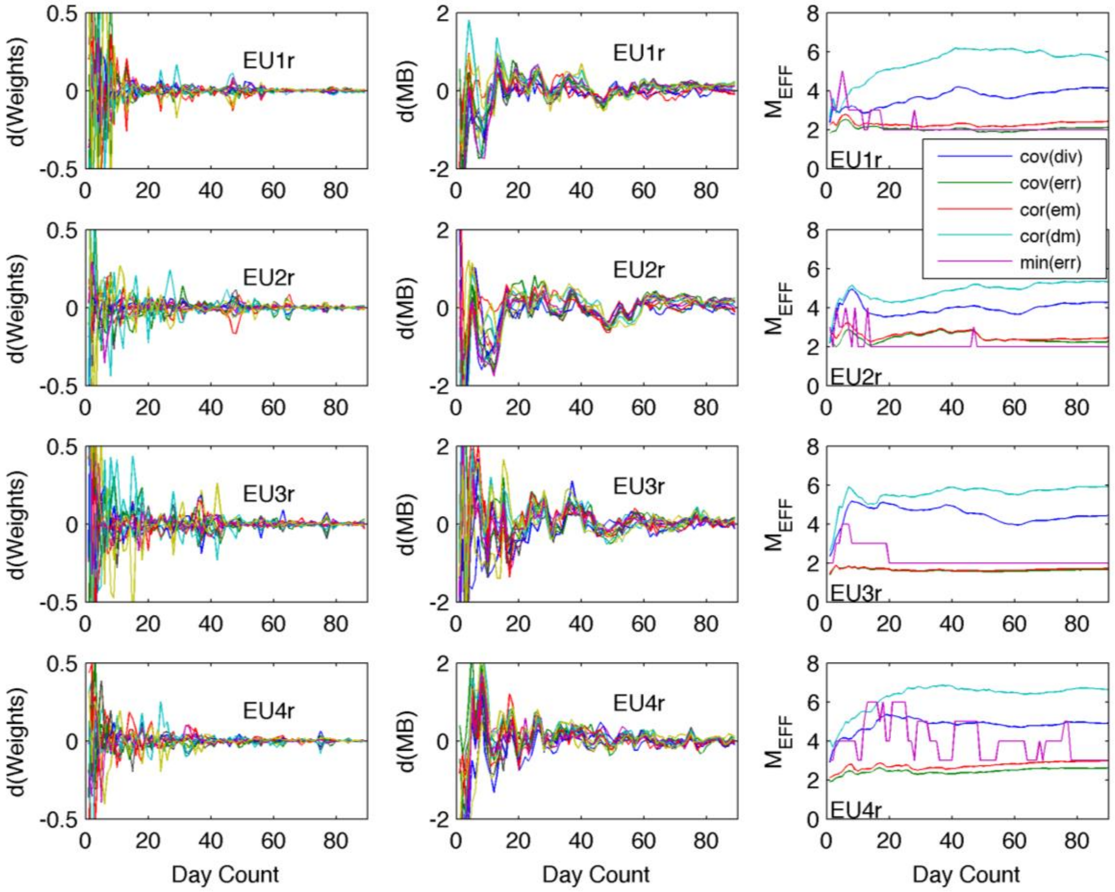

Figure 7: Variability of weights (left column), bias (middle column) and effective number of models (right column) as a function of time-series length. Each thin-line represents a different model. The effective number of models is calculated through eigen-analysis and error minimization.

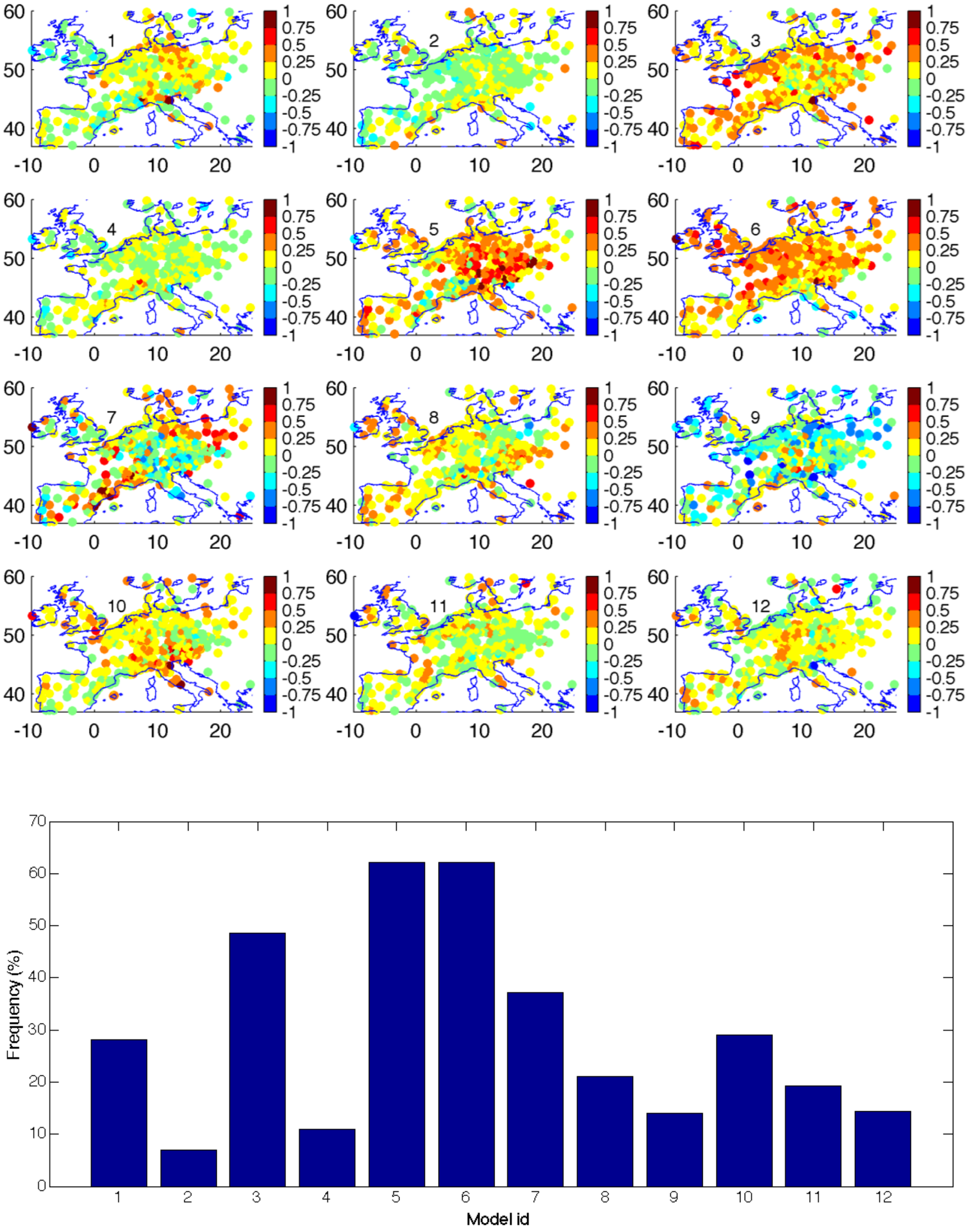

Figure 8: Ozone spatial weights (mmeW) calculated for each model (1-12) for JJA (1 segment) at the observed rural sites (segment de-bias) and aggregated frequency of model use in mmeS, from all the 451 stations for the test dataset.

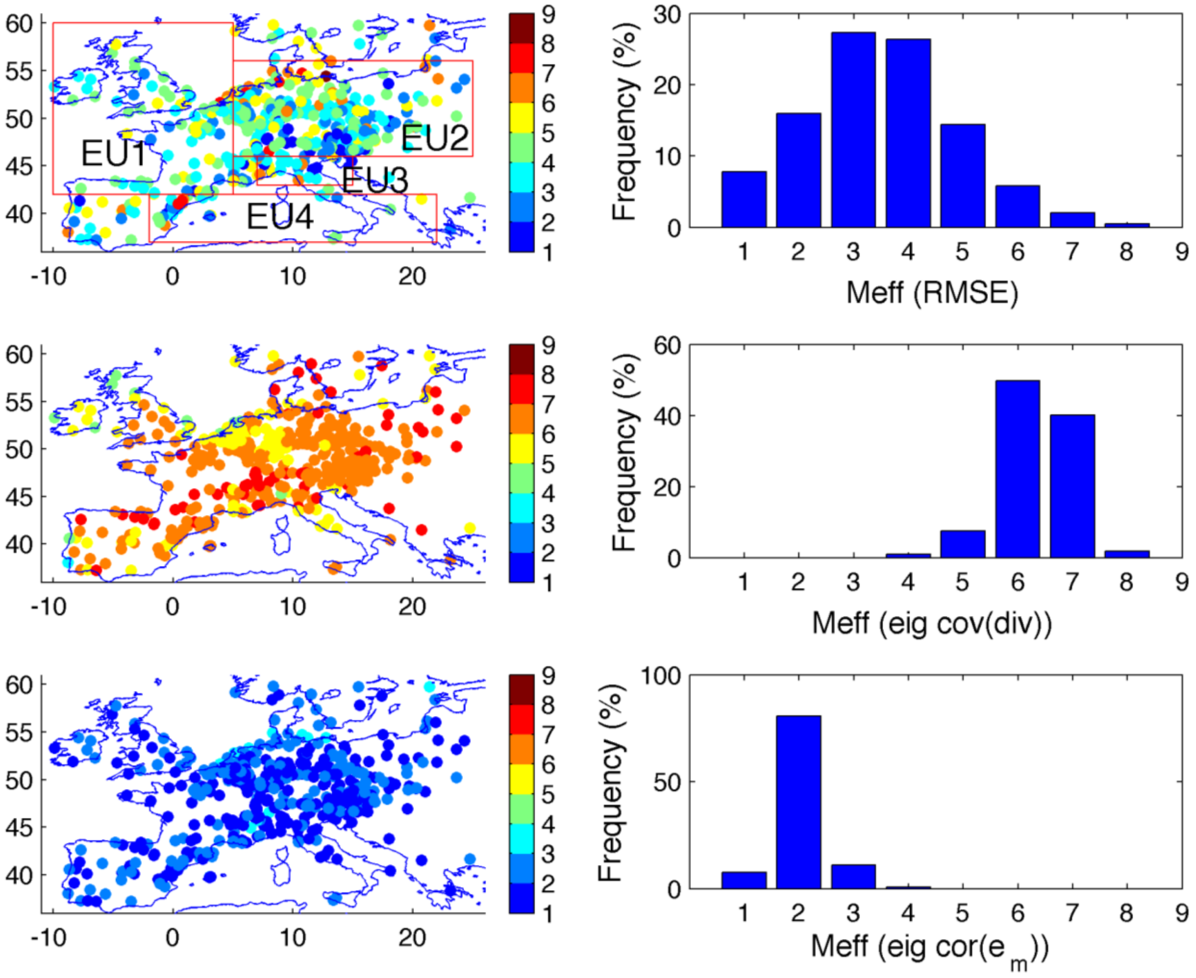

Figure 9: [Top] Spatial distribution of M_{eff} based on minimum error combination (left) and its histogram (right). [Middle] like top but for M_{eff} based on the eigenvalues of the covariance of the diversity matrix. [Bottom] like top but for M_{eff} based on the eigenvalues of the cor(e_i,e_j) matrix. Please note the different range of the y-axis between the different histograms.

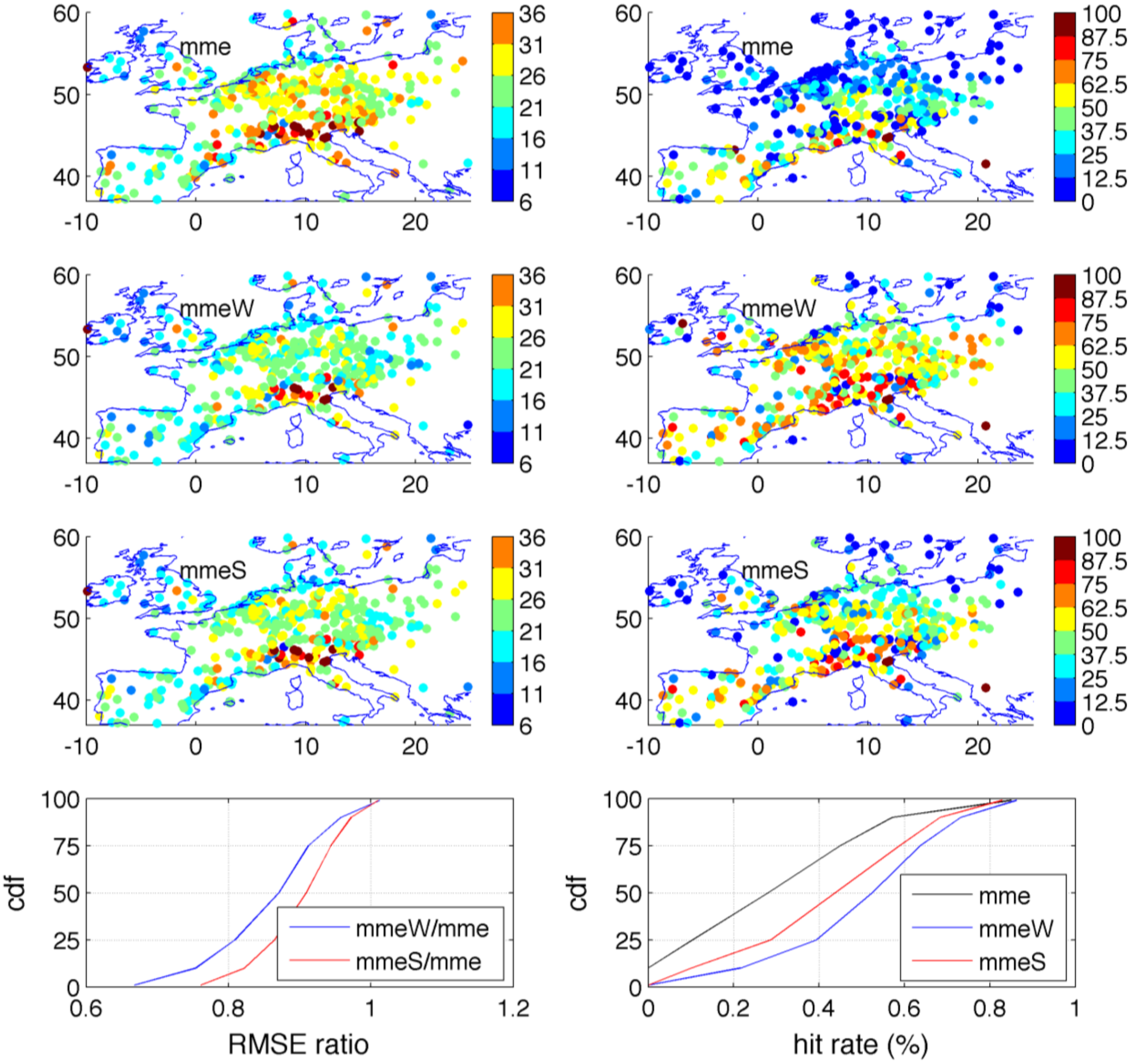

Figure 10: [Top row] The RMSE of ozone at each observed site for mme (left). The behavior at the upper tail of the distribution; percentage of correct hits for events > 120 µg/m3 for mme (right) [2nd, 3rd row] Like top row but for mmeW and mmeS. [Bottom row] The cdf of each spatial plot.

1

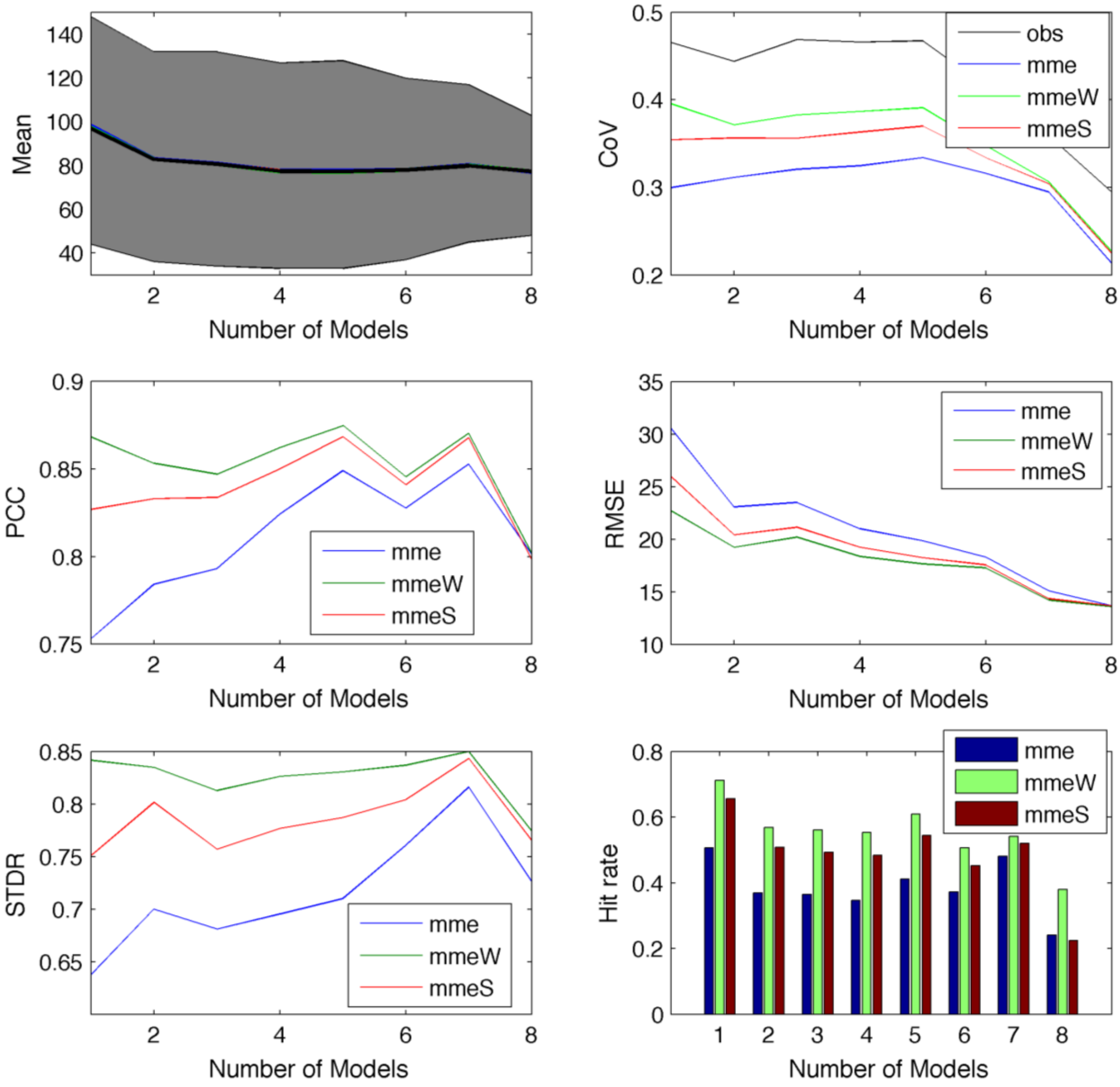

Figure 11: [Top] Statistical properties of mme, mmeS and mmeW forecasts versus observations from the 451 stations for the test dataset as a function of *Meff:* (a) mean and (b) coefficient of variation (StandardDeviation/Mean). The shadow area in the mean plot shows the 10th and 90th percentile of the observed concentrations. [Middle, Bottom] Forecast Skill of mme, mmeS and mmeW from the 451 stations for the test dataset as a function of *Meff*: (c) PCC, (d) RMSE, (e) STDR (standard deviation ratio) and (f) Hit Rate.

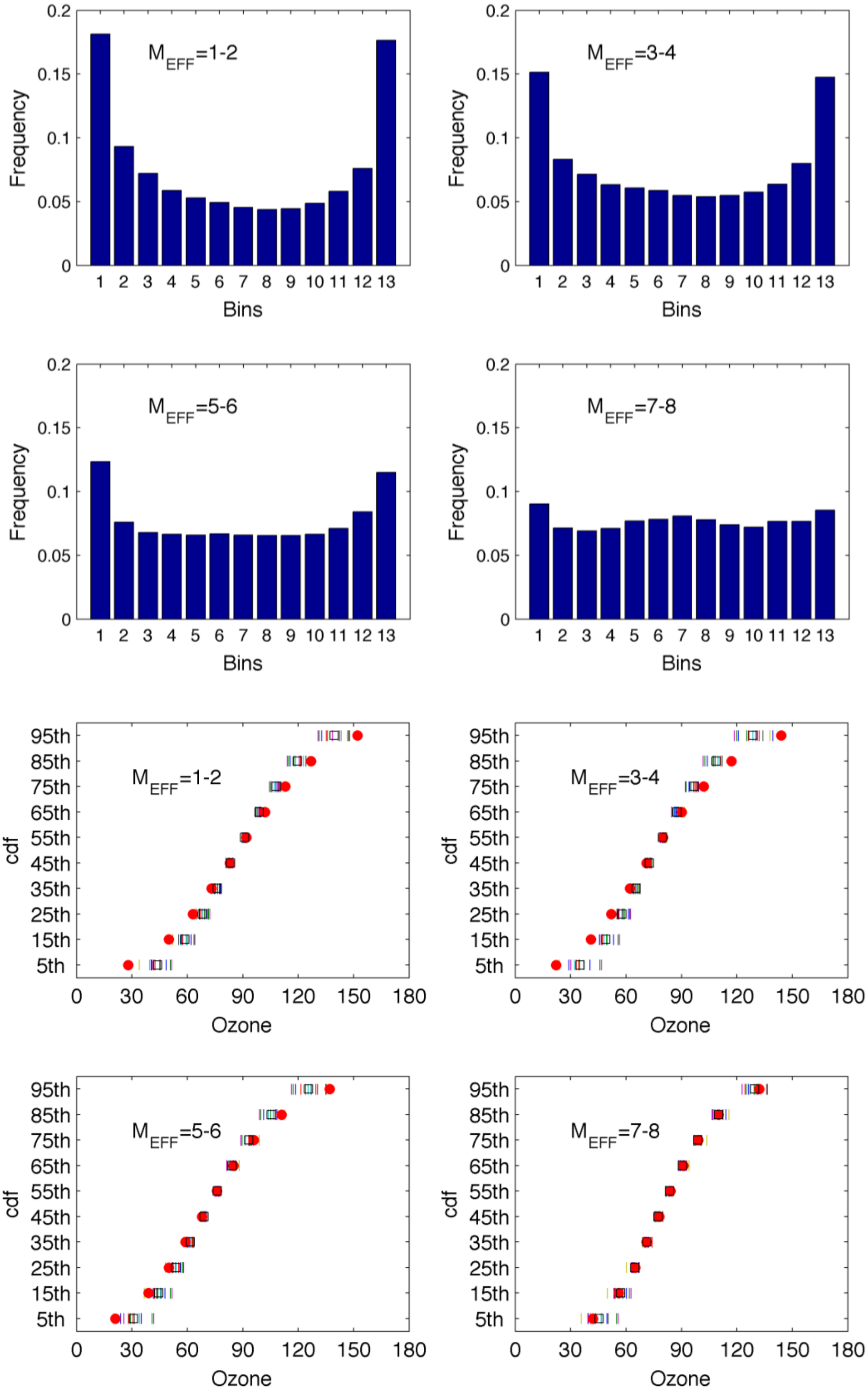

Figure 12: [Top] Talagrand diagram of the full ensemble aggregated at the stations as a function of M_{EFF}. [Bottom] Cumulative density function of observations (red circle) and models (coloured lines), aggregated at the stations as a function of M_{EFF}. The ensemble mean is displayed with a square.

1

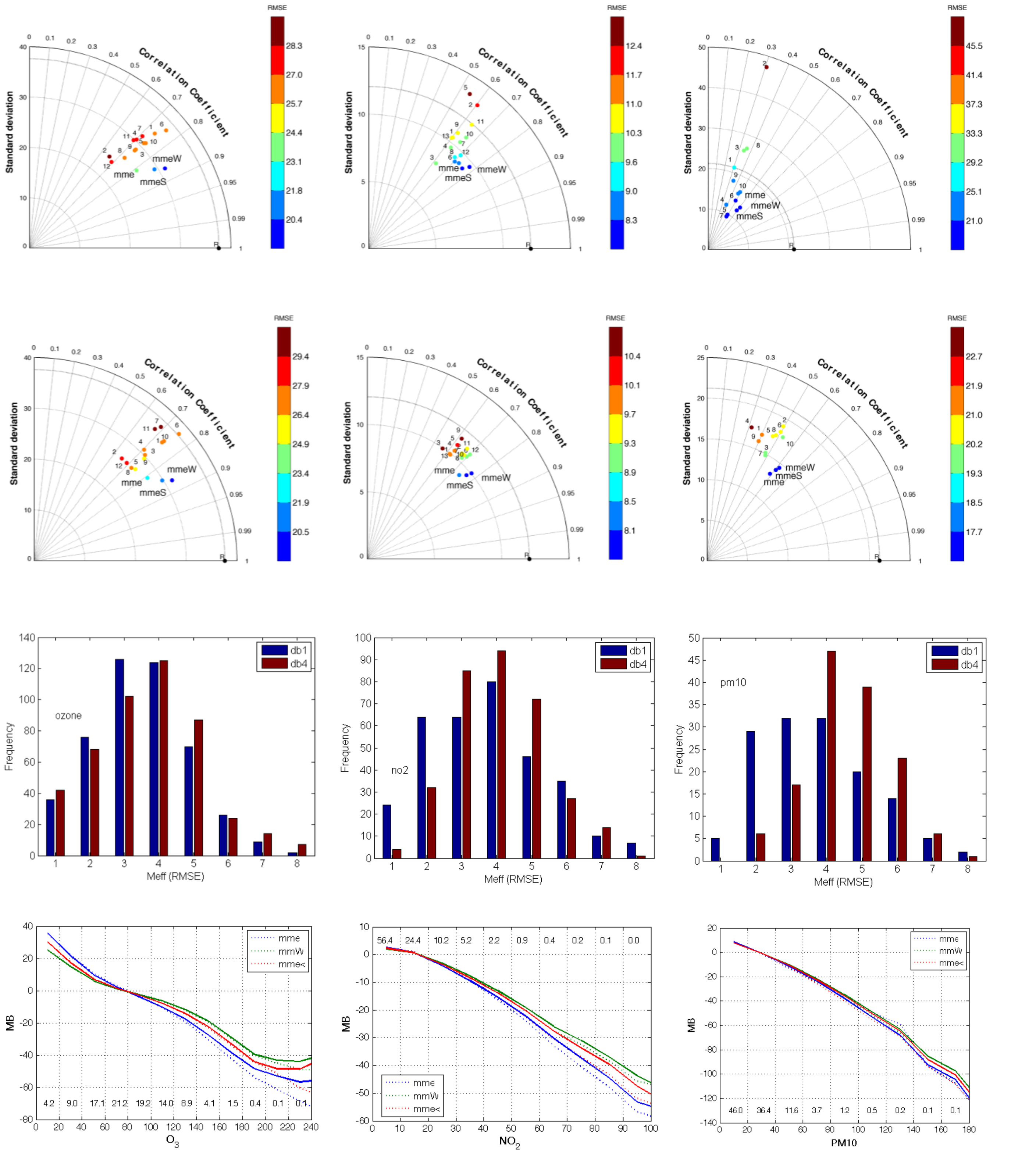

Figure 13: The Taylor diagrams on the first row refer to 1st order bias correction (db1) while on the second row refer to 2nd order bias correction (db4). The bar plot on the 3rd row show the distribution of the effective number of models in the two schemes. The line plots at the last row compare the 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, PM10). The plots have been produced from the aggregated time series incorporating all the stations of the test dataset. Please note the different range of the x and y-axis between the different panels.