Atmos. Chem. Phys. Discuss., 13, 4989–5038, 2013
www.atmos-chem-phys-discuss.net/13/4989/2013/
doi:10.5194/acpd-13-4989-2013

© Author(s) 2013. CC Attribution 3.0 License.

Atmospheric Chemistry and Physics
Discussions
Open Access

This discussion paper is/has been under review for the journal Atmospheric Chemistry and Physics (ACP). Please refer to the corresponding final paper in ACP if available.

Pauci ex tanto numero: reducing redundancy in multi-model ensembles

E. Solazzo[1], A. Riccio[2], I. Kioutsioukis[1], and S. Galmarini[1]

[1]European Commission, Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy
[2]Department of Applied Science, University of Naples “Parthenope”, Napoli, Italy

Received: 9 January 2013 – Accepted: 9 February 2013 – Published: 21 February 2013

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

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

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◂I | ▸I
◂ | ▸
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Abstract

We explicitly address the fundamental issue of member diversity in multi-model ensembles. To date no attempts in this direction are documented within the air quality (AQ) community, although the extensive use of ensembles in this field. *Common biases and redundancy* are the two issues directly deriving from lack of independence, undermining the significance of a multi-model ensemble, and are the subject of this study. Shared biases among models will determine a biased ensemble, making therefore essential the errors of the ensemble members to be independent so that bias can cancel out. Redundancy derives from having too large a portion of common variance among the members of the ensemble, producing overconfidence in the predictions and underestimation of the uncertainty. The two issues of common biases and redundancy are analysed in detail using the AQMEII ensemble of AQ model results for four air pollutants in two European regions. We show that models share large portions of bias and variance, extending well beyond those induced by common inputs. We make use of several techniques to further show that subsets of models can explain the same amount of variance as the full ensemble with the advantage of being poorly correlated. Selecting the members for generating skilful, non-redundant ensembles from such subsets proved, however, non-trivial. We propose and discuss various methods of member selection and rate the ensemble performance they produce. In most cases, the full ensemble is outscored by the reduced ones. We conclude that, although independence of outputs may not always guarantee enhancement of scores (but this depends upon the skill being investigated) we discourage selecting the members of the ensemble simply on the basis of scores, that is, independence and skills need to be considered disjointly.

cc BY

1 Introduction

Geophysical modelling nowadays relies, among other techniques, on ensemble methods to improve predictive skills, assess performance, and quantify uncertainties. This is particularly the case for atmospheric sciences, where climate and air quality models are often treated as ensembles of an arbitrary collection of models results belonging to the same family, sharing similar structure and resolution ("ensembles of opportunity", as defined, e.g. by Tebaldi and Knutti, 2007). Just like human beings normally consult a number of sources prior to making a decision (see for example the "trillion dollar garden party" analogy adopted by Knutti, 2010), the advantage of treating the information from several sources into ensembles relies on the fundamental assumption that information coming from multiple sources allows an estimation of the quality of the former, in line with the "Principle of multiple explanations" proposed by the Greek philosopher Epicurus (341 BC–270 BC) which says that for an optimal solution of a concrete problem we have to take into consideration all the hypotheses that are consistent with the input data. The fundamental aspect that makes multiple estimations a better one is the fact that the sources of the former must be independent. In our view, multi-model (MM) ensembles practices, as they have been developed over the years, lack of this fundamental consideration due to the fact that models are *phenotypically similar* (Potempski and Galmarini, 2009) and need therefore caution in their applications. Serving the scope of removing misconceptions and ambiguous interpretations of the paper, we define here:

- *Independence*, a formal property, when the joint Probability Distribution Function (PDF) of two or more model results is derived from the product of single PDFs (Cover and Thomas, 2006). This is the rigorous definition of independence, though the joint PDF is difficult to estimate in practice;

- *Un-correlation*, referring to the situation when model's outputs are linearly independent. This is the most applied proxy to independence. The outputs of independent models are un-correlated, but un-correlation does not guarantee independence;

- *Diversity,* a qualitative property. Models are said diverse when they are developed starting from different conceptual basis and are based on different causal assumptions. Their outputs (and errors) can be correlated. Proving diversity has the same practical difficulties than proving independence (in general models are the numerical coding of fundamental physical processes, and it is likely that all models have at least this in common). *Similarity* is the opposite of diversity, and can be defined when models are developed from the same conceptual basis or share a number of elements that make them similar. Outputs and errors of similar models are expected to be highly correlated;

- *Redundancy*, when two or more models, dependent or not, have correlated outputs. It is more informative than correlation as redundancy is related to the amount of explained variance (Legendre and Legendre, 1998). In the case of mutual correlations of model pairs, the redundancy reduces to the coefficient of determination, R^2, the square of the correlation coefficient. Redundancy is the primary effect of model *similarity*, and applies to both model outputs and their errors.

The lack of independence of members in ensemble treatment is not at all new. Despite the empirical evidence of the superior performance of average of models in some cases (Fiore et al., 2009; van Loon et al., 2007; Vautard et al., 2009; Pierce et al., 2009; Galmarini et al., 2004; Potempski et al., 2008), it is known that models share similar deficiencies. Several studies have demonstrated that similarities of model errors are statistically significant beyond doubt, thus questioning the effectiveness of "blindly" combine models into ensembles. Nonetheless, the problem of member (and error) similarities has received little attention by the climate modelling community, as recently recognized by Pirtle et al. (2010), and even less by the air quality community, where the theoretical work by Potempski and Galmarini (2009) and the attempts by

Riccio et al. (2012) and Solazzo et al. (2012b) remain the only studies, to the best of our knowledge, dedicated to the issue.

Independence of models can be sought in the form of different structures (proportion of parameterisations shared by the models), or, from an information science point of view, as the possibility to express the combined error PDF in terms of product of single PDFs (Abramowitz, 2010; Potempski and Galmarini, 2009). Ideally, perturbation of model parameters and associated uncertainty on model output could serve this scope, as suggested by Tebaldi and Knutti (2007), but this is often impractical. The strategy common to the (few) studies that directly investigate model diversity consists in attributing model independence only from the analysis of the output they produce. In particular, Potempski and Galmarini (2009) showed that, by relaxing the condition of model independence to that of model *associativity*, a robust theoretical framework could be built from which precise mathematical formulations could be drawn. *Associativity* is measured by the covariance or by the correlation of pairs of model outputs. However, caution is needed as "*it is possible that two models could agree with respect to outputs despite being based on different casual assumption*" (Pirtle et al., 2010). Thus, when looking at the correlation of model outputs as metric for defining independence, different models producing the same output would be erroneously considered as dependant. Further to that, the similarity of the results from two independent models is a valuable information that tells about model accuracy and uncertainty. Un-correlation of the outputs is a necessary but not sufficient condition to guarantee independence.

In the impossibility of an a-priori assessment of ensemble members' independence, model biases are excellent parameters to investigate the ensemble member interdependence. Models are intrinsically wrong due to their numerical nature, imprecise input data and limited understanding of the atmospheric chemical-physical processes. What is important is that models have independent systematic errors so the biases cancel out when combining models into ensembles. Should that not be the case, combining more and more similar models into an ensemble is not a solution, results do not improve! Moreover, a MM ensemble for which all biases have the same sign and value

cc BY

may give the false impression of accuracy, which is often confused with precision. The agreement of models to precisely predict the same (biased) result is confused with accuracy of models which implies homogeneously distributed biases around measurements (Potempski and Galmarini, 2009).

A further fundamental aspect is that of the uncertainty of the measurements used to calculate the bias, to evaluate the models, and/or to weight the ensemble members. Annan and Hargreaves (2010, 2011) have shown that, due to the uncertainties in the measurements, model's deviations from the observations can be strongly correlated, even in case of independent models. Thus, the independence of models does not necessarily translate into independence of their deviations from observations. At the same time, similar models have correlated errors but correlation of errors does not imply similarity of models. Furthermore, the conceptual assumption that models are drawn from a distribution centered around the truth (meaning that measurements and model output are not biased, or biased-corrected) might lead to wrong conclusions as recently reported by Annan and Hargreaves (2010, 2011).

To date, the link between ensemble accuracy and diversity of members is still unclear, the reason being that there is no unique way to decompose the ensemble's error in terms of bias and variance (Potempski and Galmarini, 2009). This is the fundamental problem of the ensemble techniques, whose error obeys to the bias-variance-covariance decomposition (e.g. Brown et al., 2005). The trade-off between bias and variance involves indeed three terms, and there is no way to simply minimize the co-variance without affecting other component of the error. The error of the ensemble mean increases linearly with the correlation of the members through the co-variance term. Techniques promoting diversity (or penalizing commonalities) do exist (negative correlation and other, Liu and Xao, 1999) and are an active area of research in the field of information science, though they are not the goal of this study.

The Latin expression *Pauci ex tanto numero* is extracted from the De Bello Gallico (The Gallic wars, book 7, chapter 88) by G. J. Ceasar (100 BC–44 BC) and refers to the battle of the roman army against the Gauls. The complete citation reads "*pauci* *ex tanto numero incolumes se in castra recipient*" and that translates "*few [Gauls] from a large number returned safely back to the camps*". More peacefully we decided to take the first part of the citation to stress the fact that only few from a multitude of models will be the ones that will make the ensemble result and will metaphorically survive the treatment in the end. Unfortunately the Gauls shared a different destiny.

The paper is structured as follow. In Sect. 2 the scopes are highlighted and the dataset and methodology are presented. In Sect. 3 we introduce an appropriate metric that allows detecting similarities beyond the overarching ones, and use this metric to quantify the level of redundancy of the dataset. To reduce the redundancy we then apply several techniques of dimension reduction (Sect. 4), which serve the scope of identifying the minimum number of elements necessary to explain the variance of the observational data. Once the dimension of the minimum set is established, we apply a number of member selection criterion (Sect. 5). The methods of member selection have the purposes of identifying the members (or the weights) that (i) have poorly correlated errors (thus non- redundant) and (ii) whose ensemble mean is skilful in terms of accuracy and precision. Conclusions are drawn in Sect. 6.

2 Scopes, data and method

We want to address here some fundamental questions: to what extent an ensemble of different models put together on the grounds of opportunity and convenience is indeed producing a better result? How can one quantify the information in MM ensembles that is necessary and relevant for the final result? Answers to these questions were already anticipated by Potempski and Galmarini (2009), where the angle of attack was more on whether the composition of the ensemble could be investigated a-priori. A theoretical framework and conditions were indeed identified but cannot be put in practice for all cases. Solazzo et al. (2012b) clarified the necessity of a posterior screening of the data and heuristically identified a possible methodology. In this paper we analyse various techniques available to address the following issues:

cc BY

1. Determine the ensemble redundancy: i.e. the minimum set of members that explains the variance of the observations and maximise the accuracy;

2. Reduce the ensemble redundancy: if two models, or their errors, are highly correlated one can be expressed in terms of the other by a simple scaling factor. If many redundant models are combined together, there would be loss of valuable information due to dependant biases.

The fundamental idea is thus to investigate different methodologies viable to achieve the above mentioned goals and verify the pros and the cons of each of them trying to produce a generalized concept and methodology applicable to any other kind of ensemble and variable.

For our analysis we will investigate the correlation between errors produced by AQ models run by twelve groups in the context of the Air Quality Modeling Evaluation International Initiative (AQMEII) (Rao et al., 2011; Galmarini et al., 2012). For all of the analyses presented in this study we use hourly time series for the months of June July August (JJA) of the year 2006 of the gaseous species of O_3, CO, NO_2, SO_2. We apply the analysis to two distinct regions of Europe:

- region 1 (−10, 5)° W; (42, 60)° N, including the UK, France, northern Spain and Belgium;

- region 2 (5, 24.5)° W; (46, 60)° N, the continental Europe, including Germany, Poland, Austria, and Czeck Republic.

The modelled and observed time series have been spatially averaged over the region 1 and 2 defined above. These two regions have been the subjects of in-depth investigation in other AQMEII studies (Solazzo et al., 2012a,b; Vautard et al., 2012). The number of receptors – by species – in each region is reported in Table 1. The participating models are summarized in Table 2. Details about the model settings and operational evaluation against observational data can be found in Solazzo et al. (2012a,b) and Vautard et al. (2012), with the exception of the GEM-AQ model (Côté et al., 1998; Kaminsli

cc BY

et al., 2008), which did not take part of the previous AQMEII analysis. The AQMEII ensemble of models is indeed a genuine *ensemble of opportunity*, with a good level of diversity in terms of AQ models and meteorological drivers. Emission and boundary conditions are, however, largely shared, making the distribution of model errors neither systematic nor random. The history of regional scale modelling has also forcibly produced a number of common elements to all the models, which should be considered an a-priori contaminating element of the ensemble results.

Since an accurate estimation of multivariate a PDF is hard to achieve due to the computational cost it entails even for a small number of models (Peng et al., 2005), we decide to focus on quantifying the amount of information two models share, measured by the redundancy which can be computed more easily in this case. Given the output from two models, $\boldsymbol{x},\boldsymbol{y}$ organized as a two-columns table, said $\Sigma_{\boldsymbol{xy}}$ their covariance and $p(\cdot)$ their joint PDF, the redundancy can be defined either through the redundancy index $\rho I(\boldsymbol{x},\boldsymbol{y})$ (Stewart and Love, 1968), which is a metric for quantifying the portion of variance already being accounted for by other members of the ensemble (Eq. 1), or by the mutual information among models $I(\boldsymbol{x},\boldsymbol{y})$ (Peng et al., 2005; Ding and Peng., 2005) (Eq. 1):

$$\rho I(\boldsymbol{x},\boldsymbol{y}) = \frac{\mathrm{trace}(\boldsymbol{\Sigma}_{\boldsymbol{xy}}\boldsymbol{\Sigma}_{\boldsymbol{yy}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{yx}})}{\mathrm{trace}(\boldsymbol{\Sigma}_{\boldsymbol{xx}})} \tag{1}$$

$$I(\boldsymbol{x},\boldsymbol{y}) = \iint p(\boldsymbol{x},\boldsymbol{y})\log\frac{p(\boldsymbol{x},\boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}\mathrm{d}\boldsymbol{x}\mathrm{d}\boldsymbol{y} \tag{2}$$

Eq. (1) is related to the prediction of $\boldsymbol{x}$ by $\boldsymbol{y}$ by multiple linear regression. $\rho I(\boldsymbol{x},\boldsymbol{y})$ is a weighted average of the squared multiple correlation coefficient between all pairs of variables of $\boldsymbol{x}$ and $\boldsymbol{y}$. It is a measure of the quality of the prediction of $\boldsymbol{x}$ by $\boldsymbol{y}$ and represents the proportion of explained variance in the regression of $\boldsymbol{x}$ by $\boldsymbol{y}$ (see e.g. Youness and Saporta, 2010). In the case of $\boldsymbol{x}$ and $\boldsymbol{y}$ one-dimensional vectors ρI returns R^2, the squared correlation coefficient. The mutual information in Eq. (2) is more complex and involves the PDFs of multivariate variables. In practical terms I is the level of repetition of two datasets, and the PDFs are computed as the frequency of unique elements belonging to both $\boldsymbol{x}$ and $\boldsymbol{y}$. Details about the implementation of Eq. (2) are given in Peng et al. (2005) and Yoon and Kim (2009).

3 A metric for model similarities and comparability of errors

Common biases are difficult to detect, especially for AQ models where the variance of the noise can be comparable with that of the signal, in particular for low concentrations. The AQMEII database includes results from ensemble members sharing meteorological drivers, emissions, chemical boundary conditions (Table 2). It was proven that these input fields introduce systematic biases in the model results (Solazzo et al., 2012a,b). A simple error metric would not be adequate to detect any type of underlying commonality other than these overarching biases. For this reason we have to find out a metric that (a) explores hidden similarities, i.e. those underlying common modules and parameters in the model, and that (b) is robust enough to be used under a number of scenarios. Having in mind that no wonder metric exists and that different metrics produce different results (Gleckler et al., 2008), we opted for the metric d_m proposed by Pennel and Reichler (2011) (hereafter referred to as PR2011), which explores the biases of models and removes from each model the dominating similarities, thus making individual model errors more dissimilar and unveiling “hidden” trends that are masked by overarching commonalities.

Let us start by defining the standardized deviation of models (mod) from observations (obs) for the species of as:

$$e_{i,m,\mathrm{s}} = \frac{\mathrm{mod}_{i,m,\mathrm{s}} - \mathrm{obs}_{i,\mathrm{s}}}{\sigma_\mathrm{s}} \tag{3}$$

where σ_s is the standard deviation of the observed chemical species and $i = 1, \ldots, N$ is the index of the time series, m is the model index and s that of the species being considered (O_3, CO, NO_2, SO_2). The normalisation in Eq. (3) makes more comparable

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀I ▶I
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

the errors for different chemical species and units. We now define the MM-error pattern (MME) as

$$\mathrm{MME}_{i,\mathrm{s}} = \frac{1}{M}\sum_{m=1}^{M} e_{i,m,\mathrm{s}} \tag{4}$$

which contains the “bulk” of bias among models. To eliminate the dominating model similarities, we remove from all model’s errors the portion of MME associated with each individual model error. Accordingly to PR2011, the removal of the portion of MME relevant to an individual model can be accomplished by calculating the difference d_m between the standardised model error and the weighted standardised MME, with the weight being the correlation coefficient R between the m-th model error and MME:

$$d_{m,\mathrm{s}} = e^*_{m,\mathrm{s}} - R \cdot \mathrm{MME}^*_{\mathrm{s}} \tag{5}$$

where the “*” indicates *standardised* vectors, calculated, for each time series, by subtracting the corresponding mean value $\overline{e_m}$ and dividing by the standard deviation σ_{em} (we have now get rid of the index i for a more compact notation). Note that the normalisation serves the only purposes of making the results for different species inter-comparable, as the correlation is bias- and normalisation-independent. Also note that the normalisation makes the correlation and covariance interchangeable. As said above, removal of MME makes model errors more dissimilar and uncovers “hidden” trends that are outweighed by overarching commonalities. For example, corr($e^*_{\mathrm{FR3,O3}}$, $e^*_{\mathrm{FI1,O3}}$) = 0.73, while corr($d_{\mathrm{FR3,O3}}$, $d_{\mathrm{FI1,O3}}$) = 0.36. The subtraction of the correlated portion of the bulk error from the individual error emphasizes the real differences among models. On the other hand, in the case of two highly similar models, such as DE2 and US4 the correlation among e^*_i is approximately the same as that among the d_i.

We provide two graphical examples of the efficacy of d_m vs. e_m. The correlation between individual model error and the MME (corr(ei, MME)), averaged over all models as a function of variable is reported in Fig. 1. The correlations are highly positive, demonstrating the extent of large commonalities, and also show dependence on the region (correlations for SO_2 are quite different over the two regions). In Fig. 1 the correlation corr(di, dj) is also shown, averaged over all model pairs. The values for the curves for the two regions are very similar and small, indicating that the effect of MME has been largely mitigated. By removing the MME, model errors become region-independent, as shown by the similar curves of corr(di, dj). In Fig. 2 we report the associativity tree (the dendrogram, see details in Sect. 4.4) of cov(di, dj) and cov(ei, ej) for the joint time series of the four pollutants in region 2. While e_m-associations are based on the species (model errors for each species are the most correlated), d_m-associations are drastically diverse, and unexpected patterns emerge. Models are grouped by the bias underlying modules and/or parameters strictly associated with the physics and chemistry of a given compound; the diversity for d_m is higher with respect to the e_m-dendrogram and the number of disjoint clusters is, at least, of six (distance level of ∼0.9), while four e_m-clusters were identified (at an even smaller distance of ∼0.7).

3.1 Ensemble redundancy through error analysis

In Fig. 3 a graphical representation of the covariance cov(di,dj)is shown for the species of the European region 2 (plots for region 1 are omitted for brevity). Mutual model covariance is indicated by the positioning of the model codes in black with respect to models on the horizontal axis. Because the covariance matrix is symmetric, we display only half of it, for clarity. The model codes in red indicate the variance (cov(di, di)). In these plots we also report

- the redundancy measured by R^2 (blue crosses), the square of corr(d_i, d_j). R^2 represents the amount of variance already explained by the regressor model and, for model pairs, corresponds to the redundancy index ρI;

- the mutual information I (vertical segments in orange).

Because of the normalisation of the metric d_m, the covariance and redundancy can be expressed on the same scale, between −1 and 1.

Depending on the species, the mutual relationships among members vary greatly, proving that for AQ models many factors (chemistry and dispersion modules, meteorology, grid resolution) contribute to the final outcome. This was also found to be the case for climate models (PR2011; Annan and Hargreaves, 2010). Overall, errors do not seem to co-vary more in the case of two instances of the same AQ model (DE3 and UK2 for example) than for different combination of meteo-dispersion models (FR4, DE1 for O_3 and CO; HR1 and UK2 for SO_2; and many others). The sharing of routines specifically designed for certain pollutants and process could be a possible cause for this. It is often the case that model developers borrow entire model components as their use was demonstrated to be an improved, or sometimes the only, solution for simulating a process. For example, the ISORROPIA module (Nenes et al., 1998) for inorganic pollutants, the resistive scheme by Zhang et al. (2001) for dry deposition, the scavenging parameterisation for wet deposition are all examples of shared routines among the majority of the AQMEII models (see Table 1 of Solazzo et al., 2012a).

Because the redundancy measured by R^2 is simply the ratio of the squared covariance to the variance, models with a large spectrum of covariance are also the more redundant (DK1 and DE1 for O_3; US3 and US4 for CO; DK1 and DE3 for SO_2; NL1, DE2, US4 for NO_2). The redundancy measured by the mutual information is often in line with that of R^2, although in some cases higher values are estimated. For example DE2 and US4 (same models run by different groups), but also US3 for CO and NO_2, FR4 and PL1 for SO_2, due to I being calculated as a raw frequency count, whilst R derives from a regression analysis.

A further aspect of error redundancy is the amount of the observed variance explained by the MM ensemble. Following the methodology proposed by Annan and Hargreaves (2011) we projected the observation anomalies onto the principal components (PCs) of the covariance matrix of the deviation of the ensemble of models around the MM mean. We found that just the first (or the first two for ozone) component already

CC BY

exceeds the observed variance. When all components are taken into account, it results that the MM mean for the EU region 1 (region 2) can explain as much as 1.2 (1.7), 2 (4.8), 2.1 (9), 7 (18), times of the observed variability for O_3, CO, SO_2, NO_2, respectively (the large difference between region 1 and 2 for NO_2 and SO_2 is due to the much smaller variance of the observed values of these two compounds in region 2 (~ 4 and 12 times smaller for NO_2 and SO_2, respectively)). This case is depicted as a "wide ensemble" by Annan and Hargreaves (2010). A wide ensemble can be interpreted also in terms of lack of reliability with a rank-histogram (Talagrand et al., 1998) exhibiting a "central dome" pattern: the ensemble performs poorly in predicting less frequent episodes (both high and low concentrations) and lacks sharpness. Given the massive application of AQ models in regulatory applications and the more and more stringent AQ targets, the detected overconfidence can cost considerably. Dealing with a wide ensemble implies that there is a substantial amount of redundant variability already accounted for by other models. One plausible explanation is that the ensemble size, constrained by the available members, is simply too large.

4 Quantifying ensemble redundancy through dimensionality

The foremost advantage of reducing the dimensionality of large datasets by discarding redundant information is that lower dimensionality means reduced computational costs and noise, improving the accuracy of the ensemble. Data mining and data reduction are active areas of research in various fields, from genetics to ecology to machine learning. There exist a plethora of methods aiming at detecting commonalities, most of which developed ad-hoc for a specific application, such as Independent Component Analysis (Kong et al., 2008), Maximum-Relevence-Minimum-Redundancy (Peng at al., 2005), the methods reviewed by Grömping et al. (2007), and others. Though, is seldom the case that a method developed by a community passes the barrier to be adopted in a field other than the one it was originally developed for.

Here we explore some analytical techniques proposed in various flavours in the climate modelling community. Note that the outcome of dimension-reduction methods is simply a number, that is, the dimension of the subspace sought. Selecting the members belonging to that subspace is a different problem, and is addressed in Sect. 5.

4.1 Eigenvalue methods

We calculated the effective number of models (also known as the effective number of Degree of Freedom) sufficient to reproduce the variability of the full ensemble (the MM ensemble generated with all available members) as:

$$M_{\text{eff}} = \frac{\left(\sum_{k=1}^{M} \lambda_k\right)^2}{\sum_{k=1}^{M} \lambda_k^2} \tag{6}$$

with λ eigenvalue of the corr(d_i, d_j) matrix. Theoretical derivation of Eq. (6) can be found in Bretherton et al. (1999). Under the assumption that the modelled and observed fields are normally distributed, the fraction of the overall variance expressed by the first M_{eff} eigenvalues is of 86 % (Eq. 8 of Bretherton et al., 1999).

Results for M_{eff} are reported in Table 4. The sum over all eigenvalues at the nominator of Eq. (6) expresses all the variability that is attainable in a M-dimensional vectorial base of orthogonal vectors ($M = 13$). By construction $\sum_{k=1}^{M} \lambda_k = M$ and only if all eigenvalues were equal to unity, and in this case Eq. (6) would return $M_{\text{eff}} = M$, that is all directions are equally important. In reality there exist eigenvalues that are larger than unity, and consequently other that are less than unity, and since these are squared (denominator of Eq. 6) the contribution of the former outweighs that of the latter so that $M_{\text{eff}} < M$ approximately in the amount of the number of eigenvalues larger than unity (Guttman (1954) and Kaiser (1960) indeed proposed to adopt this as rule for

cc BY

determining the number of factors to retain, supposing that it makes no sense to retain components that explain less variance than the original standardized variables). Thus, we can think of replicating the full variability of the M-members by an M_{eff}-dimensional subset of these in a vectorial space whose base is generated by the eigenvectors of the largest eigenvalues. On the other hand, if all error fields were similar, only one eigenvalue would be non-zero and $M_{\mathrm{eff}} = 1$.

By applying Eq. (6) to the datasets of model errors ($\mathrm{corr}(d_i, d_j)$) we find that M_{eff} is in the range 5 to 6.5 (Table 3). If the MME term is retained (that is, M_{eff} is calculated from $\mathrm{corr}(ei, ej)$) we find much lower values for M_{eff} (Table 3) as consequence of most of the similarity among models being expressed by the MME term.

4.2 Principal Components Analysis (PCA)

PCs analysis (PCA) (Jolliffe, 2002) is probably the most well known and wide-spread unsupervised dimension-reduction technique. It is based on eigenanalysis to select uncorrelated directions associated with the largest variances. Relationships between PCA and clustering (Ding and He, 2004), redundancy (Jolliffe, 2002), Multi-Dimensional Scaling (Groenen and van de Velden, 2004), and regression analysis (Jong and Kotz, 1999) have been documented, proving the versatility of this method. For example, the ratio of the sum of leading eigenvalues to the sum of all eigenvalues obtained by means of PCA is proportional to the ratio of the regression sum of squares ($\mathrm{SS}_{\mathrm{reg}}$) (explained or signal variance) and the total sum of squares ($\mathrm{SS}_{\mathrm{tot}}$) (the total variance) in regression analysis. This latter ratio is the coefficient of determination R^2, the redundancy index (Jun et al., 2008).

The relationship 6 provides an analytical estimate of the dimensionality of the subspace of models to produce the information of the whole ensemble. Graphically, the "scree test" (Cattell, 1966) is often applied in problems of dimension reduction. We first produce a plot of the number of dimensions vs. quantities related to the amount of variability or independence, measured by appropriate metrics. Then, we use the "elbow criterion" by seeking the point at which the curve levels off to a plateau. To produce

Discussion Paper | Discussion Paper | Discussion Paper | Discussion Paper |

ACPD

13, 4989–5038, 2013

Pauci ex tanto numero

E. Solazzo et al.

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀▮ ▶▮
◀ ▶
Back Close
Full Screen / Esc

Printer-friendly Version

Interactive Discussion

a scree plot from Eq. (6), we look at M_{eff} as a dependent variable of N, the number of models. Curves are reported in Fig. 4 for the four pollutants and the two European regions. The variability scale is calculated as the cumulative variability, for each N. As for Table 3, curves have been derived from corr(d_i, d_j) and from corr(e_i, e_j). We notice that in both sub-regions M_{eff} from corr(e_i, e_j) is much lower and that variability above 80 % is reached by the first 2–3 leading eigenvalues. As noted by PR2011, the concavity of the curves over N indicates that the addition of more models to the ensemble is not compensated by a linear increase in the overall information. This is a straight consequence of commonalities among members: chances that a new member shares features with an existing one increases as the ensemble size does. This would not happen in the case of independent models.

4.3 Multi-Dimensional Scaling (MDS)

Another method, among the many, to create a scree plot is to use Multi-Dimensional Scaling (MDS) algorithm (Borg and Groenen, 2005) for determining the relationships between model errors. MDS basically searches for a spatial configuration of the objects such that the mutual Euclidean distance among them matches their proximities as closely as possible. Here, we use the corr(di, dj) matrix as proximities. The degree of correspondence between the distances among points implied by MDS map and the input matrix is measured by a suitably defined *stress* function, the minimisation of which also provides information about the dimensionality of the subspace covering the whole variability of the data. Avoiding detailing too much, in MDS theory the Euclidean distance s_{ij} between two rows of a matrix $\mathbf{X}$ is defined as:

$$s_{ij} = \left(\sum_{k=1}^{p} \left(x_{ik} - x_{jk} \right)^2 \right)^{1/2} \qquad (7)$$

The objective of MDS is to find the elements of $\mathbf{X}$ minimising the difference between s_{ij} and d_{ij} (the elements of the proximity matrix corr(d_i, d_j)):

$$\sigma^2(\mathbf{X}) = \sum_{i=2}^{n}\sum_{j=1}^{i-1} w_{ij}(d_{ij} - s_{ij}(\mathbf{X}))^2 \tag{8}$$

σ^2 is the raw stress function (with w_{ij} non-negative weights, set to unity). Minimisation of the stress function is not trivial, and thus numerical iterative methods are employed (Borg and Groenen, 2005). By running the minimisation problem for different values of p in Eq. (8), we plot the stress against the dimension. Results for the European region 2 are reported in Fig. 5 (results for region 1 are very similar and therefore not shown). The "elbow" in the scree plot indicates when more dimensions only yield a negligible improvement in terms of stress. The trend of the curves in Fig. 5 (similar for all pollutants) indicates four as the number of independent components that best fit the data, i.e. about one third of the whole sample size.

4.4 Hierarchical clustering (HC)

Given a data set of M instances $\mathbf{X} = \{X_1, X_2, \cdots, X_M\}$, a clustering algorithm generates m disjoint clusters based on a distance metric, represented as $\Pi = \{\pi^1, \pi^2, \cdots, \pi^r\}$. Each clustering solution π^i is a partition of the data set $\mathbf{X}$ into K^i $(i = 1, \ldots, r)$ disjoint clusters of instances, represented as $\pi^i = \{c_1^i, c_2^i, \ldots, c_{K^i}^i\}$, where $\cup_k c_k^i = \mathbf{X}$ (Fern and Brodley, 2004). A typical output of HC is a dendrogram or associativity tree, where redundant models are gropued together and the level of similarity among groups is reported based on the distance between the elements of the input matrix. Here, we use the Euclidean as distance metric and the corr(di, dj) as input matrix. Applications of HC and dendrogram representation for air quality ensemble modelling are documented in Riccio et al. (2012) and Solazzo et al. (2012b).

A fundamental challenge of the HC method is that different grouping is obtained by slightly changing some of the options underlying the HC algorithms (Fern and Brodley, 2004). The agglomerative method, the distance metric, the number of clusters, and the cut-off distance are aspects that need to be determined case by case. In particular, the cut-off (the threshold similarity above which clusters are to be considered disjointed) determines the dimension of the sub-space of non-redundant models and is decided by visual inspection of the dendrogram. After numerous tests, in this study the unweighted pair-group average was selected as agglomeration method (Murtagh, 1984) with the cut-off value set between 0.10 and 0.15 (1 being the maximum similarity) for all pollutants in both regions, which produced five disjointed clusters (Figs. 6) for all species. The cut-off value is chosen by looking at the structure of the dendrogram: it is convenient to break structures that are obviously disjointed, and within each structure avoid separating highly connected groups, or groups of only two models. Looking at the dendrogram for ozone for example, the two main branches at the top further split into two more at a relatively low similarity level, suggesting a plausible way to proceed. At a ~ 10 to 15 % similarity level five clusters are detected for all species in both regions.

4.5 Comparing the different methods: discussion

Given the normalisation implied by the metric d_m, we found M_{eff} to range between 5.2 (ozone in region 2) and 6, with only NO_2 and CO in region 1 requiring 6.5 components (Table 3). M_{eff} based on d_m is between 1.5 (SO_2 region 2) and 5 (CO region 1) times higher than the values based on e_m (values in parenthesis in Table 3). The variability of M_{eff} among species depends on the heterogeneity of processes and sources within the two regions, as well as on the receptors coverage. Despite having removed the commonalities among models encapsulated by the MME, we still found a level of redundancy above 50 %, being M_{eff} less than half of the size of sample.

As said above, results of HC analysis indicates that at a ~15 % similarity level five clusters are detected. The between-classes variance (weighted average of the mean distance of each cluster and the mean distance of the whole dendrogram) detected by the five components generated by the HC method is between 70 % and 80 % of the total variance (depending on the variable), which would be totally reproduced only in the case of a cut-off level at the root of the dendrogram tree (one cluster only). On the other hand, the within-classes variance (average distance within each cluster) is an estimate of the redundancy, as it is proportional to the cluster-averaged coefficient of determination R^2 (Moesa et al., 2005). This result is in line of that obtained by applying PCA: M_{eff} in that case explained 86 % of the total variance (Sect. 4.2) with a slightly larger number of models. Thus the two techniques are consistent for similar amount of variance. Dimensionality through MDS and the minimization of the stress function has returned a number of components of four. In general though, MDS fit indexes are descriptive and do not always provide an absolute criterion for selecting the best dimensionality (Tinsley and Brown, 2000).

Summarising, the ensemble of models is highly redundant even after having removed the MM error. It is possible to reduce the full datasets of more than 50 %, down to five to six components. As discussed next, this allows reducing noise and improving accuracy. The methods adopted give consistent results.

5 Identifying the members of the reduced ensembles

As many as $\sum_{i=1,m} \binom{M}{i}$ subspaces with dimension smaller than m are identified by the M members (M is the total number of available members). It is therefore difficult to univocally identify a subset of members systematically outscoring all the others for a large number of skills. Furthermore countless methods for selecting members have been developed by different communities, testifying that available methods are "fit for purpose" rather than of general applicability. Ideally, once a skill or feature is identified, one could select the best performing ensemble by extracting from the group of M members only those m that are individually performing at best when compared with the measurements. However, combinations of individually good models do not necessarily produce a good ensemble for a given feature: the m best models are not necessarily the best m (Cover, 1974). Although the selection of members based on performance might be

CC BY

justified in some cases (e.g. McSweeney et al., 2012), we discourage this approach where no assumptions about the redundancy of the information is provided (Solazzo et al., 2012b).

A number of data reduction and member selection/weighting techniques exist that could be useful to our goal, some of which have been already discussed in Sect. 4 and that will be to adopted reduce the ensemble. Further to that we will compare the new ensembles to the full ensemble mean, taken here as benchmark. Member selection techniques applied are:

- Hierarchical Clustering (HC);
- Multi-Dimensional Scaling (MDS);
- minimization of the Mean Square Error (minMSE);
- Principal Component Analysis (PCA);
- Correlation-Adjusted (marginal) coRrelation (CAR).

Not all of the methods above take into account the redundancy of members. The first two (HC and MDS) provide ensembles of low-redundant members; the minMSE technique is a heuristic method based on the minimisation of the error and thus selection is skill-driven (Solazzo et al., 2012b; Riccio et al., 2012; Knutti et al., 2010); PCA provides weights to the models along the directions of maximum variance; finally CAR is a score-based member selection method developed by Zuber and Strimmer (2011) that is hybrid of marginal correlation and regression analysis and is shortly discussed in Sect. 5.5.1.

5.1 Hierarchical clustering (HC)

With reference to Fig. 6, members from each cluster are selected according to the individual model scores for bias: the model ranked best for bias, among the models of each cluster, was the one chosen to represent the cluster. The bias was selected since it is the metric underlying d_m, the metric used to build the dendrogram. The members are reported in Table 4. Other options have been taken into account, such as selecting the model closer to the centre of each cluster, and the model whose MSE with respect to the cluster's centroid is minimum. However, the reduced ensembles generated with these selection criteria were outperformed by that of members of minimum bias (see Sect. 5.5.2) and are therefore not shown.

5.2 Multi-Dimensional Scaling (MDS)

In MDS the distance among models can be used as proxy for independence, providing the visual aid needed for interpreting the grouping and selecting the more diverse member. MDS transforms the correlation between members into a reciprocal distance, allowing a visual inspection of the mutual model positioning into a two-dimensional plane. The reciprocal distance among members is the only information this methodology offers. Application of MDS for member selection in climate ensemble modelling can be found in Jun et al. (2008); the model space of Abramowitz (2010) is an extension of MDS, where the observations are treated as a de-facto model. Figure 7 summarises the mutual model distance in 2-D for the species of region 2.

5.3 Minimum error (minMSE)

Solazzo et al. (2012b) show that the ensemble mean minimizing the MSE has also superior skills with respect to the full ensemble, both in terms of variance and, of course, error. We ran similar analysis for the present datasets. Application of this analysis yields (i) the number of dimensions to retain (the dimension of the subset), and (ii) the members to retain (the component of the subset, reported in Table 4). Finding the subset of models minimising the MSE is a heuristic practice based on the evidence that a MM ensemble whose mean minimize the error is likely to have small covariance (from the variance-bias-covariance relationship, minimum error is more probable from un-correlated members, so that the covariance term is null). Knutti et al. (2010) and Annan and Hargreaves (2010) also explained the behaviour of the curves of MSE obtained by randomly sampling the ensemble of members. In particular, when the mean of the MSE distribution decays proportionally to $\sigma_{\text{obs}}(1+1/m)^{0.5}$, as in the present study, indicates that observations and model results are extracted from distributions having the same variance (the authors refer to this case as exchangeable or indistinguishable ensembles). Moreover, the fact that the MSE, no matter how large is M, can never reaches zero, is a consequence of the variability affecting the observations (from the error decomposition relationship, the variance of the observation is the lower bound for the error). Plot of RMSE for ozone (region 2) of the mean of random subsets of the ensemble members, plotted as function of subset size, is reported in Fig. 8 (for brevity, plots for the other species are omitted). The curves show the maximum, mean and minimum of RMSE. The dash-dotted curve decays as $m^{-0.5}$ that would be the trend is the models errors were independent (Knutti et al., 2010; Annan and Hargreaves 2010). For ozone we find a minimum for $m = 3$, where the RMSE is ~37 % smaller than the full ensemble mean. Adding more members to the ensemble increases the noise and deteriorates the accuracy. This would not happen if the model errors were independent as the curve in that case would decay monotonically.

5.4 Principal Components Analysis (PCA)

Although PCA cannot be applied for selecting individual, independent, members, it can be nonetheless used to generate an artificial time series mod_{PC} obtained by projecting the original data onto its PCs. This generates a weighted ensemble, the weights being the projections of the model components onto the eigenvectors associated to the leading m eigenvalues. We have applied PCA to the matrix of covariance $\text{cov}(d_i, d_j)$, to disclose redundancy patterns (see Sect. 4.2). The reduced matrix $\mathbf{dm}_{\text{red}}$ is obtained by projecting d_m onto PC_m, the subspace of the first m eigenvectors:

$$\mathbf{dm}_{\text{red}} = \mathbf{dm} \cdot \text{PC}_m \tag{9}$$

We have discussed the scree plot of Eq. (9) (Sect. 4.2) and the dimensionality of the subspace PC_m. Ideally now we should be in a position to score the weighted ensemble obtained by retaining m components. Getting the time series back from Eq. (9) is not trivial though, since d_m is a composite metric, and no similar applications of PCA have been found in the literature. The reduced time series is given by:

$$\mathrm{mod}_{\mathrm{PC}} = \sigma_{\mathrm{obs}} \left(\sigma_{\mathrm{em}} \left(\mathrm{dm}_{\mathrm{red}} + r \cdot \mathrm{MME}^{*}\right) + \overline{e}_m\right) + \mathrm{obs} \qquad (10)$$

Some assumptions are necessary, as for example how to obtain the MME and the $\overline{e}_m$ (the mean error of each model) for the reduced space. The assumption we made consists in projecting these quantities onto PC_m too, as it is no possible to associate them to their original time series.

The use of the observational data for recreating the time series is a major shortcoming of this methodology, which can be moderated in case of applications to forecasting. In that case we could use a portion of the data to generate back $\mathrm{mod}_{\mathrm{PC}}$, and another portion for verification of the forecast. Current work is devoted to this aspect.

5.5 Comparing the different methods: discussion

5.5.1 Member selection

Members selected with MDS, HC, minMSE and CAR score are reported in Table 4, where the redundancy index ρI of each reduced ensemble is also reported.

HC analysis of region 1 highlights that there is group of two models common to the four pollutants (FI1 and FR3), and with the exception of ozone DE1 and HR1 are also in common. Furthermore, given the high level of similarity between DE2 and US4, NO_2 and SO_2 are represented by the same members (Table 4). The outputs of these models have therefore the least correlated bias, and in this sense can be considered not redundant. For region 2 we found the selection to be more sensitive to the species, with only US3 and DE1 common to three species. The spatial dependence of the bias is not entirely removed by the metric d_m, as members of region 1 and 2 are quite different, with only three members in common for ozone (FI1, FR4, UK2) and CO (FI1, DE1, HR1). Members selected by HC and MDS are, in general, different.

The combination of minimum MSE for region 1 is often achieved by combining the members that individually performs best. For example, for NO_2 and SO_2 the two models whose mean produce the minimum MSE have the highest rank, individually, in terms of error. This is not the case for O_3 and CO. For the former we need to combine the first three ranked (FR4, PL1, US3) with the last one (DE1) and a middle ranked one (DK1), whilst for the latter the combination is composed by the best individual model (US4) with two middle ranked ones (FI1 and UK2). We should point out, however, that the combination of the first three MSE-ranked models produces, for all pollutants, a MSE very close to the global minimum, as already demonstrated by Knutti et al. (2010). It remains to be explained, though, why the ensemble of global minimum for ozone encompasses the worst performing models along with the best ones. Pierce et al. (2009), in the context of climate modeling, showed that the mean of the best and that of the worst models that could be built out a large ensemble were statistically indistinguishable, and that the rank of the ensemble did not reflect that of the individual models. Similar conclusions were drawn by Solazzo et al. (2012b) for ozone in Europe and North America. The reason for this is an open issue.

Further to that, for some species, the minimum error is obtained by combining highly redundant members (Table 4), as for example SO_2 and NO_2 in EU region 2, where the two instances of WRF/Chem run by DE2 and US4 both participate to minimize the MSE. As we can see from Fig. 4, these members (in the red square) are often those maximizing the variance of the error. The presence of two highly similar models is difficult to interpret. If we look at the individual model performance for SO_2 and NO_2 we find that DE2 and US4 are not individually ranked the highest for error and variance. In general, minMSE is achieved by combining redundant and less redundant models. For example, DE3 errors are uncorrelated with the quintuplet of models minimizing MSE for SO_2 in region 2, while FR3 (which also belongs to the quintuplet) is highly redundant with respect to the others. Similar patterns are detected for the other compounds too. This is an additional indication that independence and skills need be investigated separately (Abramowitz, 2010).

Two similar, high redundant models are bound to score likewise under a variety of member selection techniques. This could be a possible way to read the combination of redundant members optimizing the MSE. We have applied the CAR score recently developed by Zuber and Strimmer (2011) to our dataset (available in the package "relaimpo" for the R statistical software (www.r-project.org)). This method provides a ranking based on the partial correlation between model and the observations, conditioned to all other models. The CAR methodology is related to the amount of explained variance, enforces the simultaneous selection of highly correlated predictors and penalises variables correlating with opposite signs with the observations. Models with a small CAR score contribute little to improve the prediction error or to reduce the unexplained variance. Interestingly, for the species O_3 and NO_2 of region 1 and O_3, CO, and NO_2 of region 2 two of the selected models using CAR are in common with the minMSE selection (the first four CAR-ranked models are reported in Table 4). Further to that, the overall redundancy of the ensemble built by the mean of the first four CAR-ranked models is, in some occasion, even lower than that of HC selection (O_3, SO_2, and NO_2 in region 1; CO and NO_2 in region 2). We shall further notice that the minMSE having lower redundancy than any other method is related to the fact that the subspace has dimensionality of two (SO_2 and CO region 1) and three (NO_2 region 1 and O_3 region 2). As noted above, from the error decomposition theorem in the case of low dimensionality it is straightforward to assess that the error in minimised by low redundant members, as the covariance term is null.

5.5.2 Skill scores

In Table 5 we report the scores of the reduced ensemble generated with the methods discussed above. The full-member ensemble mean is also included as reference. With few exceptions, the reduced ensembles score better than, or as well as, the full ensemble, especially in terms of variability. Overall, the minMSE selection seems to

cc BY

outperform the other techniques for a number of pollutants in both regions. It gives the best accuracy (lowest error) by definition, but scores among the best also for variability. Good performance do not seem to be related to the redundancy of the ensembles, as to low redundancy values (SO_2, CO, NO_2 of region 1) do not systematically correspond the best scores. This aspect deserves future investigations. HC, MDS and CAR methods do not consistently score high, although performing best in some occasions.

The overwhelming strength of the weighted PCA ensemble derives from having used the observations to rebuild the time series, as discussed in Sect. 5.4. In general, though, one the main drawbacks of weighted ensembles is that they are not robust enough to be applied under a variety of scenarios (species, temporal, and spatial), and in practical applications MM mean is often preferred (Pierce et al., 2009; Knutti et al., 2010).

5.6 Implications for AQ forecasting

We outline here some considerations about applying the techniques of dimension reduction and member selection to periods of time other than those used for selecting the members. It is in the ensemble forecasting applications that the low redundancy of the bias plays the most important role: since observations are not available to provide evaluation, averaging out of errors is the only means to avoid common, redundant biases to determine the direction of the (biased) agreement.

We thus ask whether any associativity among members can be inferred in the case observational data were not available. In other words, knowing the associativity among the errors, what can be deducted about the associativity of the models underlining those errors? This problem is of direct relevance to forecasting, thus worth investigating. The starting point is as usual the covariance matrix of the errors $\mathrm{cov}(d_i, d_j)$, which we assume it is known. After some basic manipulation we get:

$$\begin{aligned}\mathrm{cov}(d_i, d_j) &= \mathrm{cov}(m_i, m_j) - (\mathrm{cov}(m_i,\mathrm{obs}) + \mathrm{cov}(m_j,\mathrm{obs})) - \mathrm{var}(\mathrm{obs}) \\ &= \mathrm{cov}(m_i, m_j) - (\mathrm{var}(d_i) + \mathrm{var}(d_j)) - \mathrm{var}(\mathrm{obs})\end{aligned} \quad (11)$$

The model errors covariance and the model covariance are strictly related, thus we cannot prescind from the observations. All we can do is to infer some consideration about the covariance of the model errors for short periods of time ahead. In practical terms, we first derive a reduced ensemble from the matrix of errors cov(d_i, d_j). Then, if we trend of the error does not change drastically for a few hours or days ahead, we can deduce that the association among them does not change either, thus the reduced ensemble is still the best option. Exploitation of reducing ensembles and member selection for forecasting applications is a topical argument and a matter of ongoing work.

Recently, Galmarini et al. (2013) have investigated the possibility of forecasting air quality starting from the combination of well-behaved spectral properties extracted from the AQMEII ensemble. The results show that the approach outruns even the ensemble median. Further investigation will be devoted to determine the correspondence between the reduced set obtained here and the properties of the ensemble put together by Galmarini et al. (2013) for the sake of identifying a deeper structure inside in the model behaviour and performance.

6 Conclusions

That of the similarity of members in ensemble modelling is an outstanding issue which has recently raised awareness in the ensemble climate community but not in the air quality one. In this study we explain the risks of combining models sharing high correlated bias into ensembles. We apply our analysis to a high resolution dataset covering two regions of EU for 3 months. Along with observational data, we have treated results of 13 AQ models for four air pollutants: CO, O_3, NO_2 and SO_2.

We have provided definitions for the concepts of independence, diversity/similarity, redundancy of models and their errors, which are often used interchangeably, giving raise to misconception. Due to practical difficulties in computing independency, we used the redundancy instead, which is simpler to handle and has the advantage of expressing the amount of the accounted-for variance, regardless of the diversity of

cc BY

models. Conceptually we believe this is very important, as it allows to univocally interpreting the results.

We started by applying the metric d_m introduced in climate modelling studies to our ensemble of regional-scale pollutant concentrations. d_m serves the scope of eliminating overarching commonalities among members and to explore hidden similarities, i.e. those underlying common modules and parameters in the models. Some main results and considerations:

1. The correlation among the majority of models remained a constant feature across the two examined regions, but varied from species to species. In fact it is generally not possible to identify model similarities common to the four species. This implies a large spectrum of partially shared modules and parameterisations within the AQ modelling systems which are invoked depending on the species and on other inputs, such as meteorology and emissions. Indeed, although most of the model similarities encapsulated by the multi-model ensemble mean error were removed by calculating d_m, similarities among model errors were still found significant;

2. By projecting the observational values into the eigenvectors of the anomalies of the models about the MM ensemble, we found that the ensemble is wide, that is, accounts for more variability than that of the observations. We concluded that the ensemble size, constrained by the available members, was too large. Given the massive application of AQ models in regulatory applications and the more and more stringent AQ targets, the detected overconfidence can cost greatly. This, together with item 1 above, justify the need for the analysis of the redundancy of the datasets;

3. We therefore explored some dimension-reduction methods:

   - Eigenvalue methods – number of effective models and Principal Component Analysis;

   - Clustering analysis and dendrogram representation;

- Multi-dimensional Scaling and graphical representation of model similarities as mutual distance among models;
- minMSE heuristic investigation, determining the size of the ensemble of models whose mean minimize the MSE.

None of the aforementioned method is new, they are all well-established techniques used, in many flavours, in various branches of science. They have not been used before in an AQ ensemble context, neither compared. We also introduce, where possible, the nexus between these techniques and redundancy. We found that the optimal size of an ensemble of poorly correlated members is of about 4–6, implying that more than half of the information of the full MM ensemble is redundant.

1. We continued the investigation by applying member-selection techniques and scoring to the reduced ensembles against simple operational metrics, taking the scores of the full member ensemble mean as benchmark. We prove that subsets of models outperform the full ensemble. The minMSE selection seems to outperform the other techniques for a number of pollutants in both regions. It gives the best accuracy (lowest error) by definition, but scores among the best also for variability. HC, MDS and CAR methods do not consistently score high, although performing best in some occasions.

2. The error being minimized by highly redundant members does not justify, in our view, the use of the ensemble of those members. Skills and diversity need to be analysed in separation. This is because redundant members might share common biases which will force the agreement to be directed towards the same direction, with the risk of misjudging the results. These aspects are likely to be detected by diagnostic-type of analysis (rather that by simple operational scores based on distance metrics) and may often reveal more about the causes of model errors and the processes responsible for those errors (Dennis et al., 2010; Gleckler et al., 2008). The combination of minimum error might just arise from a favourable

numeric combination in the trade-off between covariance and bias. Indeed, if all mutual covariances between the models are negative then the optimal MSE is less than for un-correlated models. It remains to be explained why also high redundant ensembles produced high scores; a further open issue is why ensembles of individually high-ranked members might be outperformed by ensembles of high and low ranking members.

3. Application of PCA to the matrix of errors for the purpose of data reduction has proved successful. By contrast, generating the reduced time series (the time series projected on the leading eigenvectors) is not trivial and requires the use of the observational data, which masks the outcome of the procedure. As no applications of this sort have been found in the literature, our intention is to devote future work to this aspect which might be relevant in the realm of forecasting;

4. Finally, we have highlighted the steps for applying the methods of dimension reduction and member selection to a forecasting context.

We also believe the effort we spent to migrate some of the knowledge and techniques developed in other scientific areas (especially computer science, genetics, and climate modelling) will contribute to raise awareness in the ensemble AQ community about the dependency of models and the meaning of model agreement.

Acknowledgements. The AQMEII community (http://aqmeii.jrc.ec.europa.eu/) is kindly acknowledged for providing the data used in the analysis.

References

Abramowitz, G.: Model independence in multi-model ensemble prediction, Australian Meteorological and Oceanographic Journal, 59, 3–6, 2010.

Annan, J. D. and Hargreaves, J. C.: Reliability of the CMIP3 ensemble, Geophys. Res. Lett., 37, L02703, doi:10.1029/2009GL041994, 2010.

Annan, J. D. and Hargreaves, J. C.: Understanding the CMIP3 multimodel ensemble, J. Climate, 24, 4529–4538, 2011.

Azimi, J. and Fern, X.: Adaptive Cluster Ensemble Selection, Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence IJCAI-09, 992–997, 2009.

Borg, I. and Groenen, P. : Modern Multidimensional Scaling: Theory and Applications, 2nd edn., Springer-Verlag, New York, 2005.

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

Brown, G., Wyatt, J. L., and Tino, P.: Managing diversity in regression ensembles, J. Mach. Learn. Res., 6, 1621–1650, 2005.

Cattell, R. B.: The scree test for the number of factors, Multivariate Behavioural Research, 1, 245–276, 1966.

Côté, J., Gravel, S., Méthot, A., Patoine, A., Roch, M., and Staniforth, A.: The operational CMC–MRB Global Environmental Multiscale (GEM) model – Part I: Design considerations and formulation, Mon. Wea. Rev., 126, 1373–1395, 1998.

Cover, T. T.: The best two independent measures are not the two best, IEEE T. Syst. Man. Cyb., 4, 116–117, 1974.

Cover, T. and J. Thomas: Elements of Information Theory, 2nd edn., Wiley-Interscience, Hoboken, N. J., 2006.

Dennis, R., Fox, T., Fuentes, K., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S. T., Scheffe, R., Schere, K., Steyn, D., and Venkatram, A.: A framework for evaluating regional-scale numerical photochemical modelling systems, Environ. Fluid Mech., 10, 471–489, doi:10.1007/s10652-009-9163-2, 2010.

Ding, C. and He, X.: K-means clustering via Principal component analysis, Proceedings of the 21st International Conference on Machine Learning, Banff, Canada, 4–8 July 2004, 2004.

Ding, C. and Peng, H.: Minimum Redundancy feature selection from microarray gene expression data, Journal of Bioinformatics and Computational Biology, 3, 185–205, 2005.

Fern, X. Z. and Brodley, C. E.: Solving cluster ensemble problems by bipartite graph partitioning, in Proceedings of 21th International Conference on Machine Learning (ICML2004), 2004.

Fiore, A. M., Dentener, F. J., Wild, O., Cuvelier, C., Schultz, M. G., Hess, P., Textor, C., Schulz, M., Doherty, R. M., Horowitz, L. W., MacKenzie, I. A., Sanderson, M. G., Shindell, D. T., Stevenson, D. S., Szopa, S., Van Dingenen, R., Zeng, G., Atherton, C., Bergmann, D., Bey, I., Carmichael, G., Collins, W. J., Duncan, B. N., Faluvegi, G., Folberth, G., Gauss, M., Gong, S., Hauglustaine, D., Holloway, T., Isaksen, I. S. A., Jacob, D. J., Jonson, J. E., Kaminski, J. W., Keating, T. J., Lupu, A., Marmer, E., Montanaro, V., Park, R. J., Pitari, G., Pringle, K. J., Pyle, J. A., Schroeder, S., Vivanco, M. G., Wind, P., Wojcik, G., Wu, S., and Zuber, A.: Multimodel estimates of intercontinental source-receptor relationships for ozone pollution, J. Geophys. Res., 114, D04301, doi:10.1029/2008JD010816, 2009.

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 harmonised, simplified and efficient assistance to air quality model developments and evaluation, Atmos. Environ., 53, 51–59, 2012.

Galmarini, S., Kioutsioukis, I., and Solazzo, E.: E pluribus unum: ensemble air quality predictions, Atmos. Chem. Phys. Discuss., 13, 581–631, doi:10.5194/acpd-13-581-2013, 2013.

Gleckler, P. J., Taylor, K. E., and Doutriaux, C.: Performance metrics for climate models, J. Geophys. Res., 113, D06104, doi:10.1029/2007JD008972, 2008.

Groenen, P. J. F. and van de Velden, M.: Multidimensional Scaling, Erasmus University Rotterdam, Econometric Institute, Econometric Institute Report EI 2004-15, 2004.

Grömping, U.: Estimator of relative importance in linear regression based on variance decomposition, Am. Stat., 61, 139–147, 2007.

Guttman, L.: Some necessary conditions for common-factor analysis, Psychometrika, 19, 149–161, 1954.

Hadjitodorov, S., Kuncheva, L. I., and Todorova, L. P.: Moderate diversity for better cluster ensembles, Information Fusion Journal, 7, 264–275, 2006.

Jolliffe, I.: Principal component analysis, Springer, 2nd edn., 2002.

Jong, J.-C. and Kotz, S.: On a relation between principal components and regression analysis, Am. Stat., 53, 349–351, 1999.

Jun, M., Knutti, R., and Nychka, D. W.: Local eigenvalue analysis of CMIP3 climate model errors, Tellus, 60, 992–1000, 2008.

Kaiser, H.: The application of electronic computers to factor analysis, Educ. Psychol. Meas., 20, 141–151, doi:10.1177/001316446002000116, 1960.

Kaminski, J. W., Neary, L., Struzewska, J., McConnell, J. C., Lupu, A., Jarosz, J., Toyota, K., Gong, S. L., Côté, J., Liu, X., Chance, K., and Richter, A.: GEM-AQ, an on-line global multiscale chemical weather modelling system: model description and evaluation of gas phase chemistry processes, Atmos. Chem. Phys., 8, 3255–3281, doi:10.5194/acp-8-3255-2008, 2008.

Knutti, R.: The end of model democracy?, Climatic Change, 102, 395–404, 2010.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., and Meehl, G.: Challenges in combining projections from multiple climate models, J. Climate, 23, 2739–2758, 2010.

Kong, W., Vanderburg, C. R., Gunshin, H., Rogers, J. T., and Huang, X.: A review of independent component analysis application to microarray gene expression data, BioTechniques, 45, 501–520, 2008.

Legendre, P. and Legendre, L. F. J.: Numerical Ecology, chapt. 11, Elsevier Science BV, Amsterdam, 853 pp., 1998.

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

McSweeney. C. F., Jones, R. G., and Booth, B. B. B.: Selecting ensemble members to provide regional climate change information, J. Climate, 25, 7100–7121, 2012.

Moesa, H. A., Dukka Bahadur, K. C., and Akutsu, T.: Efficient determination of cluster boundaries for analysis of gene expression profile data using hierarchical clustering and wavelet transformation, Genome Inform., 16, 132–141, 2005.

Murtagh, F. : Complexities of hierarchic clustering algorithms: the state of the art, Comput. Stat. Quart., 1, 101–113, 1984.

Peng, H., Long, F., and Ding, C.: Feature selectionbased on mutual information: criteria of Max-dependency, max-relevance, and min-redundancy. IEEE T. Pattern Anal., 27, 1226–1238, 2005.

Pennel, C. and Reichler, T.: On the effective numbers of climate models, J. Climate, 24, 2358–2367, 2011.

Pirtle, Z., Meyer, R., and Hamilton, A.: What does it mean when climate models agree? A case for assessing independence among general circulation models, Environ. Sci. Pol., 799, 351–361, 2010.

Potempski, S., Galmarini, S., Addis, R., Astrup, P., Bader, S., Bellasio, R., Bianconi, R., Bonnardot, F., Buckley, R., Damours, R., Van Dijk, A., Geertsema, G., Jones, A., Kaufmann, P., Pechinger, U., Persson, C., Polreich, E., Prodonova, M., Robertson, L., Srrensen, J., and Syrakov, D.: Multi-model ensemble analysis of the ETEX-2 experiment, Atmos. Environ., 42, 7250–7265, 2008.

Potempski, S. and Galmarini, S.: *Est modus in rebus*: analytical properties of multi-model ensembles, Atmos. Chem. Phys., 9, 9471–9489, doi:10.5194/acp-9-9471-2009, 2009.

Rao, S. T., Galmarini, S., and Puckett, S.: Air Quality Model Evaluation International Initiative (AQMEII), Bulletin of the Marican Meteorological Society, 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.

Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M. D., Appel, K. W., 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., Hogrefe, C., Miranda, A. I., Nopmongco, U., Prank, M., Sartelet, K. N., Schaap, M., Silver, J. D., Sokhi, R. S., Vira, J., Werhahn, J., 15 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, 2012a.

Solazzo, E., Bianconi, R., Vautard, R., Appel, K. W., Moran, M. D., Hogrefe, C., Bessagnet, B., 5 Brandt, J., Christensen, J. H., Chemel, C., Coll, I., van der Gon, H. D., Ferreira, J., Forkel, R., Francis, X. V., Grell, G., Grossi, P., Hansen, A. B., Jericevic, A., Kraljevic, 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. T., and Galmarini, S: Ensemble modelling of surface level ozone in Europe and North AMerica in the context of AQMEI, Atmos. Environ., 53, 60–74, 2012b.

Talagrand, O., Vautard, R., and Strauss, B.: Evaluation of probabilistic prediction systems, paper presented at a seminar on predictability, ECMWF, Reading (UK), 1998.

Tebaldi, C. and Knutti, R.: The use of multi-model ensemble in probabilistic climate projections. Philos. Trans. Roy. Soc. A, 365, 2053–2075, 2007.

Tinsley, H. E. A. and Brown, S. D.: Handbook of applied multivariate statistics and mathematical modeling, Academic Press, California (USA), 334–338, 2000.

Van Loon, M., Vautard, R., Schaap, M., Bergstrom, R., Bessagnet, B., Brandt, J., Builtjes, P. J., H., Christensen, J. H., Cuvelier, C., Graff, A., Jonson, J. E., Krol, M., Langner, J., Roberts, P., Rouil, L., Stern, R., Tarrason, L., Thunis, P., Vignati, E., White, L., and Wind, P.: Evaluation of long-term ozone simulations from seven regional air quality models and their ensemble average, Atmos. Environ., 41, 2083–2097, 2007.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◂◂ | ▸▸

◂ | ▸

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Vautard, R., Schaap, M., Bergström, R., Bessagnet, B., Brandt, J., Builtjes, P. J. H., Christensen, J. H., Cuvelier, C., Foltescu, V., Graf, A., Kerschbaumer, A., Krol, M., Roberts, P., Rouïl, L., Stern, R., Tarrason, L., Thunis, P., Vignati, E., and Wind, P.: Skill and uncertainty of a regional air quality model ensemble, Atmos. Environ., 43, 4822–4832, 2009.

Vautard, R., Moran, M. D., Solazzo, E., Gilliam, R. C., Matthias, V., Bianconi, R., Chemel, C., Ferreira, J., Geyer, B., Hansen, A. B., Jericevic, A., Prank, M., Segers, A., Silver, J. D., Werhahn, J., Wolke, R., Rao, S. T., and Galmarini, S: Evaluation of the meteorological forcing used for AQMEII air quality simulations, Atmos. Environ., 53, 15–37, 2012.

Yoon, S. and Kim, S.: Mutual information-based SVM-RFE for diagnostic classification of digitized mammograms, Pattern Recogn. Lett., 30, 1489–1495, 2009.

Youness, G. and Saporta, G.: Comparing partitions of two sets of units based on the same variables, Adv. Data Anal. Classif., 4, 53–64, doi:10.1007/s11634-009-0057-4, 2010.

Zuber, V. and Strimmer, K.: High-Dimensional Regression and variable selection using CAR scores, Stat. Appl. Genet. Mo. B., 10, 1–27, doi:10.2202/1544-6115.1730, 2011.

Title Page

Abstract | Introduction

Conclusions | References

Tables | Figures

◂| |▸

◂ ▸

Back | Close

Full Screen / Esc

Printer-friendly Version

Interactive Discussion

Table 1. Number of rural receptors by species and regions.

Europe	O_3	SO_2	NO_2	CO
region 1	199	34	56	23
region 2	225	131	136	54

cc BY

Table 2. Participating models and features.

	Model		Res	No.	Emissions	Chemical BC
Code	Met	AQ	(km)	Vertical layers		
DK1	MM5	DEHM	50	29 (top: 100 hPa)	Global emission databases, EMEP	Satellite measurements
FR3	MM5	Polyphemus	24	9 (top: 12 km)	Standard[a]	Standard
HR1	PARLAM-PS	EMEP	50	20	EMEP model	From ECMWF and forecasts
UK2	WRF	CMAQ	18	34 (up to 50 hPa)	Standard[a]	Standard
DE2	WRF	WRF/Chem	22.5	36 (top: 22.5 km)	Standard[a]	Standard
US4	WRF	WRF/Chem	22.5	36 (top: 22.5 km)	Standard[a]	Standard
FI1	ECMWF	SILAM	24	9 (top: 10 km)	Standard anthropogenic In-house biogenic	Standard
FR4	MM5	Chimere	25	9 (up to 500 hPa)	MEGAN, Standard	Standard
PL1	GEM	GEM-AQ	0.2 degree[b]	28 (up to 10 mb)	Standard over AQMEII region; Global EDGAR/GEIA over the rest of the global domain	Global variable grid setup (no lateral boundary conditions)
NL1	ECMWF	Lotos-EUROS	25	4 (top: 25 km)	Standard[a]	Standard
DE1	COSMO	Muscat	24	40 (top: 24 km)	Standard[a]	Standard
US3	MM5	CAMx	15	20 (top: 24 km)	MEGAN, Standard	Standard
DE3	COSMO-CLM	CMAQ	24	30 (up to 100 hPa)	Standard[a]	Standard

[a] Standard anthropogenic emission and biogenic emission derived from meteorology (temperature and solar radiation) and land use distribution implemented in the meteorological driver (Guenther et al., 1994; Simpson et al., 1995).

[b] Corresponding to 22.2 km at the domain center.

cc BY

Table 3. M_{eff} from Eq. (6). Values have been calculated using corr(d_i,d_j) (corr(e_i,e_j)).

Europe	O_3	SO_2	NO_2	CO
region 1	5.8 (2.3)	5.7 (1.3)	6.5 (2.2)	6.5 (1.3)
region 2	5.2 (2.5)	5.3 (3.2)	5.9 (2.5)	5.6 (1.9)

Table 4. Representative models (corr(d_i,d_j) for the months of JJA). The number in parenthesis is the redundancy index ρI for each ensemble.

	MDS	HC (min bias)	MinMSE	CAR
		EU region 1		
O_3	DE2;NL1;DK1;DE1 (0.27)	FI1,FR3,FR4,UK2,US4 (0.12)	FR4,PL1,US3,DE1,DK1 (0.53)	PL1,US3,HR1,UK2 (0.07)
SO_2	US4;US3;FI1;NL1 (0.29)	FI1,FR3,DE1,HR1,US4 (0.22)	HR1,DE1 (FI1, UK2) (0.007)	FR3,DK1,FR4,UK2 (0.15)
CO	DK1;FR3;HR1;US3 (0.25)	FI1,FR3,DE1,DK1,HR1 (0.17)	FI1,DE1 (NL1, US3) (0.02)	FR4,UK2,FI1,US4 (0.29)
NO_2	FR4;FR3;PL1;DK1 (0.27)	FI1,FR3,DE1,DE2,HR1 (0.21)	FI1,UK2,US4 (DE2) (0.13)	UK2,FI1,DE3,HR1 (0.15)
		EU region 2		
O_3	US4;US3;FI1;HR1 (0.31)	FI1;FR4;UK2;US3;DE3 (0.30)	FR4;US3;DE1 (FI1) (0.23)	DE1,US3,PL1,DE2 (0.35)
SO_2	DK1;UK2;US4;FR3 (0.29)	DE3;US3;DE1;NL1;US4 (0.28)	DE3,FR3,US3,US4,DE2 (0.47)	DE3,DK1,UK2,NL1 (0.45)
CO	US3;DK1;DE1;NL1 (0.60)	FI1;DE1;NL1;PL1;HR1 (0.15)	FI1,NL1,US3,HR1,DE1 (0.59)	UK2,DE3,HR1,NL1 (0.20)
NO_2	US4;FI1;FR4;HR1 (0.29)	US3;DE1;PL1;DK1;DE2 (0.18)	NL1,US4,HR1,DE2,DE3 (0.55)	UK2,DE3,DE1,NL1 (0.08)

Table 5. Ensemble skills for regions 1 and 2 of Europe (JJA). (RMSE: root mean square error; R: Pearson correlation coefficient; NMB: Normalised Mean Bias; STDEV ratio: modeled to observed standard deviation). Results in italic are those for which the selected ensemble scores better or as good as the full member ensemble (vice versa for the values in bold).

	EU region 1	RMSE	*R*	NMB	STDEV ratio
CO	PCA	*0.03*	*0.65*	*0.25*	**5.20**
	HC	*0.06*	*0.36*	*−0.22*	*0.39*
	minMSE	*0.05*	*0.36*	*−0.09*	*0.45*
	MDS	*0.06*	**0.28**	*−0.19*	*0.51*
	CAR	*0.06*	*0.41*	**−0.29**	*0.44*
	Full Ensemble	0.06	0.38	−0.26	0.39
O_3	PCA	*2.5*	*0.99*	*0.04*	*0.99*
	HC	**12.0**	*0.96*	**0.003**	**0.63**
	minMSE	*8.1*	*0.96*	**0.03**	*0.81*
	MDS	*11.2*	**0.94**	**0.04**	*0.70*
	CAR	*10.8*	*0.97*	**−0.05**	*0.71*
	Full Ensemble	10.9	0.96	0.002	0.67
SO_2	PCA	*0.9*	*0.96*	*0.12*	*1.12*
	HC	*2.0*	*0.17*	*−0.07*	*0.57*
	minMSE	*1.9*	*0.27*	*−0.11*	*0.55*
	MDS	*2.1*	**0.16**	*0.12*	*0.60*
	CAR	*2.2*	< 0.1	*−0.03*	*0.75*
	Full ensemble	2.2	0.17	0.26	0.49
NO_2	PCA	*1.0*	*0.99*	**0.20**	**1.09**
	HC	*3.5*	*0.68*	*−0.09*	*1.05*
	minMSE	*3.0*	*0.74*	*−0.09*	*0.96*
	MDS	**4.7**	**0.61**	**0.20**	**1.43**
	CAR	*3.6*	*0.74*	**−0.24**	*0.95*
	Full Ensemble	3.7	0.67	0.18	1.06

Table 5. Continued.

	EU region 2	RMSE	*R*	NMB	STDEV ratio
CO	PCA	*0.04*	*0.99*	*0.18*	*0.98*
	HC	*0.05*	**0.48**	*−0.21*	*0.68*
	minMSE	*0.03*	**0.38**	*0.02*	*0.83*
	MDS	*0.04*	**0.45**	*−0.17*	*0.67*
	CAR	*0.07*	*0.58*	**−0.38**	*0.57*
	Full Ensemble	0.07	0.50	−0.35	0.57
O_3	PCA	*2.5*	*0.98*	*−0.005*	*0.99*
	HC	*11.6*	**0.92**	*0.005*	*0.81*
	minMSE	*7.8*	*0.95*	*0.02*	*0.91*
	MDS	**15.3**	*0.93*	**−0.14**	*0.73*
	CAR	**7.8**	*0.95*	*0.02*	*0.84*
	Full Ensemble	12.3	0.93	−0.06	0.71
SO_2	PCA	*0.7*	*0.74*	*0.03*	*1.4*
	HC	*0.7*	*0.73*	*−0.13*	**3.3**
	minMSE	*0.5*	*0.59*	*−0.07*	*1.08*
	MDS	*0.8*	**0.53**	*−0.3*	*0.86*
	CAR	*0.8*	*0.76*	*−0.4*	*1.11*
	Full ensemble	0.8	0.59	−0.4	0.77
NO_2	PCA	*1.0*	*0.99*	**0.5**	*1.03*
	HC	**3.4**	*0.66*	*0.05*	**2.12**
	minMSE	*1.9*	*0.70*	*−0.07*	*1.32*
	MDS	**3.6**	*0.67*	*0.06*	**2.12**
	CAR	*2.2*	*0.75*	**−0.26**	*1.38*
	Full Ensemble	2.5	0.59	−0.16	1.47

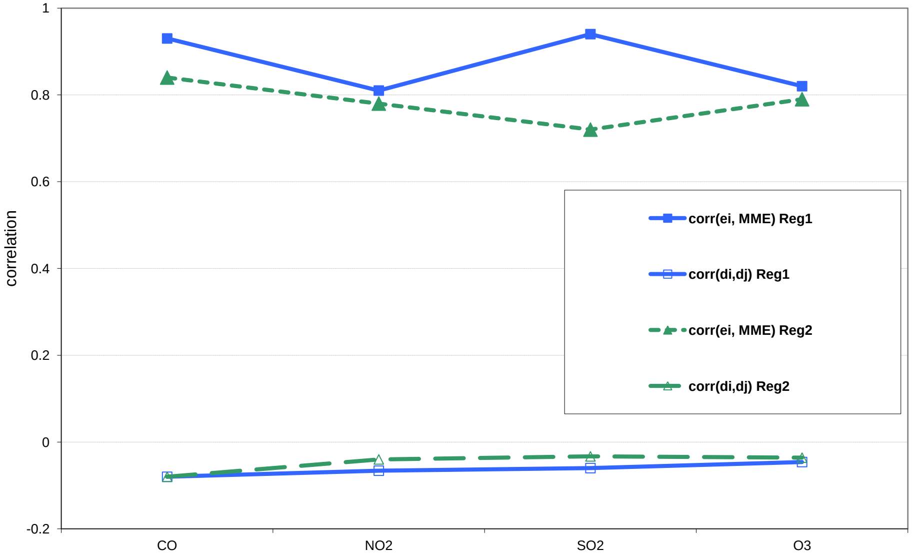

Fig. 1. Correlation of errors (between individual model and MME and between d for all model pairs) for region 1 and region 2 of Europe for the months of JJA of 2006.

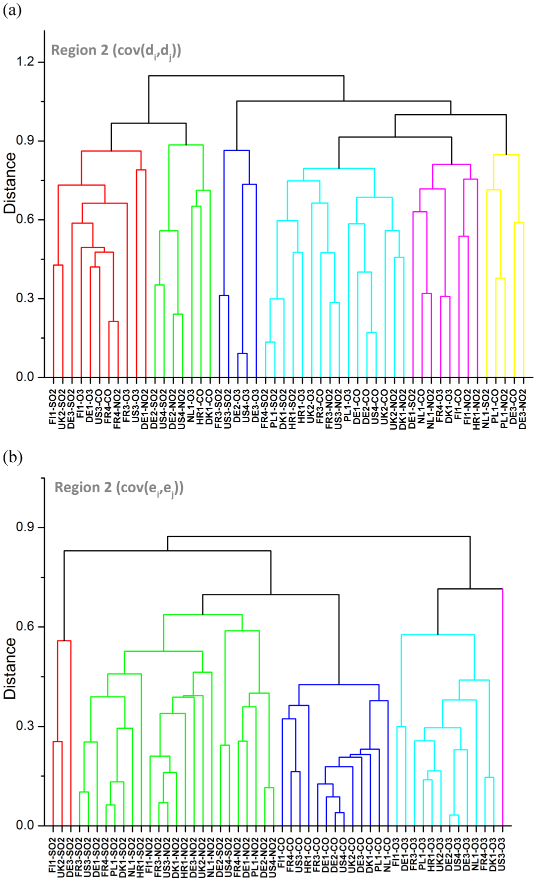

Fig. 2. Associativity trees for all models and species (European region 2) using **(a)** the cov(di,dj) and **(b)** the cov(ei,ej) as distance matrix.

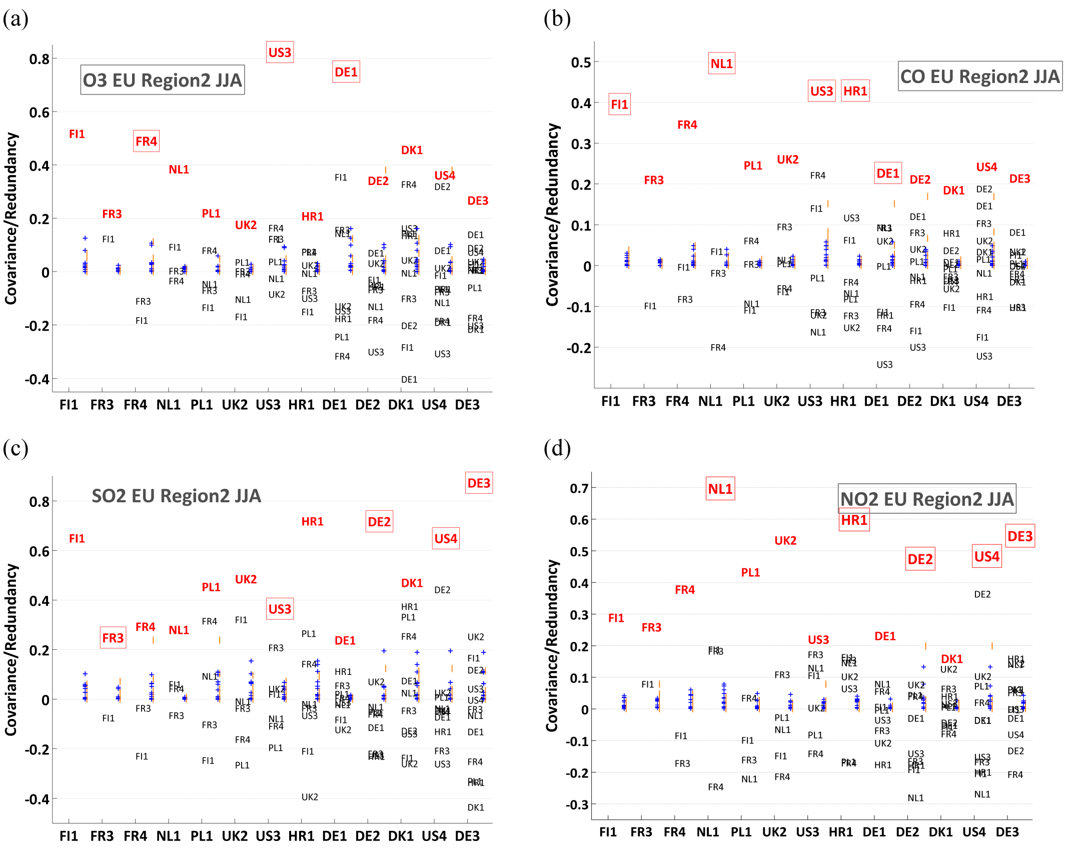

Fig. 3. cov(di,dj) by species for European region 2. In red the variance (cov(di,di)), in blue the range of redundancy (corr^2(di,dj)) and in green the range of redundancy measured by means of the mutual information (see text).The models in the square are those whose ensemble mean produces the minimum MSE (see Sect. 5.5.1).

ACPD

13, 4989–5038, 2013

Pauci ex tanto numero

E. Solazzo et al.

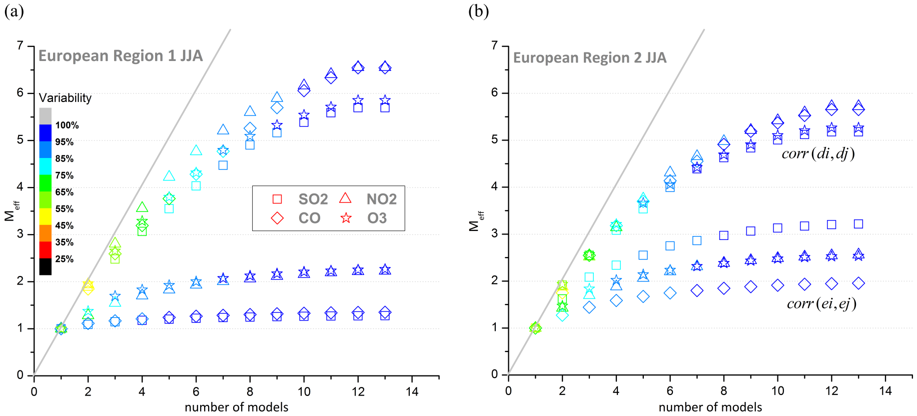

Fig. 4. M_{eff} (Eq. 6) as function of the number of models for EU region 1 and region 2. The two sets of curves have been generated from the corr(di, dj) (top curves) and the corr(ei, ej) (lower curves) matrixes. The cumulative variability is color coded. In grey is the one-to-one line.

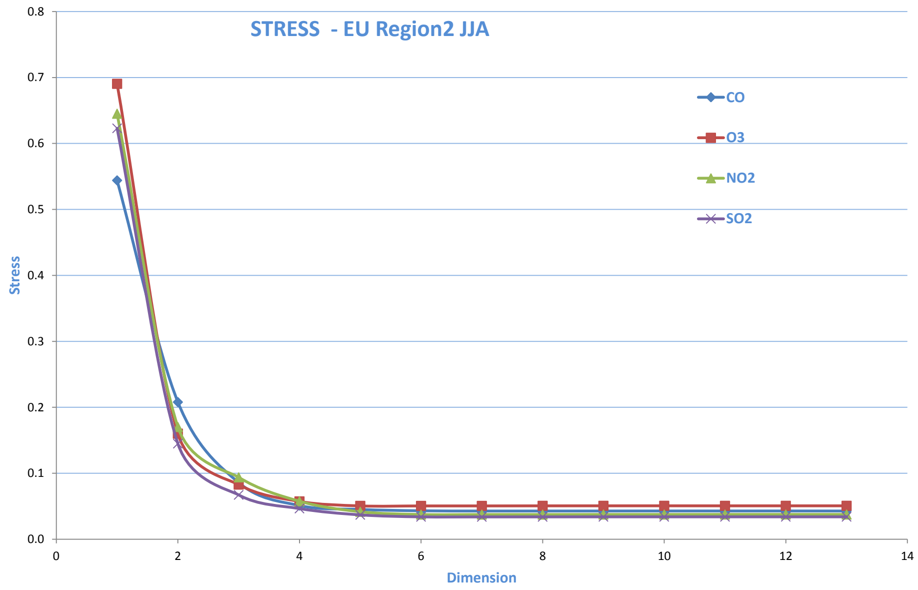

Fig. 5. Sub-space dimension calculated by minimising the stress function in the MDS methodology. The corr(di,dj) matrix is used as similarity criteria.

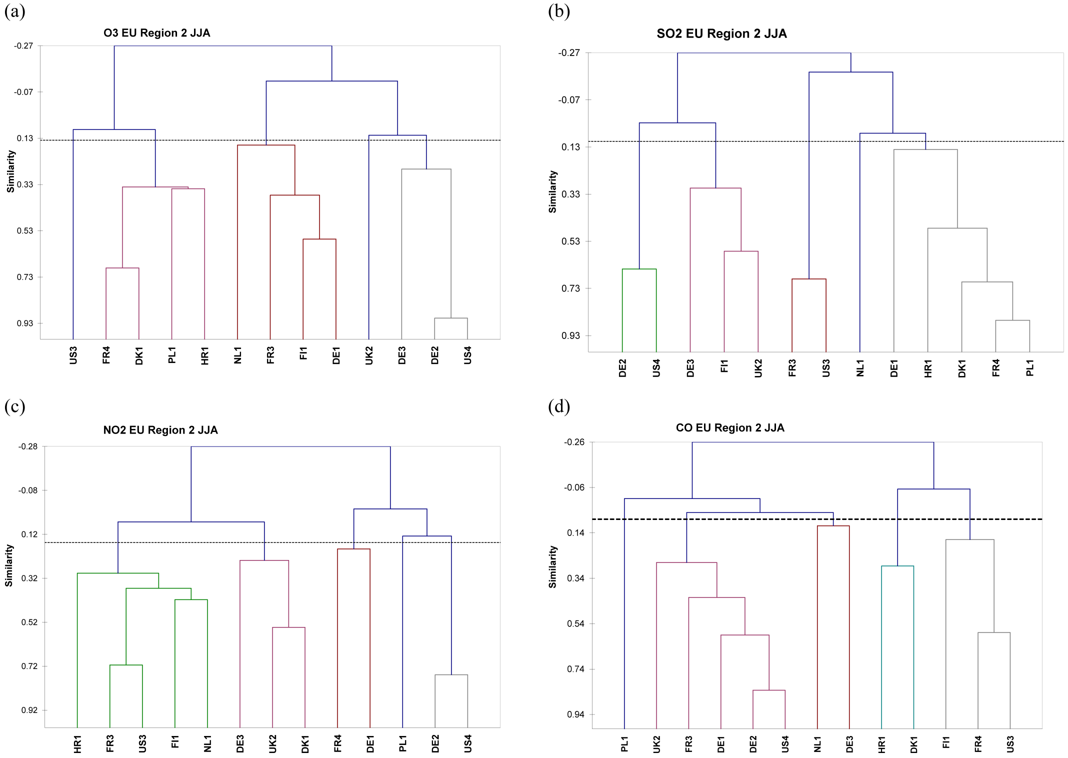

Fig. 6. Hierarchical clustering of corr(di,dj) for EU region 2. The dotted horizontal line defines the level of similarity. Disjoint clusters are identified by different colors.

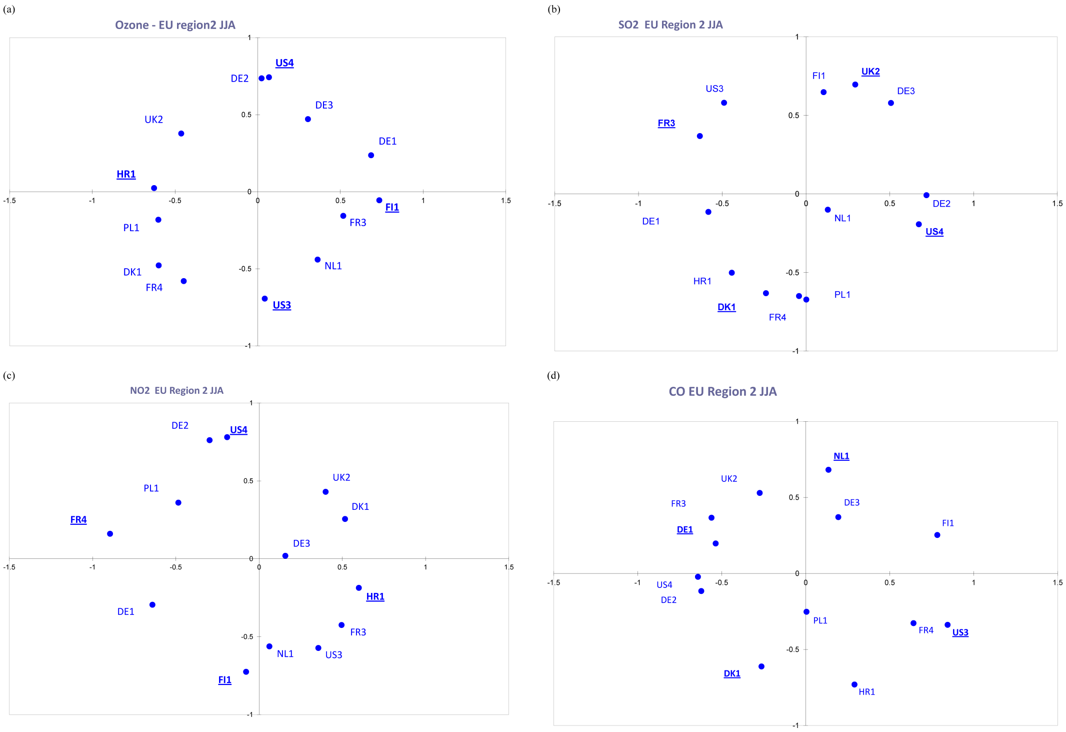

Fig. 7. 2D MDS of corr(di, dj) for EU region 2. Models underlined are those selected to generate the reduced ensemble.

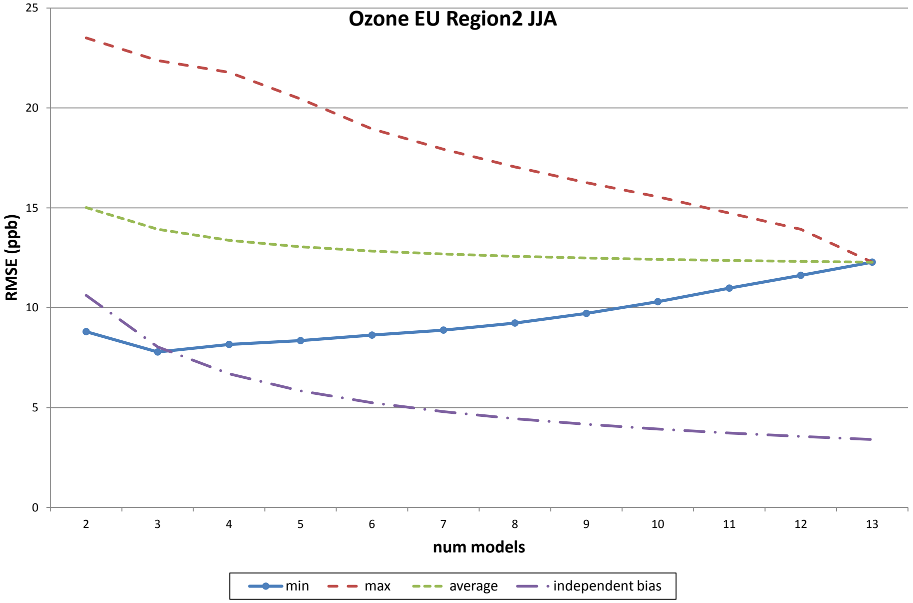

Fig. 8. Curves of maximum, mean and minimum RMSE for ozone in EU region 2. The curves are obtained by calculating the mean of randomly sampled subsets of models, as function of dimension of the subsets m. The theoretical decay that would occur if the model errors were independent, $\sim m^{-0.5}$ is also reported.