A Science-Based Use of Ensembles of Opportunities for Assessment and Scenario study

Efisio Solazzo and Stefano Galmarini

European Commission,

Joint Research Centre,

Institute for Environment and Sustainability, Air and Climate Unit,

Ispra (Italy).

KEY NOTES

Multi-model ensembles need inspection prior use

ABSTRACT

The multi-model ensemble exercise performed within the HTAP project context [Fiore et al., 2009] is used here as an example of how a *pre-inspection*, diagnosis and selection of an ensemble, can produce more reliable results. The procedure is contrasted with the often-used practice of simply averaging model simulations, assuming different models produce independent results, and using the diversity of simulation as an illusory estimate of model uncertainty. It is further and more importantly demonstrated how conclusions can drastically change when future emission scenarios are analysed using an un-inspected ensemble. The HTAP multi-model ensemble analysis is only taken as an example of a wide spread and common practice in air quality modelling.

1. INTRODUCTION

A multi-model (MM) ensemble is defined as a group of simulations of the same case study, produced by formally different models, which are statistically treated in an attempt to improve the quality of the result [Potempski and Galmarini, 2009]. Given the ever increasing collaborations of geophysical modelling communities in joint assessment studies, MM ensembles are becoming very popular and an opportunity to extend and generalize individual deterministic model results [Solazzo et al., 2012 and; 2013; Solazzo and Galmarini, 2014; Galmarini et al., 2004; Vautard et al., 2012; Evans et al., 2013; Bishop and Abramowitz, 2013; and many others].

In particular in atmospheric sciences, MM ensembles are used extensively in climate and air quality predictions and assessments. While in climate research and applications many of the concepts applied and described here are well known and correctly used, in air quality this is not always the case and several are the examples of direct use of *un-inspected* MM ensembles. We shall describe an *inspected* MM ensemble (opposed to an un-inspected one) as: a set of model results, whose properties and characteristics, have been analysed in an attempt to reduce the presence of redundant information or elements that are not relevant to the determination of an accurate result. An inspected ensemble should is expected to produce

a result that is more accurate than the simple average of the multi model results, at least in all the cases when the members of the ensemble are not independent (e.g., Kioutsioukis and Galmarini, 2014).
The motivations behind the necessity to inspect a MM ensemble are connected to the way in which MM ensembles are put together and to the nature of the participating models. In fact, the selection of the models whose results are ensembled is not, to the best of our knowledge and at least for air quality applications, regulated by any science based criteria and there is no a-priori specification that defines the characteristics of a model that should or should not take part to an ensemble. The constitution of a MM ensemble is merely based on an opportunity to provide model simulations and to participate to a community activity where anybody is welcome (*ensemble of opportunity*). Regarding the nature of the models producing results for ensemble applications, one should never forget that the best results are those produced by ensembles of independent (and accurate) models [Potempski and Galmarini, 2009; Kioutsioukis and Galmarini, 2014; Weigel et al., 2008; Pirtle et al. 2010; Knutti, 2010; Knutti et al., 2010; Riccio et al., 2012]. Formally, model m_1 is defined independent from m_2 if the joint probability p for a result of m_1 and m_2 can be expressed as $p(m_1,m_2)=p(m_1)p(m_2)$. When many independent models are combined together their bias can be randomly positive or negative increasing the probability of cancelling out and of the sampled uncertainty does not overlap (Knutti et al., 2010; Abramowitz, 2010; Solazzo et al., 2013). Models used in air quality (among others) are not independent, they are often sharing common assumptions, modules, input data, and cannot therefore be considered independent. In most of the cases the models are different (*Phenotypical model difference*, Potempski and Galmarini, [2009]), but are not independent. This leads to the possibility that results obtained from an ensemble, rather than representing a true alternative and independent solution, would just be like in music composition a *variation on the theme*, producing a false sense of variability which could lead to coinciding (diverging) biased results and a false sense of agreement (uncertainty).

MM ensembles derived from simply different models are prone to redundancy and overconfidence. The inspection is therefore primarily finalised at:

- the identification of the level of diversity (communality) shared by the model results,
- retaining only those that are contributing with original information
- removing the redundancy.

Techniques exist that allow such screenings that rely on the existence of observations and the comparison of the ensemble variability with the observational variability [Potempski and Galmarini, 2009; Solazzo et al., 2013; Riccio et al., 2012].

In this study we aim at demonstrating the importance of using existing good practices in the air quality MM ensemble context. Toward the scope we have selected a case study published in the past which does not exploit the true value of having multiple model results at hand. The case analyzed is the HTAP (Hemispheric Transport of Air Pollution) phase 1 multi-model exercise [Dentener et al. 2010] and in particular the multi-model ensemble activity performed within it and presented by Fiore et al. (2009). The study of Fiore et al. (2009) is used here as mere representative of a wide spread practice in the air quality modelling communities at all scales and it represents just an example on how things could be improved further. The MM ensemble by Fiore et al. (2009) is original in many aspects and, in particular, is used for sensitivity studies with respect to emission reduction options. The inspection of the ensemble can have important consequences also for emission scenarios as shown later, an aspect never considered before in the literature.

2. The case study and MM ensemble inspection

In 2006 the Task Force on Hemispheric Transport of Air Pollution (http://www.htap.org/) organised a comparison exercise of global and hemispheric transport models, focussing on the relationships between regional scale emission perturbations and the response in air quality, ecosystem, and climate related variables. The information was used in an aggregated form to evaluate air pollution abatement strategies and their impact across the Northern Hemisphere. Results of the comparison exercise are summarized in Dentener et al., [2010]; Sanderson [2008]; Fry et al. [2012]; Wild et al. [2012]; Jonson et al., [2010]; Anenberg et al., [2009]; Fiore et al., [2009].

We will focus on the MM ensemble analysis by Fiore et al. [2009] (from now FetA09). In FetA09, an average of 21 model results was used to investigate the monthly mean surface ozone concentration in three sub-regions of Europe (Mediterranean, Central Europe with receptors between 0 and 1 km height and Central Europe with receptors between 1 and 2 km height), five North-American sub-regions (North East, South West, South East, Great Lakes, and Mountainous) and one Japanese sub-region (EANET stations). Operational scores (bias, correlation coefficient and standard deviation) were calculated in each sub-region making use of ground-based measurements. The combined spatial and temporal average of the modelled concentration values resulted in smoothed monthly time-series. The analysis of FetA09 reveals that the distribution of the results is rather symmetric (Figure 1). Supported by the agreement with observations, the authors considered the MM ensemble mean to be the best possible estimate as it "*generally captures the observed seasonal cycle and is close to the observed regional mean*" [FetA09], thus justifying the use of the MM ensemble mean to quantify source-receptor relationships as well as ozone concentration response to changes in the emissions scenarios.

The scope of the analysis by FetA09 was not to prove the robustness of the MM ensemble mean, and provides an example of the widespread practice of averaging all available members, assuming that the average of many model results is always a better result than that of one model. That would be true if the models were independent but there is no a-priori proof of that. Some questions arise: how robust are the results if the members are not independent models? How different the result would be should some model not taking part to the activity or more outliers (like the one present in the Figure 1) would be present? How generalised is the result since the selection of the ensemble members is based on the voluntary participation to a joint activity and the MM ensemble does not contain all possible results? Is there any duplication of information? Is all the information contained in a MM ensemble relevant and necessary? Since the construction of a MM ensemble is not governed by scientific selection criteria, so it happens that the subsequent ensemble result strictly depends on *aleatory* factors and one can presume that it lacks generality as it is supported by assumptions known to be valid for independent members only.

The screening methodology we propose and that we apply as an example to the FetA09 set, is a good way to exploit an abundance of model results in the best way, to transform the aleatory gathering of information into a more robust result that is based on general selection criteria. The large ensemble of model results becomes an opportunity to *cherry-pick* those models whose combination produce the most accurate MM ensemble and use only those to drive conclusions. The analysis will help identifying the size of the non-redundant ensemble and the subsets of members to produce skilled results.

2.1 INSPECTING A MULTI MODEL ENSEMBLE

In this section the MM ensemble of FetA09 is inspected. We will concentrate on the ozone simulations over the same regions presented in FetA09 and we will make use of exactly the same model data and observations used in by FetA09 as the main point of the investigation here is to use the same available information of Fetal09 to show that results are different when, an inspected MM ensemble is adopted. The inspection is based on the following steps:

- determine to what extent the variability (standard deviation about the ensemble mean as in Fortin et al., 2014) present in the observation is reproduced by the ensemble
- determine the minimum number of models necessary to represent the observed variability
- identification of the models forming the reduced MM ensemble used for subsequent analysis.

2.1.1 THE "ACCOUNTED FOR" VARIABILITY: EIGEN-ANALYSIS AND RANKED HISTOGRAM TECHNIQUE

The goal of this first analysis is to determine to what extent the observational variability is reproduced by the ensemble. An optimal situation is the one in which the variability of observations coincides with that produced by the ensemble of models, in other words the ensemble of the results all together covers the same range of variation of the measurements. Any deviation from this condition, namely a smaller or a larger variability of the MM ensemble with respect to the observed one would show, on one side, the incapacity of the ensemble to span the observed reality, or on the other, the addition of irrelevant information to the simulation of the observed situation. Therefore considering that a MM ensemble is assembled on an opportunity basis rather than results characteristics, this first step is of primary importance to estimate to what extent the gathered set is appropriate for the case study.

A technique to assess the variability and to estimate the redundancy of the MM ensemble with respect to that of the observations, was suggested by Annan and Hargreaves [2010] and applied in several MM ensemble modelling contexts (see, e.g. Solazzo et al., [2013]; Solazzo and Galmarini [2014]). It consists of projecting the observation anomalies (the element-wise difference between the observations and their mean) onto the principal components (PCs) of the covariance matrix of the deviation of the ensemble of models from the MM mean (the element-wise difference between each model realisation and the MM ensemble mean). Principal component analysis [Jolliffe, 2002] is probably the most well-known and widespread dimension-reduction technique. It is based on eigen-analysis to select uncorrelated directions associated with the largest variances.

When applied to the HTAP 21-member ensemble analysed by FetA09, this method shows that the first (largest) eigenvalue already explains more than 90% of the observational variability in most regions, the only exception being Japan with 60%. In other words, most of the ensemble members have a significant projection onto the first eigen-vector defining the major component, thus explaining the same portion of variance. If too many models are projected on the same eigenvector, it means that there are too many models producing repeating or ‘overlapping’ solutions (thus, the MM ensemble is redundant and overconfident). A well-behaved MM ensemble (not necessarily the theoretical case of independent models) should be made of a number of models whose eigenvalues contribute to the explanation of as many different components as the observational variability and the ratio model-to-observed variance should be close to unity. In the case of the HTAP MM ensemble, when all eigen-values are taken into account (and all of the associated eigen-vectors), the MM ensemble variance is 4.7, 6.0, 8.7 times the variance of the observation anomalies for the EU Mediterranean, Central 0-1 km, and Central 1-2 km, regions respectively. Concerning the US Mountains, Great Lakes, SE, NE, SW regions, the full MM ensemble mean accounts for 25.4, 9.1, 20.6, 10.7, 5.6 times the observed variability, respectively, and finally 4.7 times for the Japanese sub-region. According to the definition of Annan and Hargreaves [2010] the ensemble is therefore *wide*, i.e. its variability is larger than the observed one. Dealing with a wide ensemble implies that there is a substantial amount of redundant variability, i.e. variability already accounted for by other models. Not all information contained in the ensemble is needed in principle and needs to be reduced.

An alternative method to diagnose the variability spanned by an ensemble of models to the eigenvalues used is the Talagrand or Ranked Histogram (RH) [Talagrand et al., 1998], which provides an evaluation of the consistency of the ensemble with an observed quantity. In a RH the observations are ranked into a number of bins equal to the number of models making up the ensemble plus one for the extremes. The ensemble members are sorted to define ranges or "bins" of the modeled variable such that the probability of occurrence of the observation within each bin is, ideally, equal. The bins are determined by ranking the ensemble member from lowest to highest. The interval between each pair of ranked values forms a bin. To a N-member ensemble correspond N+1 bins [Hamill, 2001]. The underlying assumption is that each ensemble member in principle introduces an independent degree of variability. An indication of an ill-constructed ensemble is the ratio between the number of elements and the number of data available per model. If there are N models with time series each of size n_t (elements of the time series), the implication of $N > n_t$ is that there will be at least N-n_t empty bins in the RH, indicating redundancy of the ensemble and that the ensemble is inappropriate for the case analyzed. This same result could be visualized by looking at the load factors resulting from the decomposition in PCs: many projections would be null, as the number of eigen-vector is larger than the number of data to project. The HTAP MM ensemble used in this example, $N = 21$ and n_t =12. The RH for the nine sub-regions is reported in Fig. 2. Six (NA NE) to nine (NA SW) bins out of 22 are populated, (i.e. contain non-zero values), due to insufficient data and excess of redundant information. The use of the RH reveals another important problem with the FetA09 MM ensemble. Good ensemble practice would require $n_t >> N$. The plots clearly show that there are many empty bins (so degrees of freedom in the process that are not part of the reality as no observations are present in that range). The uneven distribution of the histograms shows that much emphasis (overconfidence) is given to some aspects of the process description, while others are neglected, that is another way of representing the redundancy obtained with PC analysis presented earlier

2.1.2 EFFECTIVE NUMBER OF MODELS

Having assessed that the ensemble is redundant it is important to determine the minimum number of models from those available in the MM ensemble that would suffice to describe the observational variability. A method developed by Bretherton et al. [1999], and firstly applied to air quality models by Solazzo et al. [2013] , quantifies the effective number of models sufficient to reproduce the variability of the observation as:

$$N_{eff} = \frac{(\sum_{k=1}^{N} \lambda_k)^2}{\sum_{k=1}^{N} \lambda_k^2} \qquad \text{Eq (1)}$$

with λ eigenvalue of the $corr(d_i,d_j)$ matrix, which contains the linear correlation coefficient between any pair d_i, d_j (i,j=1,…,N). d is a metric defined accordingly to Pennel and Reichler [2011]:

$$d_m = e_m - R\ \ MME \qquad \text{Eq (2)}$$

where the index m identifies the model, MME is the multi model error (the average of all individual model's errors) and R is the Pearson correlation coefficient between e_m, the error of model m and the MME. The removal of MME in Eq. (2) makes model errors more dissimilar from one another and uncovers "hidden" trends that are outweighed by overarching commonalities. Indeed the scope of the metric d_m is to determine similarities between models beyond the dominating ones induced by shared inputs and/or common parameterisations to the extent that the former are accounted for in the average. The relationship (1) should be interpreted as: only if all eigenvalues were equal to unity, Eq. (1) would take a value of $N_{eff} = N$, which corresponds to the situation where all directions are equally important and all models add independent contributions to the explanation of the observational variability. On the other hand, if all error fields were similar, only one eigenvalue would be non-zero and $N_{eff} = 1$. Equation (1) provides an analytical estimate of the dimensions of the subspace of models necessary to produce the information of the whole ensemble.

For the HTAP MM ensemble of FetA09, Eq. (1) gives N_{eff} ranging between ~2 and 4 for the regions analysed by FetA09 compared to the original 21 models. Thus, approximately three quarter of the available members participate to the ensemble with already 'accounted for' information. This is a revealing result that indicates paradigmatically the relevance of a pre-inspection of an ensemble. What seemed like a largely populated ensemble turns out to be incapable of capturing several degrees of freedom of observations and 2 to 4 members of 21 are sufficient to describe the observational variability. One may ask: if so, why is the average of the 21 models fitting so well with the observations as presented in FetA09? The answers could be: pure chance, since finally the model results participated out of good will, and happened to be there in the right mixture. Just consider what would have happened to the mean of the models should one of the two most evident outliers in Figure 1 decide to withdraw from the exercise. Alternatively an explanation could be the massive smoothing due to the monthly averaging along with the high level of tuning of the models around specific solutions that are normally distributed around the average observed data.

2.1.3 Reducing ensembles

As demonstrated in the previous sections, the HTAP MM ensemble is redundant and in particular 2 to 4 members are sufficient to represent the observational variability while the rest do not add any new information. Similarly, the extra elements are likely to deteriorate any evaluation metrics applied to the ensemble. At this point we know that the number of models that are necessary and sufficient is smaller than 21 but we do not know which combination of members for every grouping produces the optimal ensemble.

Given N members, there are $G=N!/[r!(N-r)!]$ possible groups of r elements. A straight forward way to identify the optimal ensemble (optimal sub set) and maximize the accuracy of the ensemble is to analyse all the G combinations of subsets of models and identify the one that minimize the Root Mean Square Error (RMSE). The latter is a measure of the accuracy (the even distribution of model results from the observed value), and high accuracy also improves precision (a reduced spread/scatter of the model results around the observed value). In principle measurement errors should be also taken into account in the procedure for reducing the ensemble, but in case where they are significantly smaller than the model ones, RMSE is sufficient measure.

In Fig (3) we report the curves of minimum, mean, and maximum RMSE for the nine sub-regions used by FetA09 as a function of the number of members of ensembles (r=2,…21). The figure confirms the results on the number of models necessary to maximize the ensemble performance and tells us that which combination of the 2 to 4 models out of 21 produces such improvement. The scores of the reduced ensemble are reported in Table 2 and are compared against the ones produced by the full ensemble mean. In all cases the mean of the reduced ensemble improves the accuracy (from 31% for NA NW to 71% for NA Mountain and NA Lakes) and precision (most notably for NA SE and NA NE). As it can be seen in several regions the use of the full MM ensemble of opportunity produces a clear deterioration in the ensemble statistics. In Table 2 we report also the ranking of the models contributing to minimize the error in the sub-regions. As from the table it is often the case that the error is minimized by mix-ranked (good performing and bad performing) of members. In fact, if the two best models have a high chance of being also highly correlated then they would share some portion of information, thus resulting redundant. Therefore when considering the ensemble mean of these two models, very little decrease in error would be found compared to the individual models. Mathematically, the theorems by Elashoff et al. [1967] and Cover [1974] have proven two important results on the selection of member and evaluation of individual scores: the best two models are seldom the combination of two models that maximises the score of an ensemble average, and furthermore, that the best single model may not appear in the ensemble maximising the feature score. As a result, the simple method of making ranked combinations of models with the best individual features may prove unsuccessful, as also demonstrated by e.g. Solazzo et al. [2013], Hannan and Hargreaves [2011], Kioutsioukis and Galmarini [2014], Knutti et al., [2010], and others. This confirms the importance of the inspection of the available results prior to their use and of having at disposal a large pool of models from which optimal subsets can be extracted.

3. IMPACT ON THE RESULTS OF EMISSION SENSITIVITY ANALYSIS OF AN INSPECTED VS UNINSPECTED ENSEMBLE

An important part of FetA09 relates to the sensitivity study on emission reduction. As part of the HTAP program the consequences of an emission reduction of 20% anthropogenic NOx in specific part of the globe where investigated using the MM ensemble available. Since we have demonstrated that the MM ensemble used in FetA09 is redundant and having identified the optimal number of elements and the most accurate set of models, one may wonder how the predicted consequences of the emission reduction on ozone concentration would change if we used the reduced ensemble.

We focused the analysis on the North-American region only. In FetA09 the use the mean of the full ensemble produced an average response in ozone concentration of -0.76 ppb in the NA region as a consequence of the reduction of NOx emission by 20%. We shall note that the NA region is subjected to the emission reduction and therefore the investigation includes the whole of the US and part of Mexico (Figure 1 of FetA09), and thus it has a spatial extension that includes the five NA sub-regions described in section 2 for the evaluation. Furthermore, of the 21 models participating to the evaluation part of the exercise, only 14 models results were made available for the simulation with reduced emission scenarios. Therefore, for the sake of consistency, we repeated the redundancy inspection for the 14-member ensemble and calculated the most accurate set through the minimization of RMSE described section 2.1.3. The size of the newly calculated subsets ranges between three for the Lakes, North-East, South-West, South- East of USA, and four for the Mountainous region. The newly calculated set obtained from the original 14 member ensemble produced an ozone concentration reduction of 2.32 ppb on average across all regions. That is 300% more than that found by FetA09. The largest variation is obtained for the South-East region of USA, with an ozone concentration decrease of 5.30 ppb that is a 5-fold than what obtained by FetA09. Such an analysis demonstrates how conclusions could change if the ensemble is not inspected a priori and reduced if necessary.

In the exploration of scenario or sensitivity to ideal conditions like that presented in HTAP, one may be tempted to construct an ensemble that only groups the best preforming models results in the evaluation against measurements and using only those in the sensitivity or scenario case study grouping them in an ensemble. This would be wrong in principle or in other words would not produce the best ensemble by definition as demonstrated by the already cited theorems of Elashoff et al. [1967] and Cover [1974].

4. CONCLUSIONS

Multi-model ensemble is becoming very popular in geophysical studies. In this paper we have been contrasting the results from an *ensemble of opportunity* where casually assembled model *phenotypical different* are the driving elements, with the results obtained when the same pool of model is screened to eliminate redundancy and the optimal combination is used.

The case of HTAP phase 1 is taken here as an example of a practice that is wide spread, especially in the realm of air quality, atmospheric dispersion at all scales. A very limited amount of studies apply correctly the technique. The HTAP case has been selected for two main reasons:

- The very large number of models that participated to the initiative and that were available for the ensemble analysis;
- the ensemble results were also used as basis to assess the consequences of an emission reduction strategy on ozone in several regions of the world.

The HTAP ensemble has been assessed against available measurements and the following conclusion were obtained:

- In spite of the large number of participating models, the scarcity of time steps produces an important level of redundancy as from the simple analysis of a ranked histogram.
- At smaller subset of model perform much better when compared to measurements and it is statistically more significant.
- In the case of HTAP [FetA09] the objective of the study was to determine, through a MM ensemble, the impact of emission changes produced in one continent on another. The analysis conducted on the impact over the same continent where the emissions

are produced, reveals that the conclusions remain the same as those produced by FetA09 but the values found are between 3 to 5 times higher when using a non-redundant ensemble.

These are problems that are common to many multi model studies and for which a minimum set of good practice rules should be taken into account (Kioutsioukis and Galmarini, 2014). Among these, we point out that in order to have any reasonable statistics the number of measurement should be much greater than the number of ensemble members. Otherwise rank histogram is simply not a proper tool for the analysis.

On a more general level, it is clear that the use of un-inspected ensembles of opportunities is a miss-practice that could lead to under-exploitation of the latter and in some case even wrong conclusions. Quantitative practices guarantee for the best possible diagnosis of the ensemble potential and its full exploitation. The availability of monitoring information is essential for the performance of the analysis presented here and it could be argued that the optimal ensemble identification is prone to the time and spatial representativity of the observations. This is true but as much as it is for the evaluation of any individual model result that depends on the space and time distribution of observation and the phenomenology represented.

The hemispheric transport case analyzed here brings to the attention also the issue of the space and timescale at which a set of model verified in a certain area could be used. The verification of the effect of the selection of an optimal set out of an ensemble based on data pertaining to a specific region and time frame, produces over another region, remains an important element of research. In other words, whether an optimal set selected for region *A* using observation in region A can be used for a region *B* and in a scenario or sensitivity analysis mode. Scale dependence of the atmospheric processes involved could become an issue in this case and will have to be verified. On the other end we consider the use of the optimal set for scenario and sensitivity study in the area where the observation used for its selection have been collected much more appropriate than the use of a full ensemble of opportunity. The selection of the optimal set through observations on a base case scenario is equivalent to the evolution of a single deterministic model and its application for speculative scenario analysis or forecast applications.

The representativity of the ensemble compared to observation and the minimization of the redundancy remain an important issue. In the light of that we speculate here, the use of multi-scale multi-model ensembles, constructed with the combinations of models covering different portions of the atmospheric power spectrum, could greatly improve the representativity and provide coverage of the problem in a much more detailed form. The combination of global and regional scale results, for example, in one ensemble is a possibility that will be explored in the framework of the next phase of HTAP.

ACKNOWLEDGMENTS

Dr Arlene Fiore (Columbia University) and the HTAP modeling community are acknowledged for making the model and observational data available for the current analysis (http://www.htap.org/) and for the openness to our investigation. Dr Frank Dentener (JRC) is acknowledged for the valuable comments that greatly improved this manuscript. The authors also thank Dr Brigitte Koffi (JRC) for having retrieved some of the data used in this paper.

REFERENCES

Anenberg S.C., J. J. West, A. M. Fiore, D. A. Jaffe, M.J. Prather, D. Bergmann, K. Cuvelier, F. J. Dentener, B. N. Duncan, Michael Gauss, Peter Hess, Jan Eiof Jonson, Alexandru Lupu, Ian A. MacKenzie, Elina Marmer, Rokjin J. Park, Michael G. Sanderson, Martin Schultz, Drew T. Shindell, Sophie Szopa, Marta Garcia Vivanco, Oliver Wild, Guang Zeng, (2009), Intercontinental Impacts of Ozone Pollution on Human Mortality, Environ. Sc.Tech, 43, 6482–6487

Annan, J.D., Hargreaves, J.C., (2010), Reliability of the CMIP3 ensemble. Geophys. Res. Lett., 37, p. L02703

Bishop, C.H., Abramowitz, G., (2013), Climate model dependence and the replicate Earth paradigm. Clim. Dyn 41, 885-900

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

Cover, T. T. (1974). The best two independent measures are not the two best, IEEE Trans. System Man. and Cybernetics, 4, 116–117.

Dentener F, T. Keating and H. Akimoto, (eds) , (2010), Hemispheric Transport of Airpollution, Part A, Ozone and Particulate Matter, Edited by F, Economic Commission for Europe, Air Pollution Studies, 17, ISBN, 978-92-1-117043-6, UNECE, Geneva.

Elashoff, J.D., Elashoff, R.M., Goldman, G.E., (1967), On the choice of variables in classification problems with dichotomous variables. Biometrika 54, pp. 668–670

Evans, J.P., Ji, F., Abramowitz, G., Ekstrom, M., (2013), Optimally choosing small ensemble members to produce robust climate simulations. Environ. Res. Lett. 8, 044050 (4pp).

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., (2009) Multimodel estimates of intercontinental source-receptor relationships for ozone pollution, J. Geophys. Res., 114, D04301, doi:10.1029/2008JD010816.

Fortin, V., Abaza, M., Anctil, F., Turcotte, R., (2014). Why should ensemble spread match the RMSE of the ensemble mean?. J.Hydrometeor 15, 1708-1713.

Fry M.M., V. Naik, J. J. West, M. D. Schwarzkopf, A.M. Fiore, W.J. Collins, F.J. Dentener, D. T. Shindell, C. Atherton, D. Bergmann, B. N. Duncan, P. Hess, I. A. MacKenzie, E. Marmer, M. G. Schultz, S. Szopa, O.Wild, G Zeng, (2012), The influence of ozone precursor emissions from four world regions on tropospheric composition and radiative climate forcing, J. Geophys. Res., 117, D7, doi:10.1029/2011JD017134.

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 (2004). Ensemble dispersion forecasting—Part I: concept, approach and indicators. Atmos. Environ., 38 (28), pp. 4619–4632

Hamill, T.M., (2001), Interpretation of rank histograms for verifying ensemble forecasts. Mon. Weather Rev., 129 (3), 550–560

Jolliffe, I., (2002), Principal component analysis, Springer, 2nd edition.

Jonson J.E., A. Stohl, A.M. Fiore, P. Hess, S. Szopa, O. Wild, G. Zeng, F.J. Dentener, A. Lupu, M.G. Schultz, B.N. Duncan, K. Sudo, P. Wind, M. Schulz, E. Marmer, C. Cuvelier, T.j. Keating, A. Zuber, A. Valdebenito, V. Dorokhov, H. De Backer, J. Davies, G.H. Chen, B. Johnson, and D.W. Tarasick, (2010), A multi-model analysis of vertical profiles, Atmos. Chem. Phys. 10, 5759-5783.

Knutti, R. , (2010), The end of model democracy?, Climate Change, 102, 395–404.

Knutti, R., Furrer, R., Tebaldi, C., Cermak, J., Meehl, G.A., (2010). Challenges in combining projections from multiple climate models. American Meteorological Society 23, 2739-2758.

Kioutsioukis I. and S. Galmarini, (2014), De praeceptis ferendis: good practice in multi-model ensembles, Atmos. Chem. Phys. Discuss., 14, 15803-15865

Pennel, C., Reichler, T., (2011), On the effective numbers of climate models J. Clim., 24, 2358–2367

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

Potempski, S., Galmarini, S., (2009), Est modus in rebus: analytical properties of multi-model ensembles. Atmos. Chem. Phys. (2009), pp. 9471–9489

Riccio, A., Ciaramella, A., Giunta, G., Galmarini, S., Solazzo, E., Potempski, S., 2012. On the systematic reduction of data complexity in multi-model ensemble atmospheric dispersion modelling. Journal Geophysical Research 117, D05314.

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 S. Galmarini, (2012): Ensemble modelling of surface level ozone in Europe and North America in the context of AQMEI, Atmos. Environ., 53, 60–74.

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

Solazzo, E., Galmarini, S., (2014), The Fukushima-^{137}Cs deposition case study: properties of the multi-model ensemble. J. Environ. Radioact. 139, 226-233 http://dx.doi.org/10.1016/j.jenvrad.2014.02.017.

Talagrand, O., Vautard, R., and Strauss B., (1998), Evaluation of probabilistic prediction systems, paper presented at aa seminar on predictability, Eur. cent. For Medium Weather Forecasting, Reading (UK).

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., (2012), Evaluation of the meteorological forcing used for AQMEII air quality simulations, Atmos. Environ., 53, 15–37

Weigel, A.P., Liniger, M.A., Appenzeller, C., (2008). Can multi-model combination really enhance skill of probabilistic ensemble forecast? Q.J.R. Meteorolo. Soc. 134, 241-260.

Table 1. Number of effective models N_{eff} for the sub-regions object of the analysis (with reference to Figure 2 of Fiore et al (2009) top panel, based on corr(*di,dj*)). *nrec* is the number of surface receptors used for evaluation

Sub-region	N_{eff}
EU Mediterranean region (nrec=6)	4.0
EU central region 0-1 km (nrec=24)	3.1
EU central region 1-2 km (nrec=11)	3.5
NE-USA (nrec=13)	1.9
SW USA (nrec=5)	1.8
SE USA (nrec=6)	1.9
Great Lakes USA (nrec=8)	2.0
Mountainous USA (nrec=10)	1.8
Japan EANET (nerc=10)	2.6

Table 2. RMSE-ranking and scores of the reduced MM ensemble mean for the sub-regions object of the analysis (RMSE: Roor-Mean-Square-Error; PCC: Pearson Correlation Coefficient; σ: ratio of the modelled to the observed standard deviation)

Domain	**Ranking of the MinRMSE combination**	**score**
EU central 0-1 km	1,15,19	RMSE=1.69 (2.65) PCC=0.98 (0.96) σ=0.99 (1.10)
EU central 1-2 km	7,17,18	RMSE=3.35 (9.2) PCC=0.98 (0.95) σ=1.03 (1.25)
EU medit	4,6,13,15,19	RMSE=0.76 (1.44) PCC=0.99 (0.98) σ=1.0 (1.13)
NA SW	8,10,11,15	RMSE=2.0 (2.9) PCC =0.95 (0.96) σ=0.87 (0.86)
NA SE	1,2,4,8	RMSE=3.61 (10.27) PCC=0.77 (0.62) σ=0.83 (1.81)
NA NE	3,5,6,7	RMSE=3.01 (7.8) PCC=0.93 (0.90) σ=0.90 (1.56)
NA Mountain	1,5,12	RMSE=1.53 (5.33) PCC=0.93 (0.90) σ=1.04 (1.44)
NA Lakes	1,5,6	RMSE=1.89 (6.58) PCC=0.97 (0.91) σ=1.03 (1.45)
Japan EANET	12,15	RMSE=3.11 (5.70) PCC=0.96 (0.79) σ=0.66 (0.51)

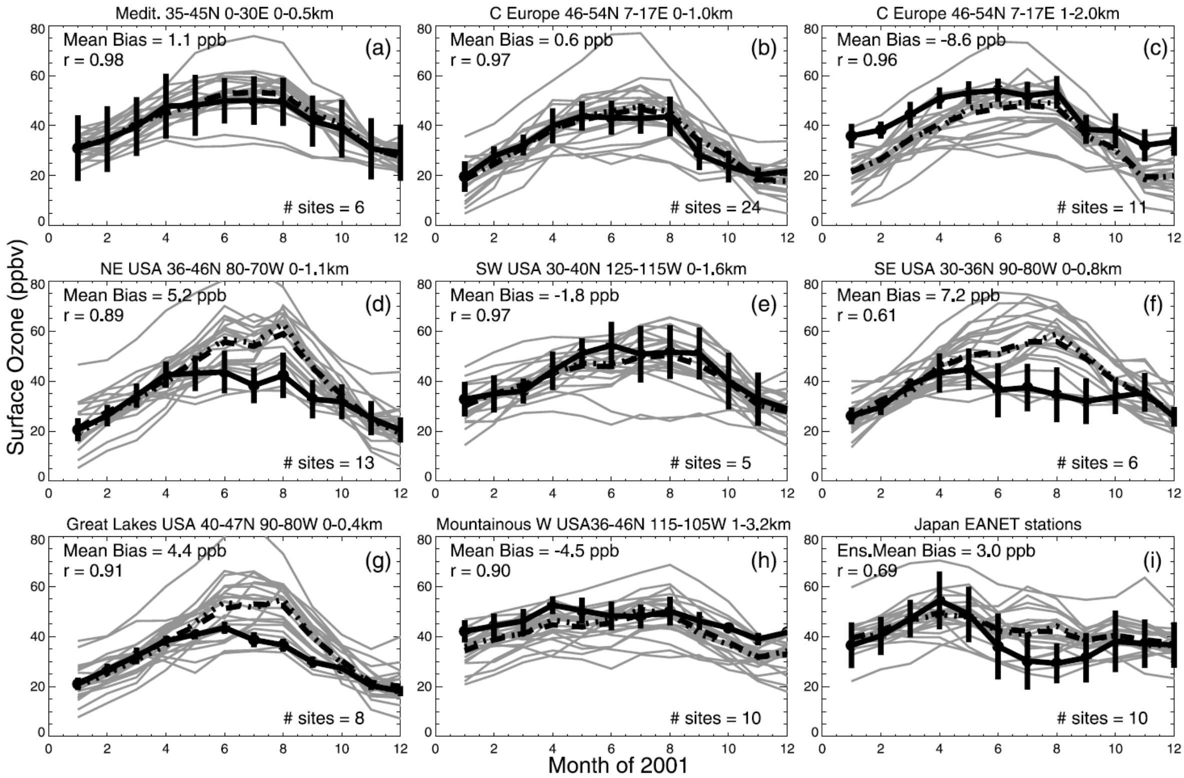

Fig 1: From Fiore et al. (2009): Monthly mean surface O3 concentrations (ppb) for the year 2001. Observed values (black circles) represent the average of all sites falling within the given latitude, longitude, and altitude boundaries and denoted by the symbols in Figure 1; vertical black lines depict the standard deviation across the sites. Monthly mean O3 in the surface layer of the SR1 simulations from the 21 models are first sampled at the model grid cells containing the observational sites and then averaged within subregions (gray lines); these spatial averages from each model are used to determine the multimodel ensemble median (black dotted line) and mean (black dashed line). Observations are from CASTNET (http://www.epa.gov/castnet/) in the United States, from EMEP (http://www.nilu.no/projects/ccc/emepdata.html) in Europe, and from EANET (http://www.eanet.cc/eanet.html) in Japan.

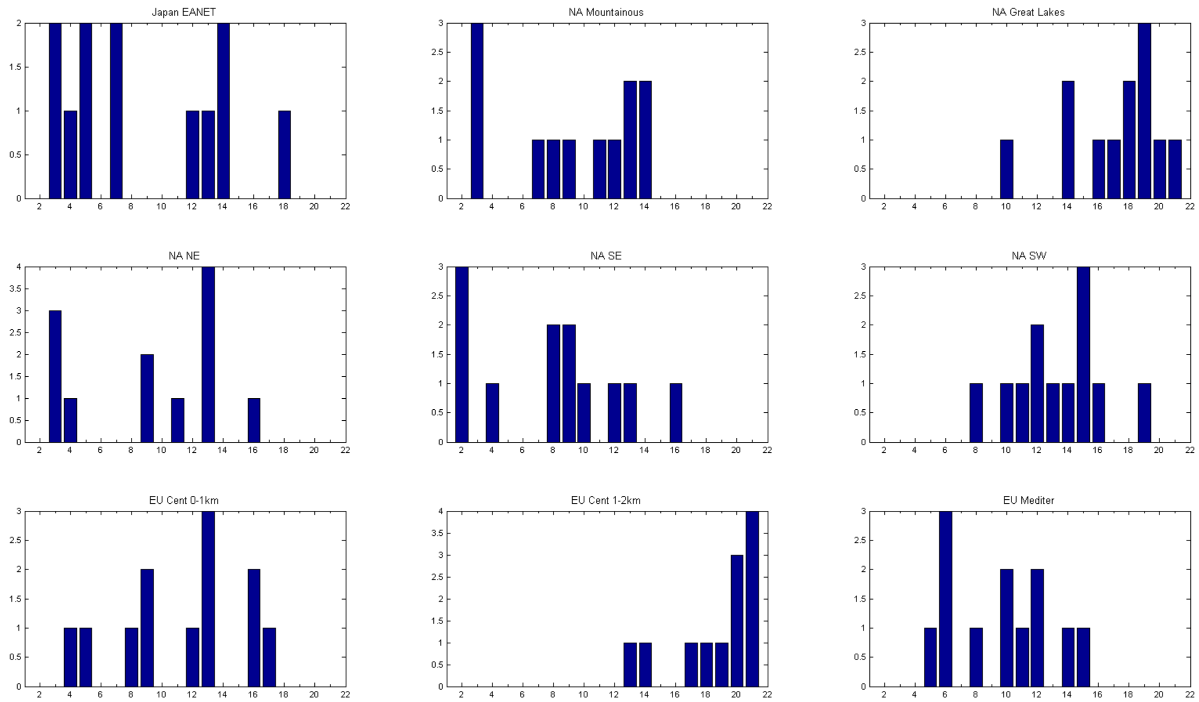

Fig.2 Ranked histogram for the nine sub-regions subject to MM ensemble evaluation

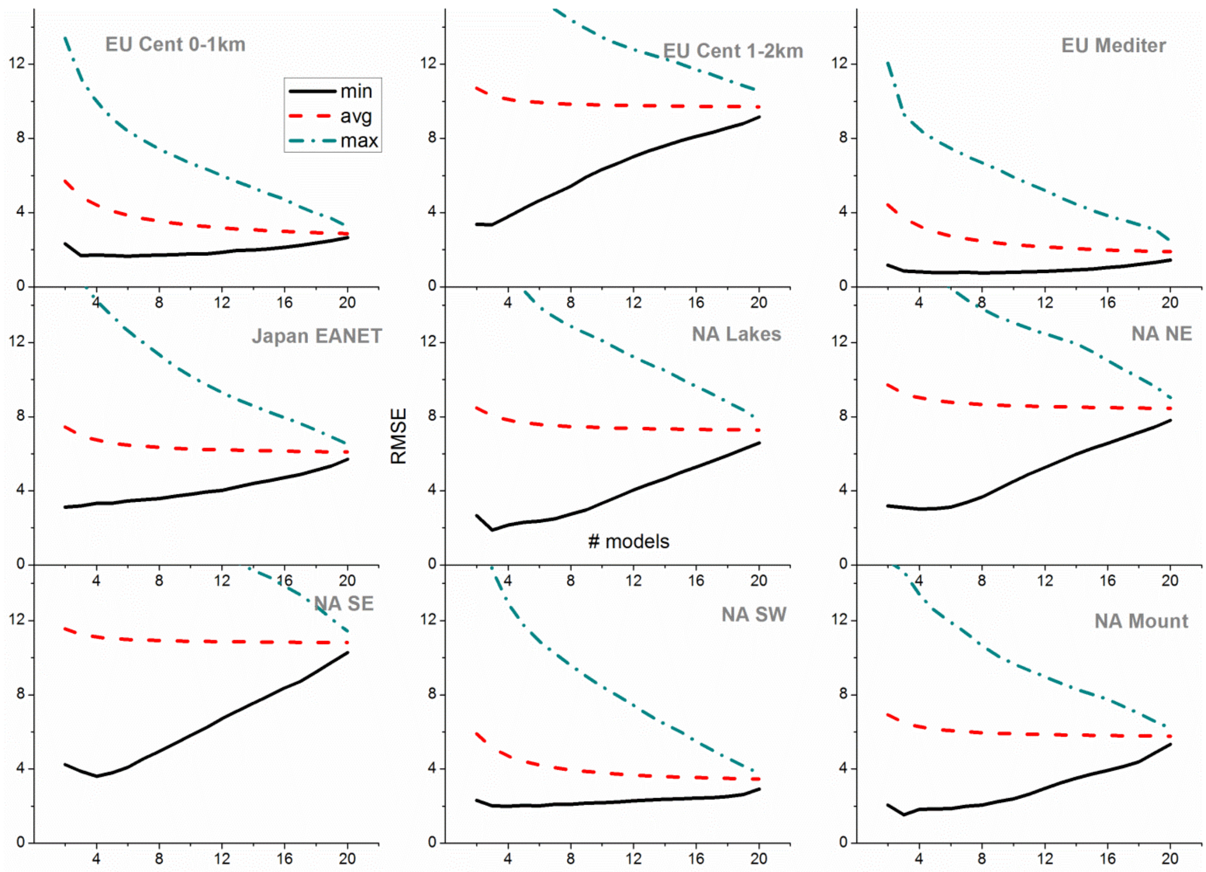

Fig 3 Maximum (dash-dot), average (dashed), and minimum (continuous line) RMSE for all subsets of MM combinations and for the nine sub-regions subject to MM ensemble evaluation.