Atmos. Chem. Phys., 7, 6085–6098, 2007
www.atmos-chem-phys.net/7/6085/2007/

Atmospheric Chemistry and Physics

Seeking for the rational basis of the Median Model: the optimal combination of multi-model ensemble results

A. Riccio[1], **G. Giunta**[1], **and S. Galmarini**[2]

[1]Dept. of Applied Science, University of Naples "Parthenope", Napoli, Italy
[2]European Commission – DG Joint Research Centre, Institute for Environment and Sustainability, Ispra, Italy

Received: 23 March 2007 – Published in Atmos. Chem. Phys. Discuss.: 27 April 2007
Revised: 21 November 2007 – Accepted: 1 December 2007 – Published: 11 December 2007

Abstract. In this paper we present an approach for the statistical analysis of multi-model ensemble results. The models considered here are operational long-range transport and dispersion models, also used for the real-time simulation of pollutant dispersion or the accidental release of radioactive nuclides.

We first introduce the theoretical basis (with its roots sinking into the Bayes theorem) and then apply this approach to the analysis of model results obtained during the ETEX-1 exercise. We recover some interesting results, supporting the heuristic approach called "median model", originally introduced in Galmarini et al. (2004a, b).

This approach also provides a way to systematically reduce (and quantify) model uncertainties, thus supporting the decision-making process and/or regulatory-purpose activities in a very effective manner.

1 Introduction

Standard meteorological/air quality practice, such as the prediction of the future state of the atmosphere, typically proceeds conditionally on one assumed model. The model is the result of the work of many area-expert scientists, e.g. meteorologists, computational scientists, statisticians, and others.

Nowadays, several models are available for the forecast of variables of meteorological and/or air quality interest, but, even when using the same ancillary (e.g. initial and boundary) data, they could give different answers to the scientific question at hand. This is a source of uncertainty in drawing conclusions, and the typical approach, that is of conditioning on a single model deemed to be "the best", ignores this source of uncertainty and underestimates the possible effects of a false forecast.

Correspondence to: A. Riccio
(angelo.riccio@uniparthenope.it)

Ensemble prediction aims at reducing this uncertainty by means of techniques designed to strategically sample the forecast pdf, e.g. the breeding of growing modes (Toth and Kalnay, 1993) or singular vectors (Molteni et al., 1996) in the weather forecasting field.

Recently, a number of works in air quality modeling (Delle Monache and Stull, 2003; Pagowski et al., 2005, 2006a; Pagowski and Grell, 2006b; Mallet and Sportisse, 2006; Delle Monache et al., 2006a, b, c; Zhang et al., 2007) successfully applied different techniques to demonstrate the advantage of deterministic ensemble forecasts compared with forecasts provided by individual models.

The advantages of ensemble prediction are twofold:

- ensemble estimates average out non-predictable components, and,
- provide reliable information on uncertainties of predicted parameters from the diversity amongst ensemble members.

Recently, the multi-model ensemble prediction system (Krishnamurti et al., 1999) has been introduced. Instead of conditioning on a single (ensemble) modeling system, the results from different climate forecasting models are combined together. The so-called "superensemble" system demonstrated to be far superior, in terms of forecasts, to any ensemble mean.

The multimodel approach has been successfully applied also to atmospheric dispersion predictions (Galmarini et al., 2001, 2004a, b) where the uncertainty of weather forecast sums and mixes with that stemming from the description of the dispersion process. The methodology relies on the analysis of the forecasts of several models used operationally by national meteorological services and environmental protection agencies worldwide to forecast the evolution of accidental releases of harmful materials. The objectives are clear: after the release of hazardous material into the atmosphere, it

is extremely important to support the decision-making process with any relevant information and to provide a comprehensive analysis of the uncertainties and the confidence that can be put into the the dispersion forecast. Galmarini et al. (2004a) showed how the intrinsic differences among the models can become a useful asset to be exploited for the sake of a more educated support to decision making by means of the definition of ad-hoc parameters and treatments of the model predictions. Among them the definition of the the so-called Median Model defined as a new set of model results constructed from the distribution of the model predictions. The Median Model was shown to be able of outperforming the results of any single deterministic model in reproducing the cloud measured during the ETEX experiment (Girardi et al., 1998).

At the end of their paper Galmarini et al. (2004b) mention: "At present we are not in the position of providing a rigorous explanation on why the median model should perform better than the single models.". . ."Furthermore the conclusions presented in this paper should be generalized and placed in a more rigorous theoretical framework".

This work moves its steps from the above mentioned sentences. In particular we will focus on the second statement as the first seems to fish deep in the conundrums of theoretical statistics. More explicitly the questions tackled here are:

1. is it possible to place the multimodel ensemble approach within a sound theoretical framework?
2. how to quantify the discrepancies between each ensemble member and observations?
3. And between ensemble-based predictions and observations?
4. In the case of ensemble-based simulations, predictions are obtained by merging results from each member. It is reasonable to suppose that ensemble member predictions are correlated. Even in the case of multimodel simulations, it is expected that results from different models are correlated, since they often share similar ancillary data, e.g. input data, physics parameterizations, numerical approaches, and so on. In the case of "correlated models", we expect that data are "clustered", thus biasing the ensemble-based results and producing too much optimistic confidence intervals. How to work around these problems?
5. Can some of the parameters described in Galmarini et al. (2004a) be presented in a coherent theoretical framework?

In this work we used a well-known statistical approach to multimodel data analysis, i.e. Bayesian Model Averaging (BMA), which is a standard method for combining predictive distributions from different sources. The BMA predictive probability density function (pdf) of any quantity of interest is a weighted average of pdfs centered on the individual bias-corrected forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts.

More specifically the objectives of this work consist in the:

- evaluation of the BMA weights, in order to sort the predictive skill of models;
- quantification of the systematic bias of each model;
- estimation of some useful statistical indexes introduced in Galmarini et al. (2004a; 2004b),
- exploration of similarities and differences between our approach and the "median model",
- quantification of the correlations between models, as a measure of interdependency.

First, we introduce the theoretical context (the Bayesian framework), under which ensemble modeling, and much other, can be placed. In Sect. 3 the BMA approach is described; this approach provides the way to interpret the weights used to combine the ensemble members results. Next (Sect. 4), we introduce the notion of independence and advance some suggestions about how to take into account the relations among models. In Sect. 5 a Bayesian hierarchical model, implementing the procedure to calculate the weights and the bias of each model, is derived and applied to the test case of the ETEX-1 experiment. The results are analyzed and discussed, bringing the "median model" heuristically introduced by Galmarini et al. (2004a, 2b) into a theoretical framework.

2 Bayes theorem and ensemble prediction

The Bayes theorem plays a fundamental role in the fields of ensemble modeling, data assimilation, sensitivity and uncertainty analysis. The Bayesian view has been acknowledged to be the most natural approach for combining various information sources while managing their associated uncertainties in a statistically consistent manner (Berliner, 2003).

The optimal combination of ensemble members has its roots in the Bayes theorem. Essentially, the Bayes theorem may be expressed as

$$p(\text{final analysis}|\text{ens data}) \propto p(\text{ens data}|\text{final analysis}) \times p(\text{final analysis}).$$

The power of the Bayes theorem relies on the fact that it relates the quantity of interest, the probability that the 'final analysis' is true given the data from the ensemble, to the probability that we would have observed the data if the final analysis were true, that is to the likelihood function. The last term on the right side, p(final analysis), the prior probability, represents our state of knowledge (or ignorance) about

the "true state" (the final analysis) before data have been analyzed; p(ens data|final analysis) is the likelihood function; the product of the two yields the posterior probability function, that is our state of knowledge about the truth in the light of the data. In a sense, the Bayes theorem can be seen as a learning process, updating the prior information using the data from the ensemble predictions.

For sake of clarity, it is useful to briefly review the key equations in an ensemble prediction system. The practical implementation of Bayes theorem requires the specification of a suitable probability model for each ensemble member. For example, consider two ensemble members. If each $p\times 1$ ensemble member state, $x_{\{1,2\}}$, is (multivariate) normally distributed

$$\begin{cases} x_1 = x + \varepsilon_1 \\ x_2 = x + \varepsilon_2 \end{cases} \tag{1}$$

where the $p\times 1$ vector x is the "true" (final analysis) state and ε_1 and ε_2 are (multivariate) normally distributed errors with mean zero and covariances Σ_1 and Σ_2, respectively, then the Bayesian posterior solution equals to

$$x|x_1, x_2 \sim \mathcal{N}(x_a, \Sigma)$$

with the final analysis x_a, and corresponding error covariance Σ, given by

$$\begin{cases} \Sigma^{-1} x_a = \Sigma_1^{-1} x_1 + \Sigma_2^{-1} x_2 \\ \Sigma^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1} . \end{cases} \tag{2}$$

The notation "$\sim \mathcal{N}(\mu, R)$" means distributed as a multivariate normal distribution with mean μ and covariance R.

Therefore, the data from the two ensemble members, x_1 and x_2, can be merged into an optimal estimate, the final analysis, x_a, provided that the linearity and gaussianity assumptions in (1) are a realistic representation of the process and one can estimate the matrices Σ_1 and Σ_2. Moreover, the combination of the two members is optimal in the log score sense, i.e.

$$-\mathrm{E}\left[\log p\,(x_a)\right] \leq -\mathrm{E}\left[\log p\left(x_{\{1,2\}}\right)\right]$$

since the precision (i.e. the inverse of the covariance matrix) of the final analysis is the sum of the precision of each member. In other words, the optimal combination makes the posterior distribution sharper and the MAP (maximum a posteriori) estimate less uncertain.

We can put a step forward this analysis, by using the Bayes theorem to combine the results of a multimodel ensemble prediction system into a skillful and well-calibrated final analysis. Krishnamurti et al. (2000) has defined this entity a "superensemble approach".

3 The BMA approach

Consider the following scenario: instead of relying on one assumed model, a researcher gathered data concerning the state of the atmosphere from different meteorological centers. The advantages of comparing different models are evident: each model is an imperfect representation of the real world and contains several approximations/parameterizations/lack of physics representations, etc.. Inferences obtained from a single model is risky, since they do not take into account for the model uncertainties. On the other hand, the comparison among several models may highlight the models' deficiencies, since it is highly unlikely that each physical phenomenon is equally represented by all models. The drawbacks of ignoring model uncertainties have been recognized by many authors a long time ago (e.g., see the collection of papers in Dijkstra, 1988), but little attention has been devoted until now.

The problem is how to combine the results from different models in a skillful summary. In the statistical literature the problem of comparing/combining results from different models is a long-standing approach. In his seminal book, Theory of Probability, Jeffreys (1961) developed a methodology for quantifying the evidence in favor of a given model/hypothesis. He introduced the Bayes factor which is the posterior odds of two hypotheses when their prior probabilities are equal.

In order to introduce the Bayes factor, assume that data x have arisen from two competing hypotheses/models, M_1 and M_2, according to a likelihood function $p(x|M_1)$ and $p(x|M_2)$. Given a priori probabilities $p(M_1)$ and $p(M_2)$=1−$p(M_1)$, the data produce a posteriori probabilities $p(M_1|x)$ and $p(M_2|x)$=1−$p(M_1|x)$. From the Bayes theorem, we obtain

$$p(M_k|x) = \frac{p(x|M_k)p(M_k)}{p(x|M_1)p(M_1) + p(x|M_2)p(M_2)} \text{for } k=1,2\,, \tag{3}$$

so that,

$$\frac{p(M_1|x)}{p(M_2|x)} = \frac{p(x|M_1)}{p(x|M_2)}\frac{p(M_1)}{p(M_2)}\,,$$

and the transformation from prior to posterior odds is simply the multiplication by the Bayes factor

$$B_{12} = \frac{p(x|M_1)}{p(x|M_2)}\,.$$

In other words,

posterior odds = Bayes factor × prior odds .

If the two models are equally probable a priori, the Bayes factor immediately provides the evidence for the first model with respect to the second one, by transforming the prior opinion through considerations on the data.

In the case of multiple competing models, Eq. (3) can be easily generalized to

$$p(M_k|x) = \frac{p(x|M_k)p(M_k)}{\sum_{k=1}^{K} p(x|M_k)p(M_k)} \text{ for } k=1,2,\ldots,K\ , \tag{4}$$

and, as usual in any Bayesian analysis, the posterior inference of a quantity of interest, say θ, e.g. a future observation or a model parameter, can be obtained from its ppd (posterior predictive distribution), i.e.

$$p(\theta|x) = \sum_{k=1}^{K} p(\theta|M_k, x)\, p(M_k|x)\,. \tag{5}$$

In this case, the ppd is the average of the posterior distribution over all models, each weighted by their posterior probabilities. The weights come from (4) and can be used to assess the usefulness of ensemble members, i.e. as a basis for selecting the most skillful model ensemble members: high (close to one) posterior model probability, $p(M_k|x)$, provides the quantitative basis to estimate the usefulness of model k in predicting the parameter of interest, thus playing the same role as Bayes factors for multiple competing models.

Model (5) is known as BMA (Bayesian Model Average) in the statistical literature. BMA works around the problem of conditioning on a single model, taking into account for the information from different models.

Recently, Raftery and Zheng (2003) reviewed the properties of BMA. There also several realistic simulation studies on the performance of BMA in different contexts, e.g. in linear regression (Raftery et al., 1997), loglinear models (Clyde, 1999), logistic regression (Viallefont et al., 2001), wavelets (Clyde and George, 2000) and medium-range weather forecasting models (Raftery et al., 2005).

3.1 The properties of BMA

In their paper, Raftery et al. (2005) developed an EM-based (Expectation Maximization) algorithm to estimate the parameters in Eq. (5). They were interested in the calibration of the University of Washington mesoscale short-range multimodel ensemble system (Grimit and Mass, 2002). They used normal distributions to model the uncertainty of each ensemble member, but different distributions may be used, as well. A plug-in implementing BMA is freely available for the R statistical software.

Apart from implementation details, several analytical results can be derived. It can be shown that the posterior BMA mean and variance are:

$$\begin{cases} E\,[\theta|x] \;\; = \sum_{k=1}^{K} \hat{\theta}_k\, p\,(M_k|x) \\ \\ \mathrm{Var}\,[\theta|x] = \sum_{k=1}^{K} \left\{ \left(\hat{\theta}_k - \sum_{i=1}^{K} \hat{\theta}_i\, p(M_i|x)\right)^2 + \right. \\ \qquad\qquad \left. +\mathrm{Var}\,[\theta_k|M_k, x] \right\} p\,(M_k|x)\;\;, \end{cases} \tag{6}$$

where $\hat{\theta}_k = E\,[\theta|M_k, x]$, i.e. the expected value of θ conditional on model k alone, i.e. having assumed $p\,(M_k|x) = 1$.

As can be seen from Eq. (6), the expected value is the weighted average over all models, and the variance is decomposed into two terms: the first term takes into account the between-models ensemble variance, i.e. the spread of the ensemble prediction, while the second term the within-models ensemble variance, i.e. the internal uncertainty of each model. Verbally,

Predictive variance = between ens. variance
+withinens. variance

It can be presumed that within-ensemble variance does not capture all the sources of uncertainty. In an ensemble approach, the estimation of confidence intervals, based only on the ensemble spread, may be optimistic, because they do not properly take into account the internal variability of the model, so that the output of any predicted variable may be not calibrated. By calibrated we mean simply that intervals or events that we claim to have probability p happen a proportion p of the time on average in the long run. For example, a 90% prediction interval veryfing at a given time and place is defined so that 90% of verification observations effectively lay between the 90% upper and lower bounds. Uncalibrated ensemble predictions tend to be under-dispersive, and this behavior has often been observed (see Coelho et al., 2004, as an example of an application of a model ensemble approach to a climatological problem). Of course, BMA is well calibrated on the training dataset, but it has been shown that it also gives satisfactory results for the predicted observations (Raftery et al., 2005).

Another interesting result is the correlation of the model ensemble error with the ensemble spread. Equation (6) provides a theoretical basis for this finding, since it relates the predictive model ensemble variance to the between-model ensemble variance. Whitaker and Loughe (1998) provide several examples from real-world meteorological ensemble data, showing the relationship between error and spread; see also Raftery et al. (2005) for a more-in-depth discussion of error-spread correlation in BMA modeling.

4 Independence and correlation

If different models are used to simulate the same phenomenon, e.g. weather, climate or the dispersion of radioactive material, they probably will give similar responses. Now, suppose that all model results agree in giving a wrong prediction; without any observational support, this situation cannot be discerned. Potentially, model ensemble results may lead to erroneous interpretations, and this is more probable if models are strongly dependent (i.e. all biased toward the wrong answer). We can say that a dependent model does not convey "newly fresh information", but it replicates the (wrong/right) answer given by the previous models.

Technically, independence can be defined by the joint/marginal probability densities. Let us denote by $p(y_1, y_2)$ the joint pdf of two random variables, y_1 and y_2; denote by $p_1(y_1)$ the marginal pdf of y_1, and similarly for y_2. Then y_1 and y_2 are independent if, and only if, the joint pdf

is factorizable in the product of the corresponding marginal pdfs, i.e.

$$p(y_1, y_2) = p_1(y_1)p_2(y_2)\,. \tag{7}$$

The extension to any number K of random variables can be straightforwardly defined, in which case the joint density is the product of K terms.

This definition can be used to derive an important property of independent random variables. Given two functions, f_1 and f_2, we have

$$E[f_1(y_1)f_2(y_2)] = E[f_1(y_1)]E[f_2(y_2)]\,. \tag{8}$$

This can be easily proved by applying (7).

$$\begin{aligned} &E[f_1(y_1)f_2(y_2)] \\ &= \int\int f_1(y_1)f_2(y_2)p(y_1, y_2)dy_1dy_2 \\ &= \int f_1(y_1)p(y_1)dy_1 \int f_2(y_2)p(y_2)dy_2 \\ &= E[f_1(y_1)]E[f_2(y_2)]\,. \end{aligned} \tag{9}$$

Equality in Eq. (7) means that the statistical properties of any random variable cannot be predicted from the others; for example, if a relationship such as $y_2{=}f(y_1)$ holds, the joint pdf is not factorizable because $p(y_2|\,y_1) \neq p(y_2)$.

In the case of independent random variables the interpretation of BMA weights is meaningful. For example, if we have three independent models, then

$$\begin{aligned} &E[\pi_1 y_1 + \pi_2 y_2 + \pi_3 y_3] \\ &= \pi_1 E[y_1] + \pi_2 E[y_2] + \pi_3 E[y_3]\,. \end{aligned} \tag{10}$$

But, if we suppose that the third model is linearly related to the others, i.e. $y_3{=}a_{31}y_1{+}a_{32}y_2$, it is straightforward to show that

$$\begin{aligned} &E[\pi_1 y_1 + \pi_2 y_2 + \pi_3 y_3] \\ &= (\pi_1 + a_{31}\pi_3)E[y_1] + (\pi_2 + a_{32}\pi_3)E[y_2]\,. \end{aligned} \tag{11}$$

This example shows the difficulties in the interpretation of BMA weights: if models are linearly dependent, they cannot be strictly identified.

The concept of independence is central in information theory, and several measures of independence has been developed, as for example mutual information or negentropy, e.g. see Cover and Thomas (1991) or Papoulis (1991).

Usually variables are not independent, but it is possible to find a proper transformation, say $z_1{=}g_1(y_1, y_2)$ and $z_2{=}g_2(y_1, y_2)$, so that the transformed variables are independent. Unfortunately, there is no general way to select the proper transformation, nor the mutual information or negentropy can be easily calculated, but, if the definition of independence is relaxed, some general and interesting results can be obtained.

A weaker form of independence is uncorrelatedness. Two random variables are uncorrelated if their covariance is zero:

$$E[y_1 y_2] = E[y_1]E[y_2]\,, \tag{12}$$

which follows directly from (8), taking $f_1(y_1){=}y_1$ and $f_2(y_2){=}y_2$. On the other hand, uncorrelatedness does not imply independence. For example, as shown by Hyvarinen and Oja (2000), assume that (y_1, y_2) are discrete-valued variables and follow such a distribution that the pairs are, with probability 1/4, equal to any of the following values: (0,1), (0,−1), (1,0), (−1,0). Then y_1 and y_2 are uncorrelated, as can be simply calculated, but

$$E[y_1^2 y_2^2] = 0 \neq \frac{1}{4} = E[y_1^2]E[y_2^2]\,.$$

Because the condition in Eq. (8) is violated, y_1 and y_2 are not independent.

In some special cases, uncorrelatedness implies independence. This is the case for normally (or lognormally) distributed data. For example, denote by $\mathbf{\Sigma}$ the covariance matrix of K-dimensional normally distributed data, then

$$p(\mathbf{y}) \propto \exp\left\{-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\bar{\mathbf{y}})\right\}\,. \tag{13}$$

If the $\mathbf{y}$s are uncorrelated, $\mathbf{\Sigma}^{-1}$ is a diagonal matrix. Then, by the properties of the exponential function, Eq. (13) can be written as the product of K functions, each dependent on only one component, i.e.:

$$\begin{aligned} &\exp\left\{-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\bar{\mathbf{y}})\right\} = \\ &= \prod_{k=1}^{K}\exp\left\{-\frac{1}{2}(y_k-\bar{y}_k)^T\Sigma_k^{-1}(y_k-\bar{y}_k)\right\} \end{aligned} \tag{14}$$

satisfying the definition of independence in Eq. (7). Even if variables are correlated, they can be made uncorrelated if the frame of reference is properly roto-translated. Let $\mathbf{U\Lambda U}^T{=}\mathbf{\Sigma}$ the eigendecomposition of the covariance matrix. The projection of the original variables onto the directions represented by the eigenvectors of $\mathbf{\Sigma}$, i.e. $(\mathbf{z}-\bar{\mathbf{z}}){=}\mathbf{U}^T(\mathbf{y}-\bar{\mathbf{y}})$, allows to obtain independently distributed variables, as can be easily proved:

$$\begin{aligned} &\exp\left\{-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathbf{\Sigma}^{-1}(\mathbf{y}-\bar{\mathbf{y}})\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{y}-\bar{\mathbf{y}})^T\mathbf{U\Lambda}^{-1}\mathbf{U}^T(\mathbf{y}-\bar{\mathbf{y}})\right\} \\ &= \exp\left\{-\frac{1}{2}(\mathbf{z}-\bar{\mathbf{z}})^T\mathbf{\Lambda}^{-1}(\mathbf{z}-\bar{\mathbf{z}})\right\}\,. \end{aligned} \tag{15}$$

See Fig. 1 for a fictitious example of bivariate, normally distributed, data.

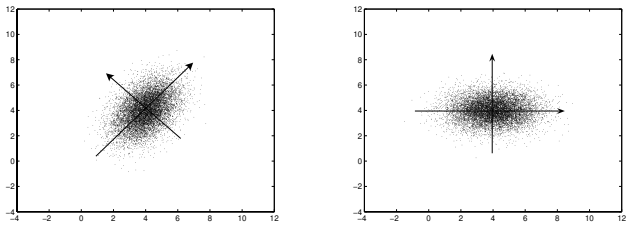

Fig. 1. An example of bivariate normally distributed data. On the left the data in the original frame of reference; on the right the same data, projected onto the eigenvectors of the covariance matrix, so that the two new directions are uncorrelated. The arrows indicate the axes of the ellipsoid.

Other measures, such as mutual information or negentropy, are much more difficult to calculate than correlations; so the eigendecomposition of the covariance matrix may be seen as a viable approximation to explore dependences between data or highlight the role of systematic deficiencies of model results, as will be shown in Sect. 6.

5 The estimation procedure

Now we have all the elements to proceed with the analysis of the results of the multi-model ensemble that will constitute our case study. The ensemble analysed in this work is an extended version of that originally analysed by Galmarini et al. (2004b). To summarize we will be looking at 25 simulations of the ETEX-1 release (Girardi et al., 1998) performed by independent groups world wide. Each simulation and therefore each ensemble member is produced with different atmospheric dispersion models and is based on weather fields generated by (most of the time) different Global Circulation Models (GCM). All the simulation relate to the same release conditions. For details on the groups involved in the exercise and the model characteristics refer to Galmarini et al. (2004b). Nine additional sets are presently available for this analysis. These include one set of results from the Danish Meteorological office (DMI), one set from the Korean Atomic Energy Agency, three sets from the Finnish met service (FMI), one set from UK-Metoffice, three sets from Meteo-France. In this study we also took care to mask the origin of the sets as we are not interested in ranking the model results. However in order to allow for the inter-comparability of the present results with those previously obtained by Galmarini et al. (2004b) we have kept the same coding for the original 16 members (m1-m16) that were used therein and added 9 additional codes (m17-m25) for the newly available sets randomly associated to the new models listed above.

Using the Bayes' theorem, model parameters can be estimated from the posterior pdf. Hereafter z_i denotes the ith observation and y_{ik} the corresponding predicted value from

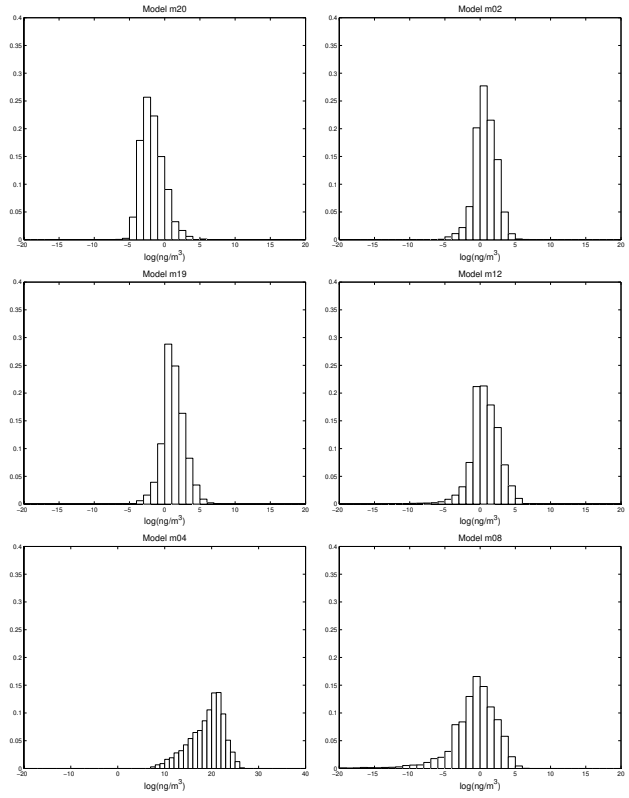

Fig. 2. Histogram of the differences between model results and corresponding observations for some selected models. From left to right, and then from top to bottom: m20, m02, m19, m12, m04 and m08. Logarithms were taken for both the model results and observations.

the kth model. The BMA posterior pdf reads

$$p(\theta|\,\pi_{\cdot}, y_{\cdot\cdot}, z_{\cdot}) = \sum_{k=1}^{K} \pi_k\, p(\theta_k|\, y_{\cdot k}, z_{\cdot}) \tag{16}$$

$p(\theta_k|\, y_{\cdot k}, z_{\cdot})$ is the posterior pdf based on model k alone, and π_k is the posterior probability (weight) of model k being correct given the data, and reflects how well model k fits the data. θ_k is the vector of parameters charactering the posterior pdf of model k.

In BMA it is customary to choose the functions $p(\cdot|\cdot)$ from the same family; in this work we selected log-normal functions; so, prior to any analysis, we log-transformed observations and model-predicted concentrations, originally expressed as ng/m^3. The motivation for this choice was based on the consideration that "errors" appeared to be log-normally distributed. In Fig. 2 the histogram of the differences between (log-transformed) model results and observations is shown; as can be seen, some models behave reasonably well, with data approximately log-normally distributed around the observations. Moreover, the choice of log-normal distributions automatically avoids the problem of getting finite probabilities for negative concentration

values. However, there are some models for which deviations from log-normality are pronounced; for example, m08 is extremely diffusive, with a large fraction of results less than observations (resulting in the negative skewness of the empirical pdf). Also, note that all these distributions are not exactly centered on zero, i.e. there is a model-dependent bias. This is particularly relevant for m04, whose results are systematically higher than observations.

In order to avoid that a large number of small values exert a disproportionate influence on BMA results, we discarded all observations with values less than 10^{-2} ng/m^3, close to the threshold (10^{-3} ng/m^3) of the analytical technique; moreover, model values equal to zero were substituted with very small values (in order to avoid "-Inf" warnings due to the application of logarithms).

Markov chain Monte Carlo (McMC) simulation (Gilks et al., 1996) was used to explore the posterior pdf. The basic procedure of Monte Carlo simulation is to draw a large set of samples $\left\{\theta_k^{(l)}\right\}_{l=1}^{L}$, from the target distribution (the posterior pdf in this work). One can then approximate the expectation of any function $f(\theta)$ by the sample mean as follows:

$$\mathrm{E}(f) = \int p(\theta|\cdot)\, f(\theta) d\theta \approx \frac{1}{L}\sum_{l=1}^{L} f(\theta^{(l)})\,, \tag{17}$$

L is the number of samples from the target distribution.

In this work we exploited a Gibbs sampler (Geman and Geman, 1984) to explore the posterior pdf. The Gibbs sampler alternates two major phases: obtaining draws for parameters from the posterior pdf of each model, and obtaining draws for the weights given the model parameters.

In the first phase, we drew a sequence of samples $\left\{(b_k^{(l)}, \sigma_k^{(l)})\right\}_{l=1}^{L}$ for each model k.

The Gibbs sampler was implemented as follows:

for $k = 1 : K$
  Initialize $b_k^{(1)}$ and $\sigma_k^{(1)}$
  for $l = 2 : L$
    draw $b_k^{(l)}$ from $p(b_k|\sigma_k^{(l-1)}, y_{\cdot k}, z_{\cdot})$
    draw $\sigma_k^{(l)}$ from $p(\sigma_k|b_k^{(l)}, y_{\cdot k}, z_{\cdot})$
  end
end

By its construction (Gilks et al., 1996), the Gibbs sampler algorithm guarantees that the chain generates a sequence of values $\left\{(b_k^{(l)}, \sigma_k^{(l)})\right\}_{l=1}^{L}$ which are $p(b_k, \sigma_k|\cdot)$ identically distributed.

Having assumed log-normal distributions and spatio-temporally independent data, the posterior pdf for model k is

$$p(b_k, \sigma_k|\, y_{\cdot k}, z_{\cdot}) \sim \prod_{i=1}^{n} \mathcal{N}\,(y_{ik} - z_i, \sigma_k)\; p(b_k) p(\sigma_k)\,. \tag{18}$$

$p(b_k)$ and $p(\sigma_k)$ are the prior probabilities for the bias and its covariance.

We placed the customary flat prior on the bias and assumed a fairly vague prior for the variance, i.e. we assumed that the prior variance was inverse-gamma distributed with a mean of 9 and variance of 36. In this case Gibbs sampling is easy to apply because it can be demonstrated that the conditional posterior distributions of the Gibbs sampler in the previous algorithm have canonical forms, i.e. a normal distribution for the bias and an inverse-gamma for the variance; for a definition of these functions, and how to draw from them, see Gelman et al. (2003).

In a preliminary test we run three chains in parallel; the Gelman and Rubin test (Gelman and Rubin, 1992) suggested that convergence is reached almost immediately (after a few iterations). We then run a single long (5500 iterations) chain and conservatively discarded the first 500 iterations, well beyond the "burn-in" period suggested by the Gelman and Rubin test. The sample means were estimated from the remaining iterations using Eq. (17), and errors were computed by batching, to account for the correlation in the Markov chain (Roberts, 1996). Table 1 shows the posterior values for the bias and standard deviations, along with their errors (i.e. standard deviations calculated from the McMC sequence).

In the second phase, we sampled the posterior distribution to get a sequence of model weights. If we look at Eq. (16) as the mixture of K competing models, the estimation process can be simplified with the introduction of the binary random variables, ζ_{ik}, with

$$\zeta_{ik} = \begin{cases} 1 \text{ if the } k\text{th model is the 'best' model in predicting} \\ \quad \text{the } i\text{th observation} \\ 0 \text{ otherwise.} \end{cases}$$

If $\theta_k^{(l)}$ indicates a shorthand notation for $(b_k^{(l)}, \sigma_k^{(l)})$, and $\zeta_i = (\zeta_{i1}, \ldots, \zeta_{iK})$, then the selection of the 'best' model in explaining the ith observation can be viewed as the outcome of a multinomial random process (Gelman et al., 2003), i.e.

$$\begin{aligned} p(\zeta_i) &= \mathrm{Multin}(\zeta_i|\, p_{i1}, \ldots, p_{iK}) \\ &= \binom{1}{\zeta_{i1}\zeta_{i2}\cdots\zeta_{iK}} p_{i1}^{\zeta_{i1}} \cdots p_{iK}^{\zeta_{iK}}\,. \end{aligned} \tag{19}$$

The factors p_{ik}s in Eq. (19), are the posterior pdf values of each model, re-normalized so that their sum over index k is equal to 1, i.e.

$$p_{ik} = \frac{p(\theta_k^{(l)}|\, y_{ik}, z_i)}{\sum_{k=1}^{K} p(\theta_k^{(l)}|\, y_{ik}, z_i)}\,, \tag{20}$$

which coincides with the Bayes' factor for the kth model in explaining the ith observation in Eq. (4).

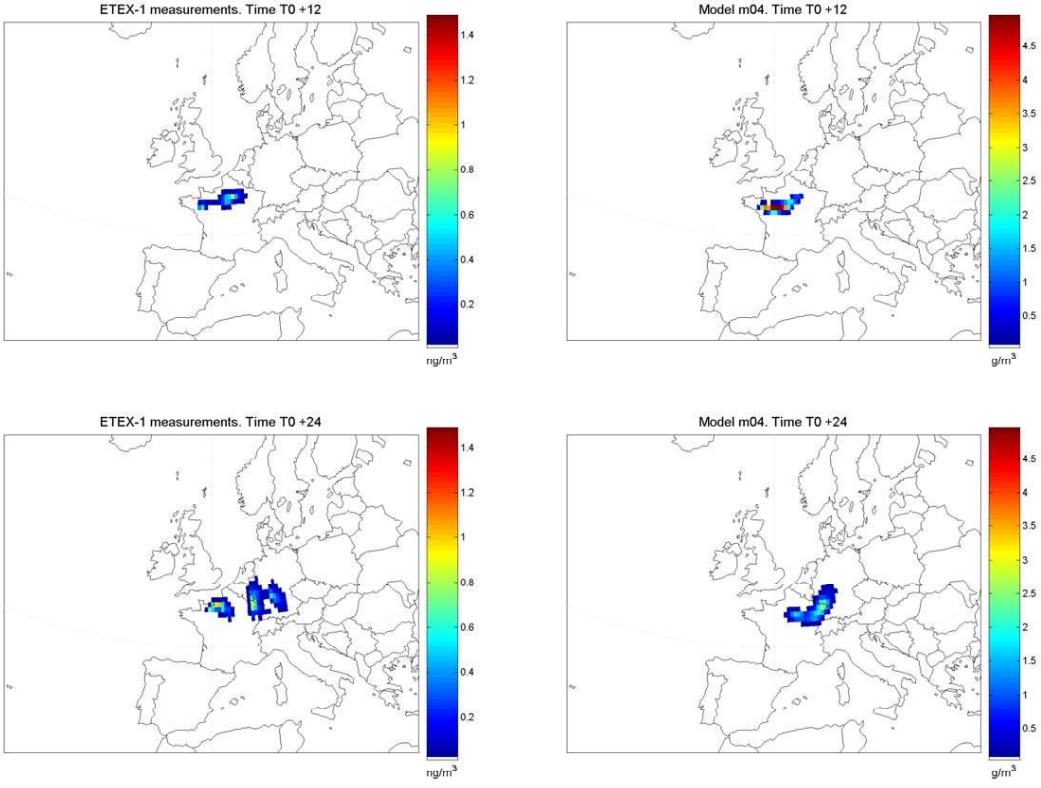

Fig. 3. Comparison between observations (left) and predictions (right) made by m04 at hours T0+12 and T0+24. Note that observed concentrations are expressed as ng/m3, while m04 results as g/m3.

From the properties of a multinomial random process, a draw for ζ_i from (19) is a vector with $K-1$ components equal to zero, and one component (that corresponding to the "best" model) equal to one, i.e. $\sum_{k=1}^{K} \zeta_{ik} = 1$ for any i. Each model has a probability to be selected as the 'best' model equal to p_{ik}, given by Eq. (20).

The selection process was repeated for each observation and iterated for each $\theta_k^{(l)}$ sample, as implemented in the following algorithm:

for $l = 1 : L$
  Set $\theta_k^{(l)} = (b_k^{(l)}, \sigma_k^{(l)})$
  for $i = 1 : N$
    set p_{ik} for any k as in Eq. (20)
    draw $\zeta_i^{(l)}$ from $p(\zeta_i \mid p_{i1}, \ldots, p_{iK})$
  end
end

Table 1 shows the expected values (with their standard deviations) of the fraction of times each model is selected as the "best" model, averaged over all McMC iterations.

Computational costs are negligible; 100 iterations took about 12 s (using Matlab as computing environment installed on a PC with an Intel Centrino Core2 T7200@2GHz CPU and 2048M of main memory), so that the whole estimation process can be accomplished within a few minutes. This makes this data analysis framework suitable for real-time applications, too.

6 Results

Essentially, the objectives of this work consist in the:

- evaluation of the BMA weights, in order to sort the predictive skill of models;
- quantification of the systematic bias of each model;
- estimation of some useful statistical indexes, e.g. APL (Above Percentile Level) or ATL (Above Threshold Level), introduced in Galmarini et al. (2004a, b),
- exploration of similarities and differences between our approach and the "median model",
- quantification of the correlations between models, as a measure of interdependency.

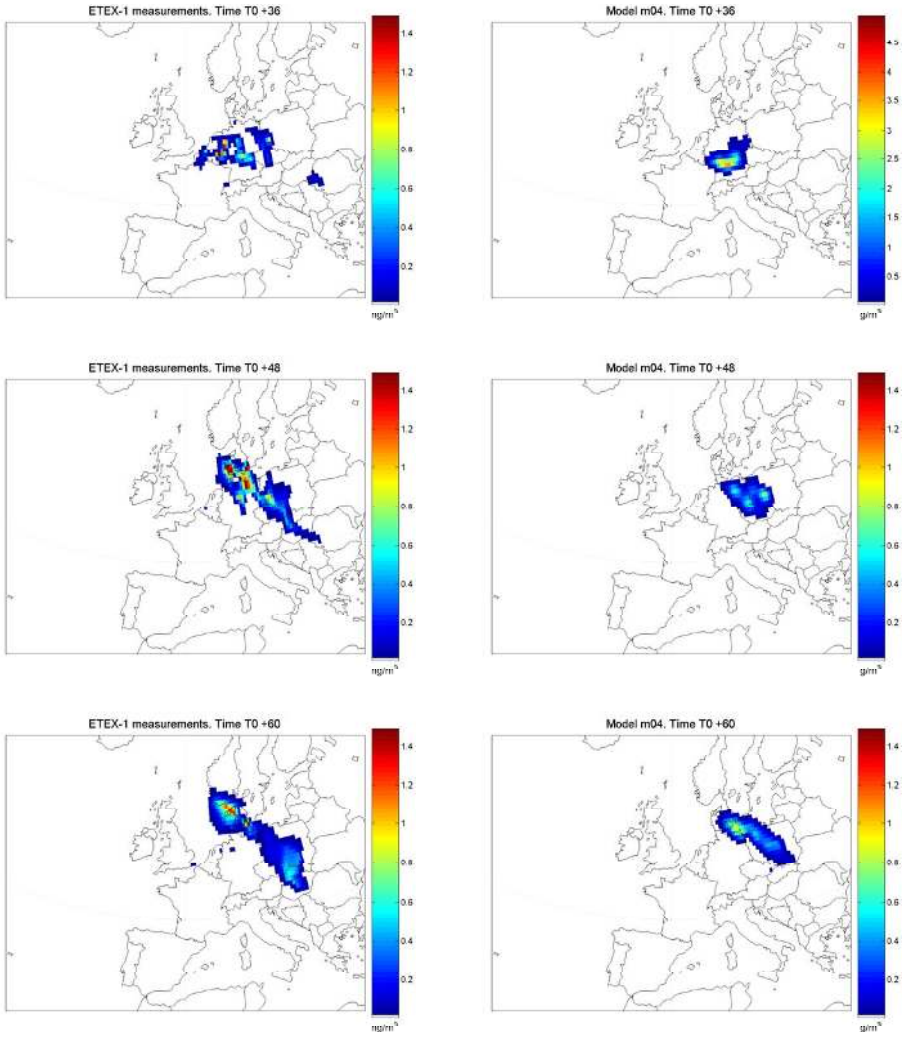

Fig. 3. (Continued.) Comparison between observations (left) and predictions (right) made by m04 at hours T0+36, T0+48 and T0+60. Note that observed concentrations are expressed as ng/m3, while m04 results as g/m3.

We will show that the results of our theoretical framework provides an answer to all these items. The results of the optimization procedure are reported in Table 1.

As can be seen, the a posteriori values of the weights can be clustered in several groups: the majority of model weights are close to the a priori value (1/25=0.04); a second group (models m04 and m08) present a below-average value. Correspondingly, there is a group of three models: m02, m19 and m20 (and to a lesser extent model m12, too), for which the weights are significantly higher than the a priori value.

The bias reported in Table 1 is a measure of how much (on the log-scale) the model predicted values should be shifted so that their mean value coincide with the mean value of observations. It can be noted that model m04 largely overestimates the observations, with a mean bias of about 11.6 on the log scale (remember that an additive bias on the log-scale is equivalent to a multiplicative bias on the linear scale). Also, note that the standard deviation of this bias is considerably larger than those of other models, suggesting that probably something went wrong with this model. As Fig. 3 shows, the physics of dispersion has been qualitatively captured, but, during the first hours after release, the predicted values are extremely high (with a concentration as high as 6 g/m^3 close to the site of release), due to a problem with the source emission strength as pointed out in Galmarini et al. (2004b). The differences between model results and observations tend to disappear during the day after the release, but the highest concentration is predicted over Poland instead of Denmark, as shown by Fig. 3.

Models tend to underestimate observations: the overall mean bias, excluding model m04, is -0.91, corresponding to a shrinking factor of about 0.4; even if m04 is included, the overall mean bias remains negative, i.e. -0.32. It can also be shown that the bias is not uniformly distributed over time:

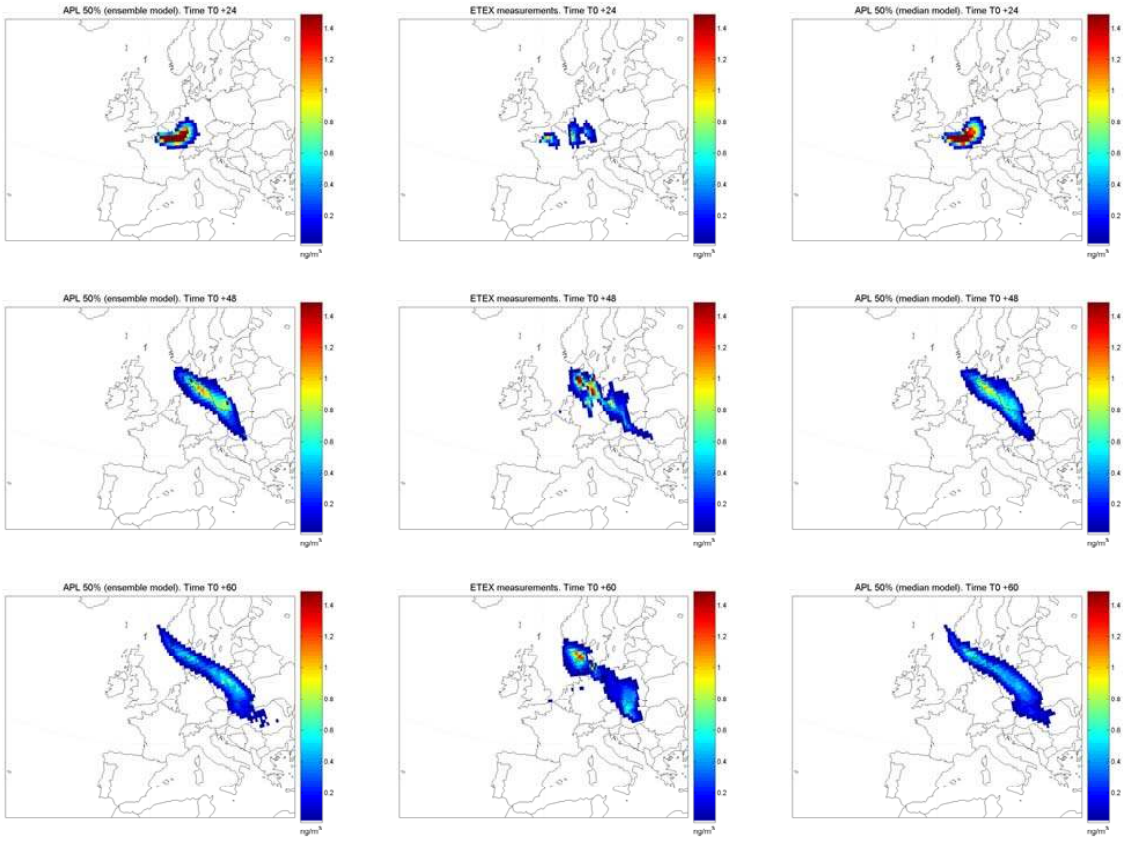

Fig. 4. 50th APL from Eq. (21) (left column), observations (middle column), and 50th APL from the Median Model (right column) adapted from Galmarini et al. (2004b), at T0+24 (uppermost row), T0+48 (middle row) and T0+60 (lowermost row).

models generally tend to overestimate observations close to time of release, and underestimate observations during the day after. We can conjecture that the well-known deficiencies of Eulerian models in correctly representing the subgrid effects, and the extra-diffusion introduced by numerical approaches, play an important role in determining the time tendency of the bias. However, our statistical analysis is not powerful enough to gain an insight into these physical/numerical aspects.

The sampled weights and parameters can be used to calculate some useful statistics, e.g. APL (Above Percentile Level) or ATL (Above Threshold Level).

In Galmarini et al. (2004a), the $\mathrm{APL}_p(x, y, t)$ is defined as the pth percentile from the K models at a specific time t and spatial location (x, y). The $\mathrm{APL}_p(\cdot, \cdot, t)$ can be graphically represented as a two-dimensional surface, e.g. see Fig. 6 in Galmarini et al. (2004a).

The expected value of this index can be straightforwardly estimated from the BMA results, too. For example, the expected APL_{50} is the concentration c' so that

$$\sum_{k=1}^{K} \pi_k \int_{-\infty}^{\log(c')} p(b_k, \sigma_k | \, y_{ik}, z_i) d \log(c) = 0.5 \qquad (21)$$

for any spatio-temporal location denoted by index i.

It is worth noticing that this value coincides with the APL50 index defined in Galmarini et al. (2004a) if a weight equal to 1/K, a bias equal to zero and small standard deviations equal for all models were used in Eq. (21), that is if the a priori values for weights and parameters were used and uncertainties were ignored.

Figure 4 shows the APL_{50} index calculated from Eq. (21), compared with observations and the APL_{50} adapted from Galmarini et al. (2004b). As can be seen, the APL_{50} index from Eq. (21) substantially gives the same results as those from Galmarini et al. (2004b); roughly speaking, this is due to the fact that weights are approximately the same for the majority of models, and there are largely compensating effects between the bias of the different models, so that this ensemble analysis indicates a complementarity between model results.

The evidence for complementarity of model results is also supported by the following result. Figure 5 plots the contribution of each model in determining the BMA median values. For each model, we calculated the following integral:

$$\frac{1}{n} \sum_{i=1}^{n} \int_{-\infty}^{\log(c')} \pi_k \, p(b_k, \sigma_k | \, y_{ik}, z_i) d \log(c) \ ,$$

Table 1. Model weights, bias and standard deviations estimated by the BMA optimization procedure. The corresponding uncertainties (standard deviations) of each parameter are reported within parenthesis. The bias and standard deviations are estimated on the log-scale. Each model is tagged with an integer number shown in the first column.

#	Weight	Bias	Std.Dev.
m01	0.0387 (±0.0041)	−0.15 (±0.04)	2.8 (±0.03)
m02	0.0642 (±0.0055)	0.53 (±0.03)	1.77 (±0.02)
m03	0.0365 (±0.0041)	−0.73 (±0.05)	2.95 (±0.03)
m04	0.0109 (±0.0022)	11.63 (±0.17)	11 (±0.12)
m05	0.0398 (±0.0043)	−2.65 (±0.05)	2.9 (±0.03)
m06	0.0415 (±0.0043)	−2.10 (±0.04)	2.77 (±0.03)
m07	0.0375 (±0.0042)	−0.64 (±0.05)	3.26 (±0.04)
m08	0.0162 (±0.0027)	−2.38 (±0.14)	9.76 (±0.11)
m09	0.0353 (±0.0041)	−1.01 (±0.05)	3.1 (±0.03)
m10	0.0413 (±0.0044)	0.59 (±0.04)	2.76 (±0.03)
m11	0.0359 (±0.0040)	−0.57 (±0.05)	3.01 (±0.03)
m12	0.0503 (±0.0048)	0.37 (±0.04)	2.27 (±0.03)
m13	0.0425 (±0.0044)	−0.61 (±0.04)	2.53 (±0.03)
m14	0.0358 (±0.0040)	−1.50 (±0.05)	3.06 (±0.04)
m15	0.0393 (±0.0043)	−2.45 (±0.05)	2.91 (±0.03)
m16	0.0430 (±0.0045)	−0.52 (±0.04)	2.56 (±0.03)
m17	0.0294 (±0.0037)	−0.59 (±0.07)	4.21 (±0.05)
m18	0.0410 (±0.0043)	−0.11 (±0.04)	2.79 (±0.03)
m19	0.0538 (±0.0049)	0.73 (±0.03)	2.09 (±0.02)
m20	0.0694 (±0.0055)	−2.00 (±0.03)	1.62 (±0.02)
m21	0.0399 (±0.0042)	−2.04 (±0.04)	2.81 (±0.03)
m22	0.0462 (±0.0045)	−0.95 (±0.03)	2.31 (±0.03)
m23	0.0357 (±0.0041)	−1.35 (±0.05)	3.42 (±0.04)
m24	0.0397 (±0.0043)	−1.87 (±0.04)	2.78 (±0.03)
m25	0.0360 (±0.0040)	−3.15 (±0.05)	3.39 (±0.04)

where c' the the median concentration calculated from (21) and n is the number of distinct spatio-temporal locations. Apart from models m04 and m08 which contribute to a lesser extent, and model m20 which contribute to a greater extent, all other models contribute with similar proportions. Therefore, at different times and/or spatial locations, models alternatively contribute to define the BMA median result, without no clear dominant subset. This result reflects very closely that found by Galmarini et al. (2004b).

Also, it should be stressed that the specific values of weights may depend on the selected database, as well as on the assumptions exploited in this work (e.g log-normal deviations of model predictions from observations); differences in the relative performance are expected using different databases and/or implicit assumptions. However, there is no reason to assume that the ETEX-1 database acts as a 'special' case, and we expect that models will continue to behave in a well-balanced manner also using other databases. Results from the ENSEMBLE project (Galmarini et al., 2004a) suggest this is indeed the case.

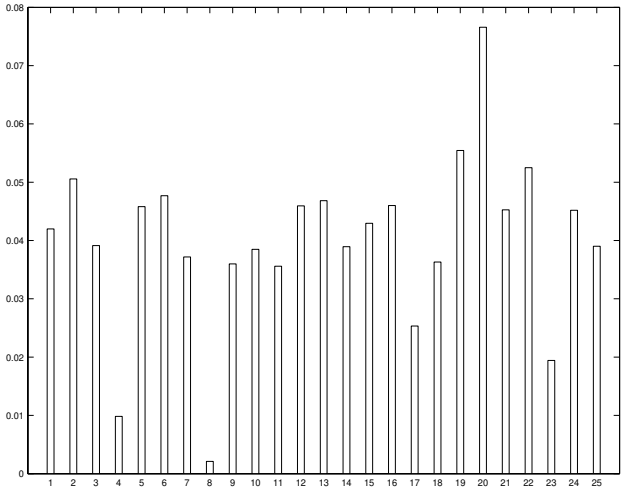

Fig. 5. Contribution of each model to the determination of the BMA 50th percentile. Values are normalized so that their sum is equal to one. The numbers of the x-axis indicate the model tags.

We can move a step forward the analysis of differences and similarities between the BMA approach and the Median Model, by exploring the distribution of the latent variables ζ_{ik}. As can be seen from Eq. (19), the vector of latent variables $\{\zeta_{i1}, \ldots, \zeta_{iK}\}$ is sampled from a multinomial distribution, where each member has a probability to be "extracted" equal to p_{ik}, given by Eq. (20). p_{ik} measures the "distance" of the value predicted by the kth model from the corresponding ith observation, so the kth model is selected with a low probability if it is farther than other models from the ith observation. We can explore the distribution of the ζ_{ik} to search for any systematic structure.

This kind of analysis provides information analogous to the ATL or Space Overlap index. In Galmarini et al. (2004b) the ATL is defined as the surface given by the normalized number of models that, at a given time, predict values above a given threshold c_t, namely

$$\mathrm{ATL}(x, y, t) = \frac{100}{K} \sum_{k=1}^{K} \delta_k$$

$$\text{where} \quad \begin{cases} \delta_k = 1 & \text{if} \quad c_k(x, y, t) \geq c_t \\ \delta_k = 0 & \text{otherwise} \end{cases} \tag{22}$$

An analogous information can be deduced from the ζ_{ik} variables, too. We define the PBS (Probability of Being Selected) index as follow

$$\mathrm{PBS}_{ik} = 1 - \frac{1}{L} \sum_{l=1}^{L} \zeta_{ik}^{(l)}, \tag{23}$$

where L is the total number of McMC iterations. This index is close to 0 if model k performs much better than the other models in explaining the ith observation, i.e if the mean value of ζ_{ik} tends to 1; conversely, it tends to 1 if model k is one

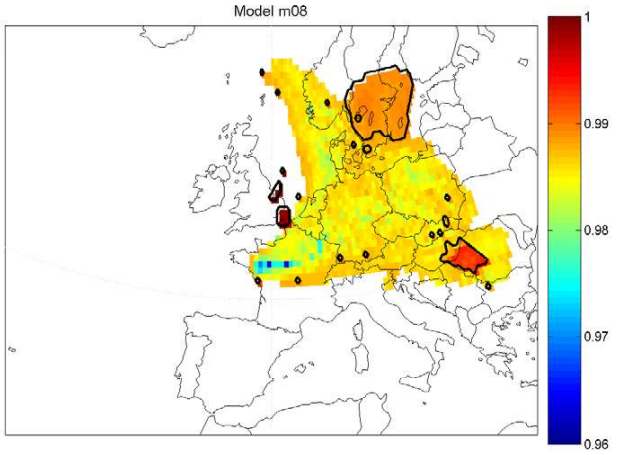

Fig. 6. The PBS index for m08. The areas for which PBS≥0.985 have been contoured with black solid lines.

of the worst model in explaining the ith observation. Figure 6 shows the PBS index for m08. A PBS average value of about $0.98(\approx 1.0-0.016)$ can be deduced for this model (see Table 1).

In Fig. 6 the areas for which PBS $\geq$0.985 have been contoured with black lines; the result is a "leopardized" structure. The leopard-like spots are due to the fact that we have not introduced any physical information is our sampling strategy: obviously model's results are spatio-temporally correlated, so we could expect a smoothly varying surface of the PBS index, but in Eq. (18) we implicitly assumed that model results are independently distributed in space and time. Notwithstanding this lack of physical coherence, there are some remarkable structures: the "bump" protruding over the Scandinavian region and that over Eastern Romania. It can be shown that these spots are due to high model concentrations which are not represented, neither by observations nor by the majority of other model results. This finding has already been outlined by Galmarini et al. (2004b) using the ATL index. They showed that the protrusion over the Scandinavian area corresponds to ATL $\approx$1, i.e. a characteristic showed only by m08 (see Figs. 3 and 4 in Galmarini et al., 2004b).

As a final example of the potentialities of this approach, we analyze the information that can be gained from the eigendecomposition of the covariance matrix. Each model in (19) can be independently selected from all others; however, models cannot be completely independent since they simulate the same phenomenon, described by well defined physical laws. As explained in Sect. 4, a viable approximation to quantify dependences among models is correlation. To this aim, we changed model Eq. (18) to

$$p(b,\,\Sigma|\,y_{..},\,z_{.})\sim\prod_{i=1}^{n}\mathcal{N}\,(y_{i\cdot}-z_{i},\,\Sigma)\;p(b)p(\Sigma)\,. \qquad (24)$$

Table 2. Components of some selected eigenvectors of the estimated covariance matrix. Values greater than 0.35 have been reported as bold. See the text for more details.

#	Eig. 1	Eig. 2	Eig. 22	Eig. 23	Eig. 24	Eig. 25
m01	−0.090	0.013	−0.001	0.129	−0.025	−0.011
m09	−0.050	−0.032	0.035	−0.003	0.032	0.037
m18	−0.067	−0.035	−0.024	-0.010	0.007	0.010
m02	−0.040	−0.037	0.070	−0.150	**0.834**	**−0.372**
m23	−0.047	−0.064	−0.000	0.040	−0.028	−0.013
m13	−0.065	−0.014	0.110	0.160	0.016	−0.040
m14	−0.090	0.011	0.018	−0.062	0.034	−0.017
m03	−0.032	−0.036	0.040	−0.022	−0.009	0.030
m17	−0.107	−0.004	−0.026	−0.025	−0.005	0.028
m04	**−0.818**	**0.532**	−0.003	−0.010	0.002	0.010
m05	−0.001	−0.041	−0.044	−0.021	−0.011	−0.020
m10	−0.092	−0.014	0.186	−0.023	−0.012	−0.001
m12	−0.052	−0.040	0.124	**0.620**	−0.198	0.014
m06	−0.021	−0.018	0.148	0.019	−0.000	−0.024
m19	−0.050	−0.038	−0.028	**−0.671**	**−0.454**	−0.261
m11	−0.065	−0.027	0.031	−0.013	0.019	0.020
m15	−0.066	0.013	0.029	−0.044	0.000	0.003
m07	−0.064	−0.014	−0.016	0.027	−0.041	−0.008
m20	−0.018	−0.019	−0.004	−0.249	0.217	**0.884**
m16	−0.076	−0.003	0.186	0.035	−0.007	−0.021
m08	**−0.498**	**−0.836**	−0.004	0.003	−0.009	0.014
m21	−0.036	−0.018	0.083	−0.099	0.000	0.031
m22	−0.054	−0.016	**−0.924**	0.107	0.052	−0.028
m24	−0.037	−0.017	0.045	0.064	−0.028	−0.030
m25	−0.035	0.010	−0.064	0.026	−0.045	−0.039

where now Σ is the $K-$dimensional matrix of covariances between models. $p(\Sigma)$ is the prior pdf for Σ, for which we chose a non-informative inv-Wishart distribution.

The analysis of the expected values of the covariance matrix says what models show correlated deviations from the observations.

As shown in Eq. (15) and Fig. 1, the eigenvectors of the covariance matrix correspond to the directions of independent components if data are normally (or log-normally) distributed. The magnitudes of the components of each eigenvector immediately say to what extent each model contributes to that independent component.

In Table 2 we report the eigenvectors corresponding to the two largest eigenvalues (Eig. 1 and Eig. 2) and to the three smallest eigenvalues (Eig. 23, Eig. 24 and Eig. 25). As can be seen, the first two eigenvectors are dominated by the components corresponding to m04 and m08, and all other models have negligible projections on these two vectors.

The first two eigenvalues (data not shown) explain about 61% of the total variance; of course this is not surprising, since, as can be seen from table 1, m04 and m08 are associated with the largest variances. This means that, not only m04 and m08 are associated with a great bias, but they also significantly co-vary (i.e. the spatio-temporal pattern of their bias is similar) and are not significantly correlated with all other models, because their projection over the successive eigenvectors is negligible.

It is worth noticing that, while models m04 and m08 are positively correlated along the direction of the first eigenvector (components with the same sign), they are negatively correlated along the direction of the second eigenvector. This is due to the fact that model m08 is extremely diffusive, so that it predicts positive concentrations even where model m04 shows zero values (remember that model m04 predicts extremely high values on the mean); the first set of data is clustered along the first eigenvector, and the second set along the second eigenvector.

There are also significant correlations between models m02 and m19 and models m02 and m20; Eig. 23 also shows that model m19 is significantly correlated with model m12. Remember that these models show the highest BMA weights (see Table 1). The data from all other models are projected more uniformly among the remaining eigenvectors.

We conjecture that models m02, m19 and m20 perform better than the others because their data share a similar spatio-temporal pattern, and this similarity is highlighted by the significant correlations between their bias.

In a model selection perspective, the analysis of the covariance matrix can be used to pick those models showing independent features. If a model would be sacrificed, it is better to discard a model with a low BMA weight and well correlated with other models.

7 Conclusions and final considerations

The results presented in the previous section highlight the advantages of the BMA framework:

1. the weights provide the quantitative basis to judge if there is an "outlier model", but, instead of disregarding its values, they are bias-corrected, weighted and included in the final analysis satisfying an optimality criterion, i.e. so that the posterior probability is maximized;

2. the McMC approach provides the way to quantify the uncertainties of each estimated parameter, so that any decision making or regulatory-purpose activity, can be supported by an adequate uncertainty analysis;

3. a deeper analysis, based on the distribution of unobserved indicators, ζ_{ik}, allows to detect the outliers among the model-predicted values, i.e. a very low mean value of ζ_{ik} indicates that the ith observation is very different from the kth model-predicted value. This analysis can be projected onto the physical space/time, thus playing a role similar to several other statistical indexes, e.g. the Agreement in Threshold Level or Space Overlap, originally introduced in Galmarini et al. (2004a; 2004b);

4. the analysis of the covariance matrix can be used to inspect the similarities and/or differences between model results. We can look at the values projected onto the eigenvectors of the covariance matrix as "orthogonal" data, i.e. data forecast by independent models, whose variations cannot be explained by the other components. In a model selection perspective, the number of independent model can be selected as those associated with the most "interesting" (uncorrelated) directions.

As outlined in Galmarini et al. (2004b), the "Median Model" results provide an estimate that is superior to any single deterministic model simulation, with obvious benefits for regulatory-purpose applications or for the support to decision making. We can look at our ensemble analysis as the a posteriori justification of the Median Model results.

Edited by: R. Vautard

References

Berliner, L. M.: Physical-statistical modeling in geophysics, J. Geophys. Res., 108, 8776, doi:10.1029/2002JD002865, 2003.

Clyde, M. A.: Bayesian model averaging and model search strategies (with Discussion), in: Bayesian Statistics 6, edited by: Bernardo, J. M. et al., 157–185, Oxford University Press, Oxford, 1999.

Clyde, M. A. and George, E. I.: Flexible empirical Bayes estimation for wavelets, J. R. Stat. Soc. Ser. A–G, 62, 681–698, 2000.

Coelho, C. A. S., Pezzulli, S., Balmaseda, M., Doblas-Reyes, F. J., and Stephenson, D. B.: Forecast Calibration and Combination: A Simple Bayesian Approach for ENSO, J. Climate, 17, 1504–1516, 2004.

Cover, T. M. and Thomas, J. A.: Elements of Information Theory, Wiley, 1991.

Delle Monache, L. and Stull, R. B.: An ensemble air-quality forecast over western Europe during an ozone episode, Atmos. Environ., 37, 3469–3474, 2003.

Delle Monache, L., Deng, X., Zhou, Y., and Stull, R.: Ozone ensemble forecasts: 1. A new ensemble design, J. Geophys. Res., 111, D05307, doi:10.1029/2005JD006310, 2006a.

Delle Monache, L., Nipen, T., Deng, X., Zhou, Y., and Stull, R. B.: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction, J. Geophys. Res., 111, D05308, doi:10.1029/2005JD006311, 2006b

Delle Monache, L., Hacker, J. P., Zhou, Y., Deng, X., and Stull, R. B.: Probabilistic aspects of meteorological and ozone regional ensemble forecasts, J. Geophys. Res., 111, D24307, doi:10.1029/2005JD006917, 2006c.

Dijkstra, T. K.: On Model Uncertainty and its Statistical Implications, Springer Verlag, Berlin, 1988.

Fritch, J. M., Hilliker, J., Ross, J., and Vislocky, R. L.: Model consensus, Weather Forecast, 15, 571–582, 2000.

Galmarini, S., Bianconi, R., Bellasio, R., and Graziani, G.: Forecasting consequences of accidental releases from ensemble dispersion modelling, J. Environ. Radioactiv., 57, 203–219, 2001.

Galmarini, S., Bianconi, R., Klug, W., Mikkelsen, T., Addis, R., Andronopoulos, S., Astrup, P., Baklanov, A., Bartniki, J., Bartzis, J. C., Bellasio, R., Bompay, F., Buckley, R., Bouzom, M., Champion, H., D'Amours, R., Davakis, E., Eleveld, H., Geertsema, G.

T., Glaab, H., Kollax, M., Ilvonen, M., Manning, A., Pechinger, U., Persson, C., Polreich, E., Potemski, S., Prodanova, M., Saltbones, J., Slaper, H.,Sofief, M. A., Syrakov, D., Sorensen, J. H., Van der Auwera, L., Valkama, I., and Zelazny, R.: Ensemble dispersion forecasting–Part I: concept, approach and indicators, Atmos. Environ., 38, 4607–4617, 2004a.

Galmarini, S., Bianconi, R., Addis, R., Andronopoulos, S., Astrup, P., Bartzis, J. C., Bellasio, R., Buckley, R.,Champion, H., Chino, M., D'Amours, R., Davakis, E., Eleveld, H., Glaab, H., Manning, A., Mikkelsen, T., Pechinger, U., Polreich, E., Prodanova, M., Slaper, H., Syrakov, D., Terada, H., and Van der Auwera, L.: Ensemble dispersion forecasting–Part II: application and evaluation, Atmos. Environ., 38, 4619–4632, 2004b.

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci., 7, 457–472, 1992.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B.: Bayesian Data Analysis, Chapman and Hall/CRC, Boca Raton, Florida, 2003.

Geman, S. and Geman, D.: Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, Transactions on Pattern Analysis and Machine Intelligence, 6(6), 721–741, 1984.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J.: Markov Chain Monte Carlo in Practice, Chapman and Hall/CRC, Boca Raton, Florida, 1996.

Girardi, F., Graziani, G., van Veltzen, D., Galmarini, S., Mosca, S., Bianconi, R., Bellasio, R., and Klug, W.: The ETEX project. EUR Report 181-43 EN. Office for official publications of the European Communities, Luxembourg, 108pp., 1998.

Grimit, E. P. and Mass, C. F.: Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest, Weather Forecast, 17, 192–205, http://isis.apl.washington.edu/bma/index.jsp, 2002.

Hyvärinen, A. and Oja, E.: Independent Component Analysis: Algorithms and Applications, Neural Networks, 13, 411–430, 2000.

Hou, D., Kalnay, E., and Droegemeier, K. K.: Objective verification of the SAMEX'98 ensemble forecast, Mon. Weather Rev., 129, 73–91, 2001.

Jeffreys, H.: Theory of Probability, 3rd Edition, Oxford University Press, 1961.

Krishnamurti, T. N., Kishtawal, C. M., Zhang, Z., LaRow, T., Bachiochi, D., Williford, E., Gadgil, S., and Surendran, S.: Multimodel ensemble forecasts for weather and seasonal climate. Mon. Weather Rev., 116, 907–920, 2000.

Mallet, V. and Sportisse, B: Ensemble-based air quality forecasts: A multimodel approach applied to ozone, J. Geophys. Res., 111, D18302, doi:10.1029/2005JD006675, 2006.

Molteni, F., Buizza, R., Palmer, T. N., and Petroliagis, T.: The ECMWF ensemble system: Methodology and validation, Q. J. Roy. Meteor. Soc., 122, 73–119, 1996.

Pagowski, M., Grell, G. A., McLeen, S. A., et al., 2005: A simple method to improve ensemble-based ozone forecasts, Geophys. Res. Lett., 32, L07814, doi:10.1029/2004GL022305

Pagowski, M., Grell, G. A., Devenyi, D., Peckham, S. E., McKeen, S. A., Gong, W., Delle Monache, L., McHenry, J. N., McQueen, J., and Lee, P.: Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts, Atmos. Environ., 40, 3240–3250, 2006a

Pagowski, M. and Grell, G. A.: Ensemble-based ozone forecasts: Skill and economic value, J. Geophys. Res., 111, D23S30, doi:10.1029/2006JD007124, 2006b.

Papoulis, A.: Probability, Random Variables, and Stochastic Processes, Mc-Graw-Hill, 1991.

Raftery, A. E., Madigan, D. and Hoeting, J. A.: Model selection and accounting for model uncertainty in linear regression models, J. Am. Stat. Assoc., 92, 179–191, 1997.

Raftery, A. E.and Zheng, Y.: Long-run performance of Bayesian model averaging, J. Am. Stat. Assoc., 98, 931–938, 2003.

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M.: Using Bayesian Model Averaging to Calibrate Forecast Ensembles, Mon. Weather Rev., 133, 1155–1174, 2005.

Roberts, W. R.: Markov chain concepts related to sampling algorithms, in: Markov Chain Monte Carlo in Practice, edited by: Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., Chapman and Hall, 45–57, 1996.

Toth, Z. and Kalnay, E.: Ensemble forecasting at the NMC: The generation of perturbations, B. Am. Meteorol. Soc., 74, 2317–2330, 1993.

Viallefont, V., Raftery, A. E., and Richardson, S.: Variable selection and Bayesian model averaging in case-control studies, Statistics in Medicine, 20, 3215–3230, 2001.

Whitaker, J. S. and Loughe, A. F.: The relationship between Ensemble Spread and Ensemble Mean Skill, Mon. Weather Rev., 126, 3292–3302, 1998.

Zhang, F., Bei, N., Nielsen-Gammon, J. W., Li, G., Zhang, R., Stuart, A., and Aksoy, A.: Impacts of meteorological uncertainties on ozone pollution predictability estimated through meteorological and photochemical ensemble forecasts, J. Geophys. Res., 112, D04304, doi:10.1029/2006JD007429, 2007.