H14-130

BIAS CORRECTION AND ENSEMBLE TECHNIQUES TO IMPROVE AIR QUALITY ASSESSMENT: FOCUS ON O_3 AND PM OVER PORTUGAL

A. Monteiro[1], I. Ribeiro[1]; O. Tchepel[1], A. Carvalho[1], E. Sá[1], J. Ferreira[1], S. Galmarini[2], A.I. Miranda[1], C. Borrego[1]

[1]CESAM & Department of Environment and Planning, University of Aveiro, Aveiro, Portugal
[2]IES/REM, Joint Research Centre, European Commission, TP 441 21020 Ispra, Italy

Abstract: After removing the bias from each model, different ensemble techniques were applied and compared. Besides the median (MED), several weighted ensemble approaches were tested and intercompared: static (SLR) and dynamic (DLR) multiple linear regressions (using less-square optimization method) and the Bayesian Model Averaging (BMA) methodology. The goal of the comparison is to estimate to what extent the ensemble analysis is an improvement with respect to the single model unbiased results. The obtained results revealed that no one of the 4 tested ensembles clearly outperforms the others on the basis of statistical parameters and probabilistic analysis (reliability and resolution properties). Nevertheless, statistical results have shown that the application of the weights slightly improves ensemble performance when compared to those obtained from the median ensemble. The same statistical analysis, together with the probabilistic measures, demonstrates that the SLR and BMA methods are the best performers among the assessed methodologies.

Key words: *air quality modelling; bias-correction;ensemble analysis; static and dynamic regression; BMA approach*

INTRODUCTION

Although widely used in operational weather prediction (e.g. Stensrud and Yussouf, 2003), ensemble forecasting of air quality has only begun to be investigated in recent years, mainly focusing on ozone (O_3) (e.g. Delle Monache et al., 2006a,b; Djalalaova et al., 2010). As shown by several authors for specific verification periods and for air quality purposes, the ensemble average or median of several independent simulations is usually closer to observations than a single simulation (e.g. Galmarini et al., 2004a,b; McKeen et al., 2005; Wilczak et al., 2006). Nevertheless, an ensemble can only give significant improvements if participating models have complementary strengths and weaknesses. This is difficult to estimate a priori and makes the use of large ensembles of models the current trend. However Potempski and Galmarini (2009) argued against the efficacy of this practice which may lead to a multiplication of overlapping model contributions to the process description, thus not leading to a real improvement of the ensemble result. Recent studies (e.g. Djalalova et al., 2010) examined the benefit of correcting the modelling results using the biases calculated by different ways. The objective is that the ensemble represents a more correct collective model perspective in which bias result from a collective misrepresentation of the system. For this purpose we have used five different regional air quality models that were applied over mainland Portugal at 5x5 km^2 spatial resolution, for July 2006, and their results were bias corrected. These models include (1) the Comprehensive Atmospheric Model with extensions (CAMx) (ENVIRON, 2008); (2) the CHIMERE model (Schmidt et al., 2001); (3) the European Air Pollution and Dispersion – Inverse Model (EURAD-IM) (Elbern et al., 2007); (4) the LOTOS-EUROS (Schaap et al., 2008) and (5) The Air Pollution Model (TAPM) (Hurley et al., 2005). These models were applied using different meteorological drivers, parameterizations, boundary conditions and also chemical mechanisms, but the same pollutant emissions inventory. The analysis focused not only on O_3, as the referred studies, but also on PM, both critical pollutants in terms of the limit values exceedances.

THE MODELLING APPROACH

The selected models include CAMx, CHIMERE, EURAD-IM, LOTOS-EUROS and TAPM. All models are regional-scale models designed for short-term and long-term simulations of oxidants and aerosol formation, and have been applied over Portugal for several times and purposes. The models have different degrees of complexity. EURAD-IM and TAPM describe the whole tropospheric column with several vertical layers, while CHIMERE and LOTOS-EUROS describe only the lower troposphere, up to above the boundary layer. LOTOS-EUROS has varying vertical layers, which follow the boundary layer diurnal evolution. CAMx vertical resolution is based and depends on the MM5 vertical layers structure. Boundary conditions are either based on observations (LOTOS-EUROS), model simulations (CAMx, CHIMERE, TAPM) or both (EURAD-IM). Driving meteorology is taken directly from the MM5 (Dudhia, 1993) meteorological model for CAMx, CHIMERE and EURAD-IM or from an optimal interpolation analysis based on observations for LOTOS-EUROS and TAPM. A summary of the modelling systems key features and additional descriptions can be consulted on the online Model Documentation System (http://pandora.meng.auth.gr/mds/mds.php). Any multi-model ensemble approach relies on model 'diversity' or in other words on the fact that different models produce with more or less emphasis specific and different aspects of the physical process they want to model (Potempski and Galmarini, 2009). We therefore used the five models in their original set up in terms of input data, numerical grid resolution, parameterizations and boundary conditions hoping with this to create an intrinsic diversity in the model results, though without having any a priori evidence of that. The anthropogenic emissions are defined on a common basis, using the national emission inventory (INERPA) spatially disaggregated (Monteiro et al., 2007) and the same horizontal resolution (5x5 km^2). O_3 and PM observed data were collected at the Portuguese air quality monitoring network, with a total of 22 background stations (www.qualar.org).

THE ENSEMBLE APPROACHES

Median ensemble (MED)

The median approach was selected since the distribution of the models results is unknown a-priori, and also based on Monteiro et al. (2011) results that confirmed the superior skill of the median ensemble comparing to the mean value. By definition, the median is less sensitive to extreme scores and it is a better measure for highly skewed distributions. The MED

ensemble thus filters extreme results and when performed at each point in space and time, reduces the deterministic character of the single realization (Galmarini et al., 2004). Given the present population of the models results, the use of the mean value lead to a large overestimation of the concentration levels and resulted less appropriate (Monteiro et al., 2011). An argument that can be raised against the use of the MED ensemble may relate to the fact that all models are equally weighted and the selection of the model that defines the ensemble, based on the ensemble distribution, can be erroneous.

Static linear regression (SLR)

A different approach to derive an ensemble (in opposition to the equally weighted median ensemble) is to use linear regression techniques to find weight coefficients for the models such that the sum of the weighted models has a minimum bias. Linear regression method has been previously used by Krishnamurti et al. (1999) to improve precipitation forecasts using a so-called superensemble which included four diverse global weather forecasting models, and after was applied for air quality modelling purposes (e.g. Pagowski et al., 2005; Djalalova et al., 2010). A system of equations is formulated as a linear combination of the simulations of ensemble members multiplied by unknown weights to be equal to the measured pollutant concentration, as shown in equation (1):

$$\begin{bmatrix} m_{11} & m_{12} & \dots & m_{1J} \\ m_{21} & m_{22} & \dots & m_{2J} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ m_{I1} & m_{I2} & \dots & m_{IJ} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \\ w_J \end{bmatrix} = \begin{bmatrix} o_1 \\ o_1 \\ \dots \\ \dots \\ o_I \end{bmatrix} \quad (1)$$

where I is the total number of observations, J the number of ensemble members, m_{ij} are ensemble member simulations, w_j the unknown weights, and o_i the observations. In the current case I is 744 (24x31), and J = 5. Since I > J the above system is over-determined and a solution to a least-square problem have to be sought. In the current approach the models weight will be estimated for each single station and not applied/assumed the same for the entire modelling domain. To calculate weights we use a technique based on the minimization of error (defined as the difference between ensemble values and measurements in the least-square sense), which allows to find the least-square solution even when severe degeneracies in matrix M occur. Matlab software with adequate libraries was used to apply this technique. In order to evaluate whether the simulation period (July month) would be sufficient to test this ensemble technique, we vary the length of the training period to determine the weights for the ensemble members from 1 to 31 days. For a specific training period, weights are calculated by solving the equation (1) written for all the events during this period. Since the model results come from a previous bias-correction technique application (Monteiro et al., 2011), only the last 26 days are available for this purpose. The dependence of the models weight averaged over this 26-day evaluation period on different lengths of the training period is shown in Figure 1.

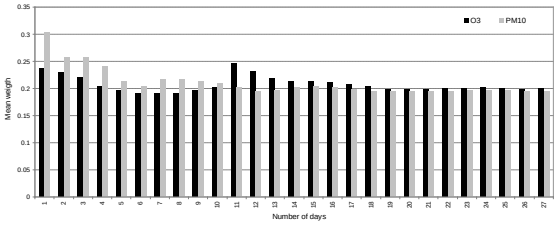


Figure 1. Relationship between the training period and the models average weight, for O_3 (black) and PM10 (grey).

The variability of weights decreases with the length of the training period. It is clear that a long training period would result in weights more appropriate for a long-term mean, while the weights calculated based on recent several days would be more responsive to the daily variability of the pollutants concentrations. It can be seen that the magnitude of weights vary significantly during the first 10 days and change little after 15 days of training. This is in agreement with other previous studies (e.g. Pagowski et al., 2005) and proves that the availability of data for only a single month is sufficient and allows testing this ensemble approach. Nevertheless, for the present exercise, we choose to use the total number of days (26 days).

Dynamic linear regression (DLR)

In contrast to static linear models, which assume that model performance does not change in time, dynamic linear models allow for temporal evolution of the characteristics of these processes. Following Pagowski et al. (2006), the dynamic linear model is calculated, similar to the SLR, using the hourly model and observed ozone values:

$$M_{i1}*w_1 + M_{i2}*w_2 + \dots + M_{i5}*w_5 = O_i \quad (2)$$

where M_{i1},., M_{i5} and O_i are the values of the 5 models and the observation for a single monitoring site; and w_1,., w_5 are the respective unknown weight coefficients. But in this case, in opposition to the SLR approach, a training period is moving in time and is similar to the dynamic bias-correction approach applied in Monteiro et al. (2011). Different length of training periods was tested in order to identify better performance. Similarly to the bias-correction exercise, a 4 and 7 days period (related to synoptic patterns lifetime) were evaluated, together with 1 day period. The over-determined matrix system is also solved using a least-squares minimization procedure which provides the optimal coefficients (w_1,., w_5). This procedure is repeated for all days in the considered time period. In Figure 2, the different training periods are compared using standard statistics to assess the ensemble performance, considering the average over all the monitoring sites.

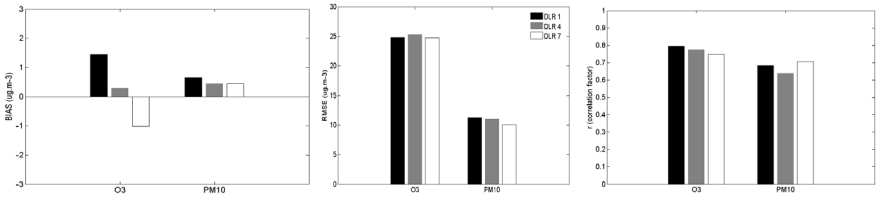


Figure 2. Validation of the DLR ensemble technique for different training periods (considering the averaged statistics BIAS, RMSE and correlation factor over all the monitoring stations).

The results show that there are no significant differences between the three different training periods, all presenting small errors (BIAS < ± 2 μg.m^{-3} and RMSE < 25 μg.m^{-3}) and high correlation measures (r>0.6), for both pollutants. The 7 days training period has the best overall performance, taking into account the three different averaged statistics and both pollutants. This training period was also selected because requires data only from the 7 previous days, and therefore this technique may still be easily applied to an operational forecast. The model weights were calculated for each observation site, allowing spatial variability of weights to be obtained. Since some of the tested ensemble methods require model weights averaged in space, this DLR7 technique will be applied using the time dependent weights only (the same set of w_i coefficients are used for all sites). Figure 3 compares the ensemble errors for these two approaches (function of time and space, as in Figure 3, and function of time only). As already expected, the DLR ensemble performance reduces when model weights are averaged in space and only vary in time. Nevertheless, there are no substantial disparities, and some exceptions are also observed.

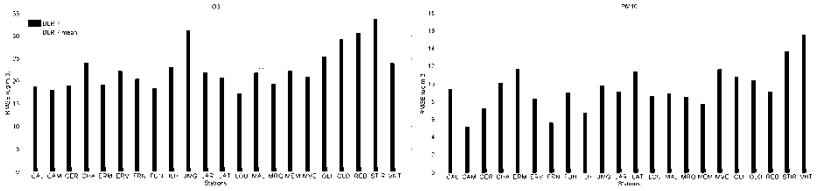


Figure 3. RMSE of the DLR7 ensemble using model weights as a function of time and space (black bars) and averaged in space (white bars).

Bayesian Model Averaging (BMA)

The BMA methodology (Riccio et al., 2007) is a statistical procedure that creates the optimal combination of the results obtained from different models by weighting individual model simulations on the basis of probabilistic measures. The weights are assigned so that the models that better predict the concentration values in a portion of the modelling domain and at a specific instant in time get the highest values and, therefore, have a great contribution to the definition of the average. More specifically, the BMA scheme describes the posterior probability density function (pdf) as a weighted average of probability distributions of individual models. The BMA technique was already applied in different contexts and presented in several publications (e.g. Riccio et al., 2007). The pdf of the BMA approach is given by (3)

$$p(x|O) = \sum_{k=1}^{j} p_k(x|M_k, O)\, w_k \qquad (3)$$

where w_k is the posterior probability of the model M_k being the best forecast in the ensemble, and$p_k$is the posterior probability that x occurs for a given model prediction M_k and measurement data O. The comparison of probability distributions of the modelling results against the observations (not shown) shows that all the models have similar performance for O_3 and the distributions of modelling results and measurements are linearly related. More complex behaviour is demonstrated for PM10. A difficulty of the models to correctly predict high concentrations is an important factor considered in BMA approach to attribute reliability, and therefore the weights, for each model. The skill levels of the ensembles are analysed using different approaches. This intercomparison was carried out using the analysis of the averaged time series (Figure 4) and classical statistical indicators (Borrego et al., 2008) (Taylor diagram; Figure 5). Appropriate probabilistic measures for ensemble evaluation, like Talagrand diagrams, are also presented (Figure 6). Figure 5 depicts the time series of observations and each ensemble results, averaged over all the monitoring sites, for each pollutant and during the simulation period.

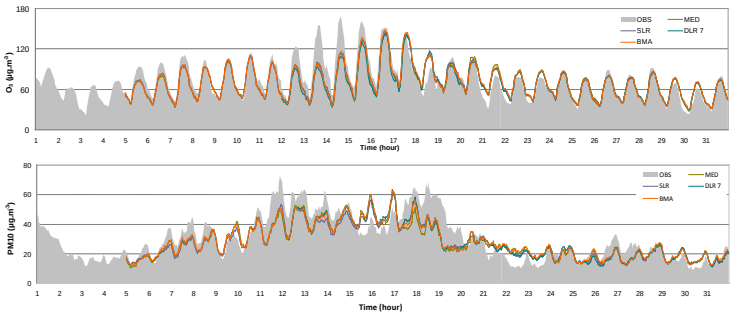
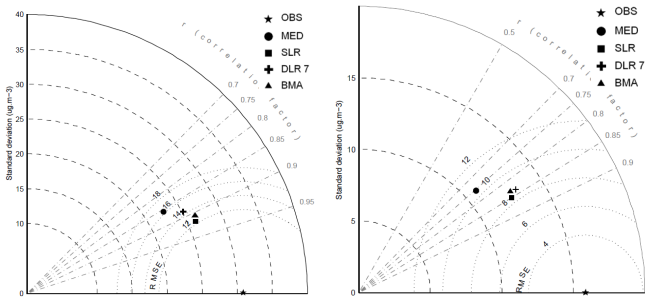


Figure 4. Hourly time series averaged over all monitoring stations, for observed and each ensemble, for O_3 and PM10.

The similarity of the four ensemble techniques is highlighted by the averaged time series that shows a similar behaviour within the ensembles for the background and peak concentrations and for both pollutants. It should be noticed that none of the ensembles was able to predict the magnitude of the peaks occurred during the 12th-16th July episode days, besides an overall good agreement between the observations and simulated values. To represent the global skill of the models and of their ensemble average, Figure 6 shows, in a single diagram, the ‘‘Taylor plots’’ for all ensemble methods and for both pollutants. The Taylor diagram is a powerful tool that summarizes standard deviation, RMSE and R in a single point on a two-dimensional plot. Together, these statistics provide a quick summary of the degree of pattern correspondence among the simulated values and the real data. In this particular case, the diagram is particularly useful in assessing the relative merits of the applied ensemble techniques. The diagram will quantify how closely the ensemble resembles the observed field (OBS). In Figure 5 five points are plotted on a polar style graph: the OBS represents the observed data and the others represent the ensemble techniques. The radial distances from the origin to the points are proportional to the pattern standard deviations, and the azimuthal positions give the correlation coefficient between the ensemble and the OBS field. The radial lines measure the distance from the OBS point and indicate the RMSE. The point representing the OBS field is plotted along the abscissa. In this case, the OBS field has a standard deviation of 30.8 μg.m^{-3} and 15.9 μg.m^{-3} for O_3 and PM10, respectively.



(a) O_3 (b) PM10

Figure 5. Taylor diagram that summarizes the overall performance of the different ensemble techniques.

From Figure 5, it should be highlighted the higher skill (correlation coefficients higher than 0.7) of the weighted ensembles in comparison to the median model. Especially significant for the median model is the increase in RMSE, while the decline in correlation is more modest. The application of the dynamic regression based on daily model outputs does not improve the ensemble, namely for O_3. It can be stated that, for both pollutants, SLR and BMA methods present the best overall performances. Comparing the poorer ensemble (the median) and the best performing ensemble (the SLR and BMA methods) we find a 22% improvement for RMSE and 7% for the correlation coefficient of O_3. For PM10, these improvements are also relevant, 18% for RMSE and 11% for the correlation coefficient. To complement this ensemble intercomparison, the capability of each ensemble to produce valuable estimates of the observed frequency associated with different simulation frequencies was determined (Delle Monache et al., 2006a). This property can be verified using reliability diagrams, also known as rank histograms which display the frequency of occurrence as a function of predicted probability. First, the ensemble methods are ranked for each prediction. Then, the frequency of an event occurrence in each bin of the rank histogram is computed and plotted against the bins. The number of bins equals the number of members plus one. A perfectly reliable ensemble shows a flat Talagrand diagram, where the bins all show the same frequency. Figure 6 shows the rank histogram for the four applied ensemble techniques, for both pollutants.

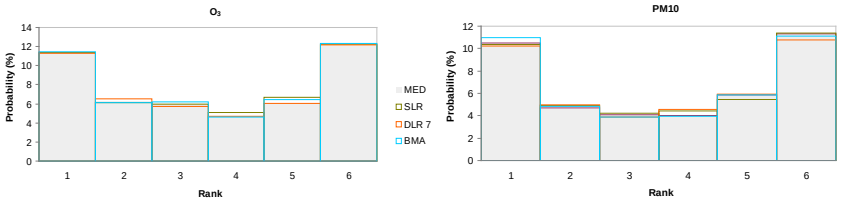


Figure 6. Talagrand (rank histogram) for the different ensemble techniques, and for both pollutants (O_3 and PM10).

The Talagrand diagram confirms the similarity of the four ensemble techniques. There is no method significantly more valuable than the others. Nevertheless, the SLR and BMA approaches have total areas (sum of each bin area) slightly superior than DLR and MED ensemble. All the histograms reveal heterogeneities in the distribution of the ensemble pollutant values relative to the distribution of observations, although smaller than the original models (Monteiro et al., 2011). The “U” shape exhibited results from an underestimation of the ensemble spread in comparison with the observations. The higher number of counts for the first and last bins (1 and 6), much larger than the other bins, reflect the ability of the ensembles to simulate low and high (peaks) levels for both pollutants. The symmetric “U” shape confirms the use of unbiased data.

SUMMARY AND CONCLUSIONS

In this study, different ensembles methods were applied to improve the estimation of surface O_3 and PM10 concentrations over Portugal, using five different air quality models, and were evaluated/intercompared against observations from the monitoring surface network. Models have been run for the period of July 2006 with the same emission data input, but different own configuration. Four different types of ensemble were generated using the bias-corrected sets of the five model simulations: median ensemble (MED); static (SLR) and dynamic (DLR) liner-regression ensembles and also the BMA

ensemble. Results pointed out that all bias-free ensembles performed very similar and significantly better than single unbiased models. This similar behaviour can be explained by the use of the bias-corrected models ensemble, either evenly or as a weighted average. Nevertheless, statistical results show that the application of the weights slightly improves ensemble performance when compared to those obtained from the median ensemble. The statistical analysis together with the probabilistic measures (histogram diagram), constructed for the four types of ensembles, demonstrate that the SLR and BMA bias-corrected ensembles present the best performance. The most appealing feature of the SLR method is its effortless implementation giving it an advantage over BMA approach, which requires a larger data processing. The above results confirm the advantage of the ensemble approach for air quality assessment. In particular, it is noted that the skill improvements from both bias correction and ensemble techniques are greater for a variable with low forecast skill (PM10) than for ozone. Results can be improved if emissions data will also be perturbed, and not only the meteorology and chemistry, since the ensemble result will have a probability distribution function with higher verification. The ensemble methodology can be particularly important for forecasting purposes, for which no monitoring data can be used for the assessment and whenever the model results are used to support decision making or regulatory purposes.

ACKNOWLEDGEMENTS

The authors acknowledge the Portuguese Environmental Protection Agency for the observational dataset support and to the Portuguese 'Ministério da Ciência, Tecnologia e Ensino Superior' for the PhD grant of I. Ribeiro (SFRH/BD/60370/2009) and post doc grant of A. Monteiro (SFRH/BPD/63796/2009) and J. Ferreira (SFRH/BPD/40620/2007).

REFERENCES

Borrego C., Monteiro A., Ferreira J., Miranda A.I., Costa A.M., Carvalho A.C., Lopes M.: 2008. Procedures for estimation of modelling uncertainty in air quality assessment. *Environment International*,**34**, 613-620.

Delle Monache L., Deng X., Zhou Y., Stull R., 2006a: Ozone ensemble forecasts: 1. A new ensemble design. Journal of Geophysical Research 111, D05307. doi:10.1029/2005JD006310.

Delle Monache L., Nipen T., Deng X., Zhou Y., Stull R., 2006b: Ozone ensemble forecasts: 2. A Kalman filter predictor bias correction. *Journal of Geophysical Research*,**111**, D05308. doi:10.1029/2005JD006311.

DjalalovaI., Wilczak J., McKeen S., Grell G., Peckhama S., Pagowski M., DelleMonache L., McQueen J., Tang Y., Leeg P. et al., 2010: Ensemble and bias-correction techniques for air quality model forecasts of surface O_3 and PM2.5 during the TEXAQS-II experiment of 2006. *Atmospheric Environment*,**44**, 455-467.

Dudhia J., 1993: A nonhydrostatic version of the PennState/NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Monthly Weather Review*, **121**, 1493-1513.

ENVIRON, 2008: User's guide to the Comprehensive Air Quality model with extensions (CAMx) version 4.50 (May, 2008), http://www.camx.com.

Elbern H., Strunk A., Schmidt H., Talagrand O., 2007: Emission Rate and ChemicalState Estimation by 4-Dimensional Variational Inversion. *Atmospheric Chemistry Physics*,**7**, 3749-3769.

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., Van der Auwera L., 2004: Ensemble dispersion forecasting, Part II: application and evaluation. *Atmospheric Environment*, **38 (28)**, 4619-4632.

Hurley P., Physick W., Luhar A., 2005: TAPM - A practical approach to prognostic meteorological and air pollution modelling'. *Environmental Modelling & Software*, **20**, 737-752.

Krishnamurti T.N., Kishtawal C.M., LaRow T.E., Bachiochi D.R., Zhang Z., Willford C.E., Gadfil S., Surendran S., 1999: Improved weather and seasonal climate forecast from multimodel superensemble. *Science*,**285**, 1548-1550.

McKeen S., et al., 2005: Assessment of an ensemble of seven real-time ozone forecasts over eastern North America during the summer of 2004. *Journal of Geophysical Research*,**110** (D21).

Monteiro A., Borrego C., Miranda A.I., Gois V., Torres P., Perez A.T., 2007: Can air quality modelling improve emission inventories?. In: Proceedings of the 6th International Conference UAQ, 26-30 March, Limassol, Cyprus, 13-14.

Pagowski M., Grell G.A., Devenyi D., Peckham S., McKeen S.A., Gong W., Delle Monache L., McHenry J.N., McQueen J., Lee P., 2006: Application of dynamic linear regression to improve the skill of ensemble-based deterministic ozone forecasts. *Atmospheric Environment*, **40**, 3240-3250. doi:10.1016/j.atmosenv.2006.02.006.

Pagowski M., Grell G.A., McKeen S.A., Dévényi D., Wilczak J.M., Bouchet V., Gong W., McHenry J., Peckham S., McQueen J., Moffet R., Tang Y., 2005: A simple method to improve ensembel-based ozone forecasts. *Geophysical Research Letters*, **32**, L07814. doi:10.1029/2004GL022305.

Potempski S., Galmarini S., 2009: Est modus in rebus: analytical properties of multi-model ensembles. *Atmospheric Chemical Physics*, **9**, 9471-9489.

Riccio A., Giunta G., Galmarini S., 2007: Seeking for the rational of the Median Model: the optimal combination of multi-model ensemble. *Atmospheric Chemistry and Physics*,**7**, 6085–6098.

Schaap M., Timmermans R.M.A., Sauter F.J., Roemer M., Velders G.J.M., Boersen G.A.C., Beck J.P., Builtjes P.J.H., 2008: The LOTOS-EUROS model: description, validation and latest developments. *Int. J. Env. Pollution*, **32(2)**, 270-90.

Schmidt H., Derognat C., Vautard R., Beekmann M., 2001: A comparison of simulated and observed ozone mixing ratios for the summer of 1998 in Western Europe. *Atmospheric Environment*, **35**, 2449-2461.

Stensrud D. J., Yussouf N., 2003: Short-range ensemble predictions of 2-m temperature and dewpoint temperature over New England. *Monthly Weather Review*,**131**, 2510– 2524.

Wilczak J., McKeen S.A., Djalalova I., et al., 2006: Bias-corrected ensemble and probabilistic forecasts of surface ozone over eastern North America during the summer of 2004. *Journal of Geophysical Research*, **111**, D23S28. doi:10.1029/2006JD007598.