ENSEMBLE and AMET: two systems and approaches to a harmonized, simplified and efficient facility for air quality models development and evaluation

S. Galmarini[1], R. Bianconi[2], W. Appel[3], E. Solazzo[1], S. Mosca[2], P. Grossi[2], M. Moran[4], K. Schere[3], S.T. Rao[3]

[1]*Joint Research Centre, European Commission, ISPRA, Italy;*

[2]*Enviroware srl, Concorezzo (MB), Italy;*

[3]*Atmospheric Modelling and Analysis Division, U.S. Environmental Protection Agency, Research Triangle Park, NC, USA*

[4]*Air Quality Research Division, Science and Technology Branch, Environment Canada, Toronto, Canada*

Abstract

The complexity of air quality modelling systems, air quality monitoring data make ad-hoc systems for model evaluation important aids to the modelling community. Among those are the ENSEMBLE system developed by the EC-Joint Research Center, and the AMET software developed by the US-EPA. These independent systems provide two examples of state of the art tools to support model evaluation. The two systems are described here mostly from the point of view of the support to air quality model users or developers rather than the technological point of view. While ENSEMBLE is a web based platform for model evaluation that allows the collection, share and treatment of model results as well as monitoring data, AMET is a stand alone tool that works directly on single model data. The complementarity of the two approaches makes the two systems optimal for operational, diagnostic and probabilistic evaluations. ENSEMBLE and AMET have been extended in occasion of the AQMEII two-continent exercise and the new developments are described in this paper, together with those foreseen for the future.

Key words: model evaluation, model community, monitoring information, information technology

1. Introduction

In atmospheric dispersion and air quality modelling, the evaluation of the model performance has long been a topic of research activities and establishing best practices (e.g. Oreskes et. 1994, Steyn and Galmarini, 2008). Before we tackle the model evaluation topic per-se, it is probably appropriate to give a couple of definitions. We will refer here to *atmospheric dispersion models* as the models that simulate the dispersion of a passive, decaying or reactive species released from a point source, be it a stack or a limited-area source. No limit exists to the scale of application which can extend as far as global. These are the models normally used for emergency preparedness and response applications. With *air quality model* we refer however to a model which deals with distributed sources of various chemical precursors and that treats the chemical transformations occurring in a volume of air that can range from the meso- to the global-scale in a three-dimensional domain. In the two cases the dynamic fields are acquired from meteorological models. The final application of these models is in the realm of support to air quality policy planning and regulation.

The verification of the model capability to adhere to experimental evidence has occupied atmospheric scientists in recent decades. The complexity of the atmospheric system in terms of spatial and temporal variability, stochasticity and scales, makes the task of evaluating a model particularly complicated. The collection of experimental evidence, representative of space and time scales, which could be used for evaluation of a time- and space-averaged model result, has always constituted the first burden in the practice of model evaluation. To date, operational networks of instruments provide organised, quality-checked information useful in this context for both atmospheric dispersion and air quality models. Routine monitoring data however are often insufficient to assess the performance of a model in depth, as they may only provide evidence of the general correspondence of model results with surface-based point measurements that are typical in such network data. These types of evaluations do not provide clear indications of the veracity of the modelled-chain of processes that leads to the model result. In other words, using a common expression, we would not know whether the right model results were obtained for the right reasons.

The problem becomes particularly complicated in the case of predictions where data are not sufficiently available. This is the case, for example, in atmospheric dispersion modelling or scenario studies in air quality. In such cases as well as in the case of a total absence of experimental evidence, a reasonable alternative to model evaluation is model inter-comparison, meaning that models are applied to a common case study and inter-compared. The relative differences in model results are useful indications of uncertainties in modelling science or data that need attention.

The Joint Research Center of the European Commission and the USEPA over the past two decades have played an important role in promoting model evaluation and bringing together the European, US and international model communities in promoting best practices in model evaluation. These activities were intended to develop and promote standards in model evaluation (e.g. Galmarini et al., 2010). Over the last decades this practice in particular has been extremely important in bringing together communities of model developers and users and has led to very important developments beyond the singular practice of model evaluation.

Given the complexity of air quality modelling systems and diversity of information necessary to run them and to evaluate them, the availability of tools that could assist modellers in the latter is very important. On the one hand one wants to guarantee that model data produced by different sources are handled in the same way. At the same time one should consider that diverse monitoring information is assembled, organised and harmonised so that all users could use the same source of information efficiently. In this context diversity pertains not only to the nature of the information, (meteorology, micro-meteorology, atmospheric chemistry), but also to the fact that same atmospheric species could be measured differently by different networks. In parallel when evaluating a model one would like to use similar metrics and a standard set of tools so that all model performances are assessed against the same ruler and terms of comparison.

Toward these ends, information technology has provided us with useful means for an easier, consistent, standardized and practical way of evaluating models. Over the last decade the European Commission Joint Research Center and US-EPA have independently invested resources in the development of two distinct systems for model evaluation, namely, the ENSEMBLE system (Galmarini et al., 2004) and AMET (Appel et al. 2011). While they both work toward the same goals they are different in philosophy and approaches and cover, in a complementary way, different aspects of the problem.

Similar systems exist for more specialised applications such as AeroCom (Textor et al. 2006; http://nansen.ipsl.jussieu.fr/AEROCOM/), which provides an online access to consultation of global model predictions observations of various kinds relating to aerosols and aerosols components. Off line model evaluation tools were developed over the years on the footprints of AMET such as BOOT (Chang, 2002; Chang and Hanna, 2004) specialized in atmospheric dispersion models and the CITY-DELTA tool still from the Joint Research Center (Thunis et al, 2008) specialized in the assessment of differences in air quality scenarios at the mesoscale models.

ENSEMBLE and AMET have been used extensively for the AQMEII activity and without them a project of this scale could not happen in the time scale in which it was performed. In this paper the two independent systems are presented in terms of their usefulness for both model evaluation and inter-comparison application, including their future developments and synergies. The perspective presented is that of the model developer/user rather than the information technology side to give an idea on the amount of resources and rapid advancements to the atmospheric science discipline such systems can provide. A model evaluation framework adopted by the AQMEII activity (Dennis et al., 2010) that considers four types of evaluation: operational, diagnostic, dynamic, and probabilistic. Operational is intended as direct comparison of model results with observations which may lead to assessment pass or fail; diagnostic is intended as investigation on the reasons for a pass or a fail; diagnostic would lead to determining the sensitivity of model results to input data, whereas the probabilistic evaluation aims at providing an assessment of confidence in model results and uncertainty. We will show that the two system serve all these type of evaluation thus allowing a thorough and complete analysis of model performance.

2. Evaluating models through a web-distributed service: the ENSEMBLE system

ENSEMBLE was developed in 2000 as a system for the real-time acquisition and consultation of the results of several models produced by different groups on the same case study. The field of application was long range atmospheric dispersion forecasting. Details on that application can be found in Galmarini et al., 2001, Bianconi et al.; 2003, Galmarini et al., 2004a and b, Potempski et al.; 2008, Galmarini et al. 2008.

Figure 1~~Figure 1~~ shows schematics of the concept behind the ENSEMBLE system. ENSEMBLE is a web distributed platform for model evaluation and inter-comparison. The advantages of a web-distributed service for model evaluation can be listed as:

- Model results can be sent to a central collection point simply by complying with a specific data format.
- The model user can concentrate on the model results only. The full set of model data can be sent to the system that will treat and transform them according to the queries of the remote data users.
- Monitoring data can be included in the system and made available to the modeling community.
- Coupling of model and monitoring data can be completely delegated to the system. All the work of collection, harmonization and upload is provided by the system administrators rather than the whole modeling community, with two clear advantages:

    - Resources can be efficiently compartmentalized
    - All the modeling community will access the same kind of information

- All modeling groups can consult, inter-compare, and evaluate their results with those of others, without having to exchange/share large sets of data. Every member sends one set and accesses the others including the monitoring data, rather than receiving all model sets and monitoring data.
- Any remote user can conduct an independent data analysis using all the information available, and a suite of statistical and graphical tools that use data on the fly.
- The case study information (input data) can be archived and distributed through a central point while at the same time being resident together with all the study results and associated monitoring data for future consultation and re-analysis. This relates to emissions/source details as well as meteorology.
- Independent communities of modelers can work with the system without interfering with one another or seeing each other results unless requested

Over the years the advantages offered by the approach became clearer and the system applicability was extended from dispersion model applications to air quality model applications. The extensions related to:

- Applicability to an unlimited number of variables (any scalar)
- Applicability to domains that could range from small to global scale and any grid resolution
- Unlimited time extensions and resolution of the simulations
- Acquisition of monitoring data of any nature from point measurements to satellite imagery.

2.1 ENSEMBLE for AQMEII

The most relevant improvements to the ENSEMBLE system have been performed for the AQMEII activity (Galmarini et al. 2010; Galmarini and Rao, 2011; Rao, et al.; 2011). All improvements were made having in mind the general applicability of the system rather than being made ad-hoc for a specific application, to guarantee continuing applicability of the system to a wider variety of cases. For the sake of synthesis we will present only the major changes and the ones that show the real advantages for the modeling community.
Figure 2.**Error! Reference source not found.** is probably the best example of the versatility and the potential of the system for model evaluation. It relates to the possibility of a direct simultaneous representation of modeling and monitoring data in a geo-referenced framework based on a Google

Maps embedded in the ENSEMBLE web interface. The figure shows the North American modeling domain of the AQMEII two continents model evaluation exercise. The monitoring data have been uploaded for a large variety of species and meteorological variables.

As an example, the sequence of plots shows the interface where the 1447 North American (troposheric) ozone monitoring stations available are presented as clusters. The more a user zooms in the domain, the more the clusters split up into smaller ones until the individual stations are visible (panel c of Figure 2**Error! Reference source not found.**). The user can diagnostically look at individual stations by clicking on the markers. Each marker is associated with a one-year time series of ozone measurements and model results which can be seen interactively within an information window. This is a very useful tool for data screening and mining and it is the first approach to the model inspection.

The user can also select a portion of the network through a *rubber band* delimiter, and within that identify stations according to physical characteristics (urban, rural, semi-rural, altitude range), species characteristics (e.g. concentration threshold levels for filtering hotspots). By dynamically dragging the mouse over the time series, a specific time period can be highlighted, and zooming can be performed down to the time resolution of the dataset. The set of markers selected geographically through the rubber band and on the characteristics of the stations can be saved in a named *data pool* that can be used afterwards to perform the other statistical evaluations offered by the system without having to repeat the selection.

These statistical evaluations can be performed by several users simultaneously and on the fly through the web interface. After selecting one of the saved data pools, the user can carry out a *pool analysis*, where data at the stations in the pool are first averaged in space at the same local time, or perform an analysis by which measured and modeled pairs are kept independent. In both cases the pairs can then be optionally averaged in time producing hourly maximum, daily maximum, daily average, and mean day. The resulting time series can then be given as input to the statistical packages of ENSEMBLE, or they can be directly represented graphically. The versatility and the practicality of these operations greatly facilitate the model data inspection and analyses as he does not have to leave the system but keeps on operating with the web interface.

Apart from on-line analyses, it is also possible to export monitoring and model data for the stations within the pool for a deeper or different ad-hoc analysis or use.

Monitoring data need to be prepared in a specific format to be included into the ENSEMBLE database. A number of heterogeneous sources and formats of data were taken into account to prepare the AQMEII dataset. Available measurements include values at ground and aloft of meteorological variables, trace gas and aerosol concentrations, and dry and wet deposition of gaseous and aerosol pollutants. ENSEMBLE now stores time series of measurements taken at thousands of monitoring stations in the two continental regions.

Measurements can be uploaded as:

- time series at receptors at fixed height or along vertical profiles
- sparse (in space and time) data representing three-dimensional trajectories

Model data can be uploaded in several forms such as:

- model data at the receptor points (to produce the maps presented above)
- vertical profiles at a given location
- as three dimensional volumes of gridded data

The upload can be made for data of any time length and frequency and number of species. Model predictions are uploaded in UTC time coordinates whereas measurements are normally local standard time (LST). The system takes care of pairing the data and it accounts for different time zones.

The three dimensional upload and consultation of modeling and monitoring data is particularly advantageous in the case of aircraft measurements. When results from a community of multiple models need to be compared with aircraft measurements, it is not very practical for the modeling groups to receive a set of concentrations at (x,y,z,t), corresponding to the aircraft trajectory and having to extract from the model the corresponding predictions or simulated values. This is particularly impractical when the number of measured profiles can be of the order of several hundreds if not thousands. Within AQMEII, for example, we were granted access to the MOZAIC dataset (http://mozaic.aero.obs-mip.fr/web/) for 2006 which allowed us to use about 2000 vertical profiles measured by instrumented commercial aircraft landing and taking off to and from 12 selected airports in North America (Portland, Vancouver, Atlanta, Charlotte, Boston, Chicago, Montreal, New York, Philadelphia, Toronto, Washington) and three in Europe (Frankfurt, Munich and Vienna). These profiles relate to a number of species and meteorological parameters. For 2006 the following measurements were available: O_3, CO, temperature, relative humidity, wind speed and direction. A special extension of ENSEMBLE was developed for this specific application that also has general applicability, illustrating a clear example of how modeling activities can be greatly facilitated when systems like ENSEMBLE can be made available. The strategy adopted consisted of plotting all the aircraft trajectories up to the altitude of approximately 13 km at all available airports for 2006 (panel d) in the two continents. Domains have been identified at each airport location that contained all the trajectories for one year. This was particularly useful in the case, for example, of the Northeast US where trajectories of aircrafts landing in Washington passed over New York airports too. Similarly in Europe, all the trajectories relating to Vienna also passed over Paris and Frankfurt. After the identification of the domain of pertinence, 13 vertical levels above ground where identified (at 0, 100, 250, 500, 750, 1000, 2000, 3000, 4000, 5000, 6000, 7500, 8500 m) which are relevant to the analysis of the model results. All modelers were then asked to deliver the model data corresponding to the grids covering the domain of pertinence of the flights at the identified altitudes with the resolution of the common grid. Once a specific airport, date, and flight were selected, the ENSEMBLE system, through the interface presented, automatically extracted the model data from the domain volume and coupled them with the MOZAIC profiles in x,y,z and t terms. The simplification on the modeling side is enormous in terms of data preparation and time for the model evaluation. In a sense all what the modelers were asked to do was to extract from the 3D volume of their AQMEII domain a sub-volume of data overlaying the red squares in Figure 3~~Figure 3~~ for all hours of 2006 and send it to the ENSEMBLE system, and then connect to the system and analyze their results. The very same procedure can be collapsed in x and y to a single point for importing vertical profiles.

The real advantage of a system like ENSEMBLE is the possibility of easily generating ensemble results given the availability of several model sets. Specifically for the AQMEII two-continent exercise a number of new analysis possibilities have been introduced. Other than the classical mean and median or percentile representation of multiple models (Galmarini et al. 2004a), the winsorized and the trimmed mean at 10 and 25% have been included. The latter two are typically ways of averaging that takes into account the presence and role of outliers. The winsorized mean (Wilcox and Keselman, 2003) is obtained by sorting a sample and then replacing a given percentage of its tails with the minimum and maximum values that are not within the tails. The trimmed (or truncated mean) is instead obtained by removing the tails for a given percentage of data within the sorted sample.

ENSEMBLE provides a wide spectrum of possibilities to the system users for numerical and graphical analyses. For example, the spread of prediction is an efficient way of representing in space multiple model results and the level of agreement of the various predictions (Figure 4~~Figure 4~~). Classical representations like Box-and-Whisker, soccer-goal, Taylor, quantile-quantile, ROC, Brier score and Talagrand diagrams are available for a quick and efficient consultation of model and – whenever available - monitoring data in space or globally, regardless of the space or time location of a data value.

ENSEMBLE can process on the fly very large datasets: an example is the Box-and-Whisker diagram made with 378 one-year hourly time series of ozone measured and predicted by 12 models on the AQMEII European domain (about 40 millions of measurement-model pairs).

Currently developments are taking place that include among other thing the coupling of the system with others for the collection of information on line. Among them the NASA's GIOVANNI system (Berrick, S.W et al, 2009; http://disc.sci.gsfc.nasa.gov/giovanni) will provide a very large sets of satellite imagery and measurements that most of all are validated and quality assured.

3. Evaluating models through a standalone service: the AMET system

The Atmospheric Model Evaluation Tool (AMET; Appel et al., 2011) was designed to aid in evaluation of meteorological and air quality model output, and has been utilized by a number of the AQMEII modeling groups to aid in the evaluation of their model output.

The AMET software is a collection of several software packages integrated together to provide a system for storing and analyzing meteorological and air quality observations and model output. AMET uses FORTRAN and Perl code to pair model output with observations in space and time. AMET uses the MySQL (http://www.mysql.com) relational database to store the paired model output and observations to allow for easy retrieval of the data, with the added benefit of being able to query the data based on various metadata, including time period (e.g. month/day), state or country, land-use category, time of day and number of other factors. The ability to query the data stored in the AMET system allows for a more in-depth analysis of the model performance, as the data can be quickly and easily subset into smaller sets of data (e.g. spatially or temporally), which may be more appropriate for some types of analysis. Finally, AMET uses the R program to compute statistics and create plots (e.g. scatter and box plots).

Although the release version of AMET air quality module (AMET-AQ) was specifically designed to compare observations and model output for simulations covering North America, it was necessary to extend the capabilities of AMET to analyze observations and model output for other regions of the world, most specifically Europe for the AQMEII project. Extending AMET to work for Europe required generally minor modifications to AMET-AQ code. The first step required was to create the paired observation and model data file that will be loaded into the AMET database. The release version of AMET includes several scripts for pairing the model output with observations in time and space. However, the included scripts are specific to networks in the United States, and therefore do not work by default with European networks, primarily due to differences in the data format. The simplest solution to this problem is to reformat the data from the European networks into a format that AMET-AQ is already familiar. The data format for the U.S. SEARCH network was found to be the most flexible format to use for the European networks. Once the data have been reformatted into a common format (e.g. SEARCH data format), the process for pairing the model and observed data is the same. In addition to reformatting the observation data file, a site file listing the site, latitude, longitude and hour offset from Greenwich Mean Time is also required.

Once the file with the paired model and observation data has been created, it needs to be loaded into the AMET-AQ database. Since AMET-AQ is not setup by default to process data for the European air quality networks, several of the included Perl scripts need to be modified slightly to include the new networks, specifically the AQ matching.

The database also requires site metadata for each network to be loaded, which allows the data to be queried based on any of the parameters included in the metadata file (e.g. land-use). Site lists already exist for the U.S. networks and can be used as examples for creating site lists for new networks. The metadata in the network site lists are then loaded into the database, which would require minor modifications to include the new network.

Once the data for the new network have been loaded into the database, the process for performing analysis of the data is the same, and would not require any modifications to the R analysis scripts, save for one minor modification that affects AMET spatial plots. By default, the function . limits the plotting area to North America, and will not work for other plotting regions of the world (e.g. Europe).

The difference in approach between ENSEMBLE and AMET is basically the different type of applications that they can serve. While ENSEMBLE can act as web-distributed large-data-collection four-dimensional screening tool, that helps to identify portion of data sets and model sets of specific interest and assists the four types of evaluation of Denis et al., (2011) at large: AMET is particularly suitable for diagnostic model evaluation where the knowledge on the model workings can be associated to the comparison of model results with data for a deeper understanding of the reasons of possible deviations. Multiple runs of different model versions can be analyzed and inter-compared whenever measured dataset are at hand. A very advantageous synergy is obviously realized by the combination of the two tools.

We shall finally notice that, although ENSEMBLE and AMET are two independent systems, data gathered in ENSEMBLE can be easily downloaded and used in AMET.

4. Conclusions

Over the past couple of decades a number of modeling consortia have been established where a large number of institutions have come together with the scope of comparing the results of their models with others in the scientific community. It is from discussions within these consortia that activities of such nature can only be beneficial to promoting progress in modeling science and application, in particular in the case of air quality modeling. The progress produced by these communities in terms of model development, improvement, understanding, information sharing just to mention some of the aspects has been tremendous. Working on the same case studies and comparing model results is an extremely useful activity, which becomes more beneficial when monitoring data are available and the models can be quantitatively evaluated.

The complexity of air quality modeling systems and the management of meteorological and air quality data provide challenges to the task of model evaluation. The development of information technology has rapidly provided the atmospheric science community with important technical supports. In this paper two analysis facilities have been presented that can greatly improve the possibilities for model inter-comparison and evaluation. The latter greatly facilitate the task of combining large amount of information in four dimensions provided by several different sources to analyze and come to a rapid assessment to the equality of the information. Just as a term of comparison in 1994 the European Tracer Experiment (ETEX, Girardi et al. 1998) model evaluation was organized and after the collection and analysis of air samples more than a year was necessary to collect model results and compare them with data. The very same kind of activity can these days be performed with the ENSEMBLE system on the fly almost in real time and tools like AMET can clearly support the preparation of individual model results prior to the submission to the scientific community. This considering that ETEX considered only one pollutant and the predictions extended to 90 hours. The unprecedented scale of the AQMEII activity (two continents, one year, over twenty groups) could not be realised in a time spam of over a year (including the model simulations) without the possibility of relying on infrastructures like ENSEMBLE and AMET. Systems like ENSEMBLE and AMET can be of very important assistance therefore freeing resources that could be dedicated to more model-relevant aspects. The organization of information on web based platforms that can be considered quality assured is a very important asset for model users. Data collection, harmonization and quality assurance is an important aspect of the very delicate process of model evaluation, but yet another important point is the correct coupling model to measured information. Resources are often not available to be spent in data collection harmonization and quality assurance. All this can be realized very cost effectively with facilities like ENSEMBLE and tools like AMET. Just to give an example of the level of data harmonization required in the case of air quality, the treatment of the monitoring data relating to PM from the mere North American networks required the treatment and harmonization of more then 70 data formats, a task that cannot be demanded to a modeler and if treated centrally will offer more guarantees of uniformity and coherence through out the modeling community.

Several terabytes of monitoring data exist that remain dormant for the simple reason that if would cost too much to treatment them and put them in a usable form by a single group. This is particularly the case if we consider datasets that go from deposition, micrometeorology, air quality and meteorology, therefore very different in nature and scale of relevance and use. Data are routinely or occasionally collected but never combined coherently and homogenously so that the modeling community can use them. This is mainly due the difficulty of find, assembling, analyze and harmonize the information. Comprehensive datasets are the missing elements in model evaluation. The task of collecting and assembling such information is huge and would scare off any modeler who would finally limit the evaluation of the model to few parameters. Information like that, when treat and made available can be as precious as a yet a brand new experimental campaign. Systems like ENSEMBLE can act as data collection point where the work can be done once and for the whole community and tools like AMET can truly help a more rapid increase in model confidence and evaluation. The potential made available by tools like ENSEMBLE, AMET and other existing is enormous and should be more exploited.

Aknowledgements

The authors would like to thank Dr C. Hogrefe (Bureau of Air Quality Analysis and Research, New York State Department of Environmental Conservation) for is constant support and advice on the collection and harmonization of monitoring information amd Drs R. Gilliam (US-EPA) and J. Zhang (Environment Canada) for their contribution in data preparation.

The authors also acknowledge for the strong support of the European Commission, Airbus, and the Airlines (Lufthansa, Austrian, Air France) who carry free of charge the MOZAIC equipment and erform the maintenance since 1994. MOZAIC is presently funded by INSUCNRS (France), Meteo-France, and Forschungszentrum (FZJ, Julich, Germany). The MOZAIC data based is supported by ETHER (CNES and INSU-CNRS).

References

Appel, K. W., Gilliam, R. C., Davis, N., and Zubrow, A., 2011. Overview of the Atmospheric Model Evaluation Tool (AMET) v1.1 for evaluating meteorological and air quality models, *Environ. Modell. Softw.* 26, 4, 434-443,

Berrick, S.W., Leptoukh, G., Farley, J.D., and Rui, H., 2009. Giovanni: A Web service workflow-based data visualization and analysis system. IEEE Trans. Geosci. Remote Sens. 47(1), 106-113.

Bianconi R., Galmarini, S., Bellasio, R., 2004. A WWW-based decision support system for the management of accidental releases of radionuclides in the atmosphere. *Environ. Model. & Soft.* **19**, 40-411, ~~2004~~.

Chang, J. C., 2002. *Methodologies for Evaluating Performance and Assessing Uncertainty of Atmospheric Dispersion Models*. Ph.D. thesis, George Mason University, Fairfax, VA 22030-4444, 277 pp

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. *Meteorol. and Atmos. Phys*. **87**, 167-196

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., (...), Venkatram, , A., 2010. A framework for evaluating regional-scale numerical photochemical modeling systems Environmental Fluid Mechanics, 10 (4), pp. 471-489.

Galmarini S., Bianconi, R., de Vries, G., Bellasio, R., 2008. Real-time monitoring data for real-time multi-model validation: coupling ENSEMBLE and EURDEP. Journal of Environmental Radioactivity 99, 1233-1241.

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, *Atmos. Environ.* **38**, 28, 4619-4632.

Galmarini S., Bianconi, R., Bellasio, R., Graziani, G., 2001. Forecasting consequences of accidental releases from ensemble dispersion modelling. *J. Environ. Radioact.* **57**, 203-219.

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, 4607-4617.

Galmarini, S. and S. T. Rao. (2011) The AQMEII Two-Continent Regional Air Quality Model Evaluation Study: Fueling Ideas with Unprecedented Data. Atmos. Environ., 45, 14, 2464.

Gamarini S., Steyn D.G., Schere K.L. and Moran M.D., (2010), Advancing the evaluation of regional-scale air quality models, EUR 24245 EN, ISBN 978-92-79-15007-4

Girardi F., G. Graziani, D. van Velzen, S. Galmarini, S. Mosca, R. Bianconi, R. Bellasio, W. Klug and G. Fraser (1998) ETEX - The European tracer experiment. EUR 18143 EN, Office for Official Publications of the European Communities, Luxembourg, ISBN 92-828-5007-2, 108 pp.

Oreskes N., Kristin Shrader-Frechette and Kenneth Belitz, Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences, Science, Vol. 263 no. 5147 , 641-646

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

Rao, S. Trivikrama, Stefano Galmarini, Keith Puckett, 2011. Air Quality Model Evaluation International Initiative (AQMEII): Advancing the State of the Science in Regional Photochemical Modeling and Its Applications. Bull. Amer. Meteor. Soc., 92, 23–30.

Steyn D.G., Galmarini S., Evaluating the Predictive and Explanatory Value of Atmospheric Numberical Models: Between Relativism and Objectivisim, Open Atmospheric Science Journal 2, 38-45, Bentham Science Publishers Ltd., 2008

Textor, C., M. Schulz, S. Guibert, S. Kinne, Y. Balkanski, S. Bauer, T. Berntsen, T. Berglen, O. Boucher, M. Chin, F. Dentener, T. Diehl, R. Easter, H. Feichter, D. Fillmore, S. Ghan, P. Ginoux, S. Gong, A. Grini, J. Hendricks, L. Horowitz, P. Huang, I. Isaksen, T. Iversen, S. Kloster, D. Koch, A. Kirkevåg, J.E. Kristjansson, M. Krol, A. Lauer, J.F. Lamarque, X. Liu, V. Montanaro, G. Myhre, J. Penner, G. Pitari, S. Reddy, Ø. Seland, P. Stier, T. Takemura, and X. Tie, 2006. Analysis and quantification of the diversities of aerosol life cycles within AeroCom, Atmos. Chem. Phys. 6, 1777-1813,

Thunis P., C. Cuvelier, P. Roberts, L. White, L. Post, L. Tarrason, S. Tsyro, R. Stern, A. Kerschbaumer, L. Rouil, B. Bessagnet, J., Builtjes, M. Schaap, G. Boersen, R. Bergstroem Evaluation of a Sectoral Approach to Integrated Assessment Modelling including the Mediterranean Sea, Eurodelta II report., EUR 23444 EN - 2008.

Wilcox, R.R., Keselman, H.J., 2003. Modern robust data analysis methods: Measures of central tendency, Psychological Methods 8 (3), 254–274

Figure Captions

Figure 1 Schematics of the ENSEMBLE concept. The web based platform act as central collection point of monitoring data of very different nature, as well as model simulations which can be thereafter be redistributed through a web interface to remote users in the form of raw data as well as statistical treatments.

Figure 2, Sequence of passages in the web browsing of monitoring/model data from point receptor network at continental scale. The example refers to the north American network of ozone. The first panel show the complete network of available monitoring points clustered in groups. The following panel show increasing levels of zooming all the way down to the single receptor that can be visualized in terms of physical location on satellite imagery. Model and measured time series can be displayed by clicking on the station. Dynamical zooming into the time series can be realized with simple and direct movement s of the mouse.

Figure 3, Schematic representation of the strategy used in ENSEMBLE for the treatment of monitoring and modeling data relating to upper air concentrations.

Figure 4. Grid plot (top) of the median value of SO2 predicted at a given time by a set of models and corresponding spread of prediction plot (bottom) showing the difference among single models in terms of range between the 10th and the 90th percentile of the predictions.

Figure 5. Box-an-Whisker plot for 12 models against one-year hourly measurements of ozone (box number 1) at 378 selected stations on the European AQMEII domain

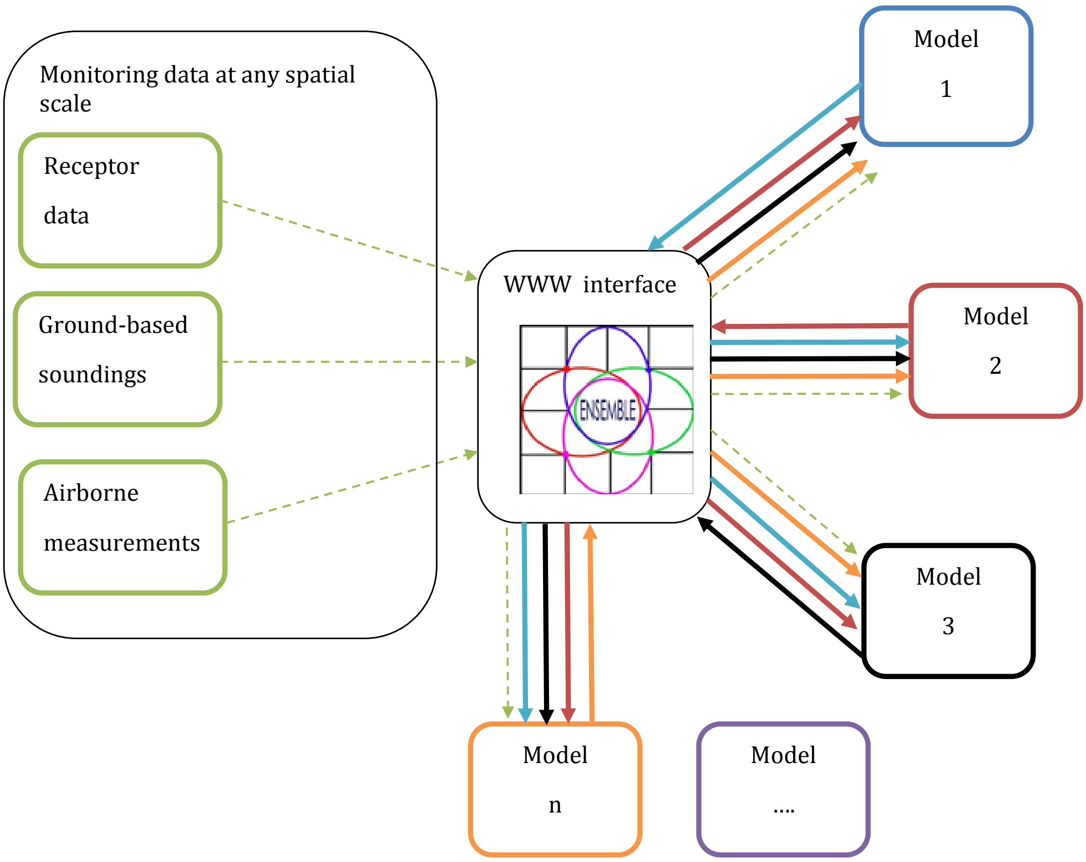


Figure 1~~Figure 1~~

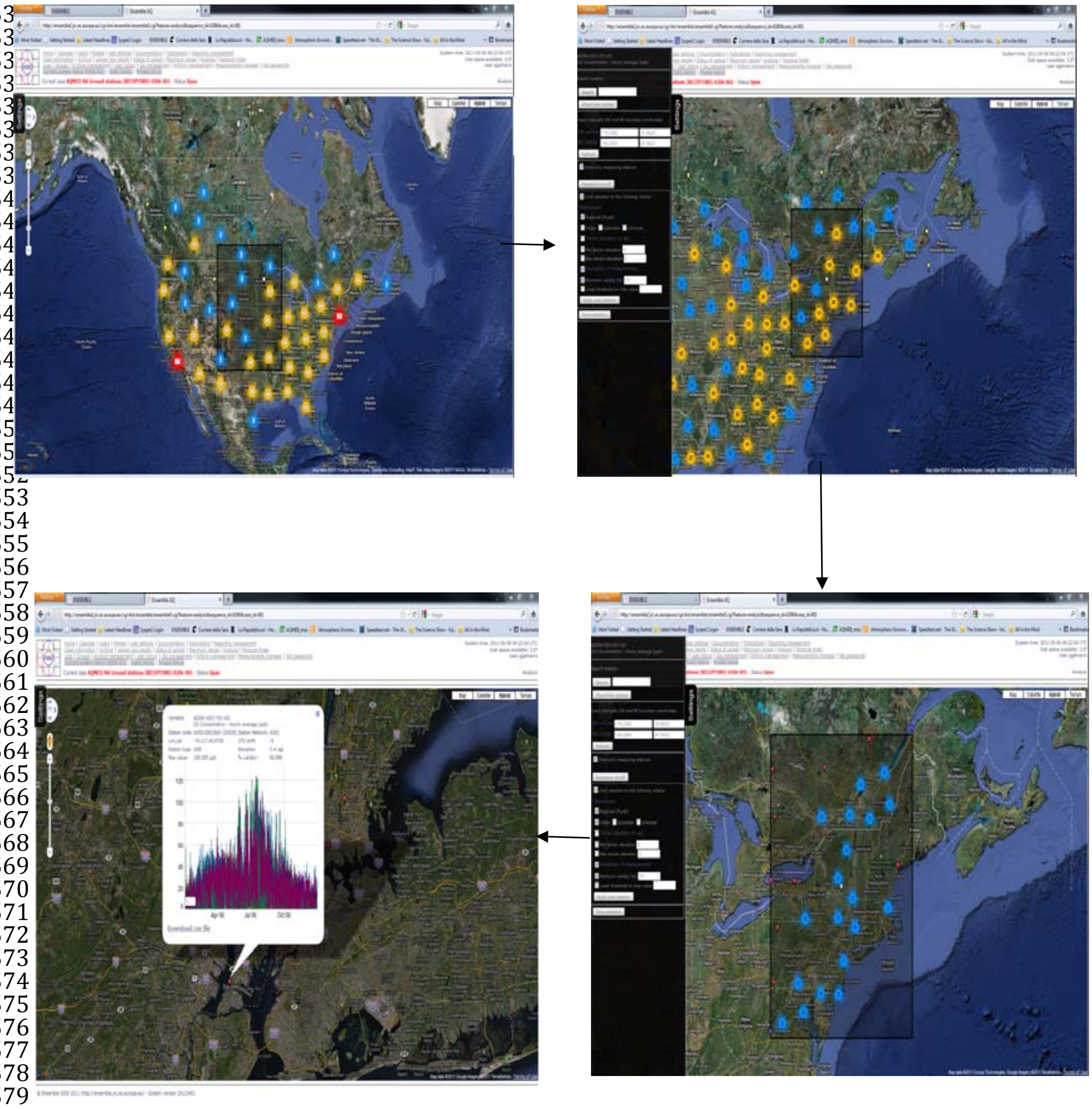
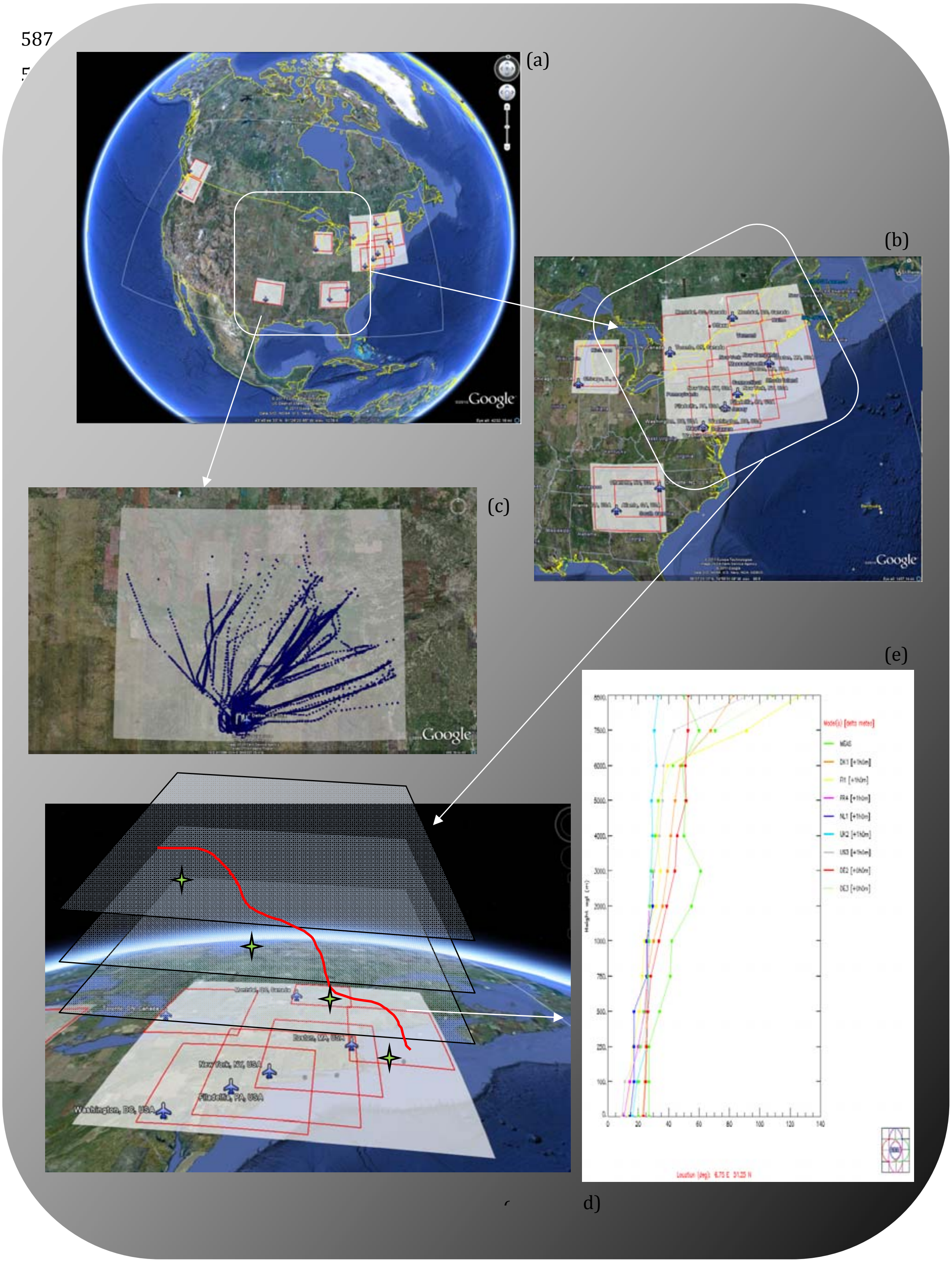
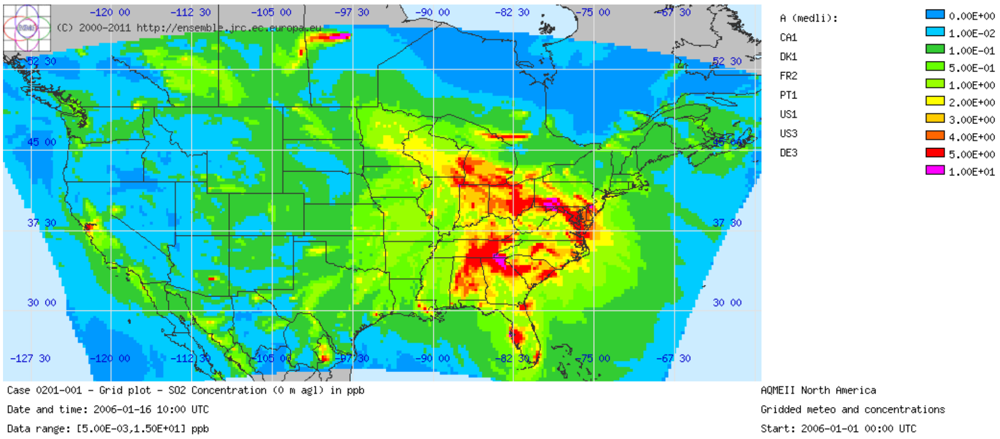


Figure 2~~Figure 2~~


(a)
(b)
(c)
(e)
d)
Google



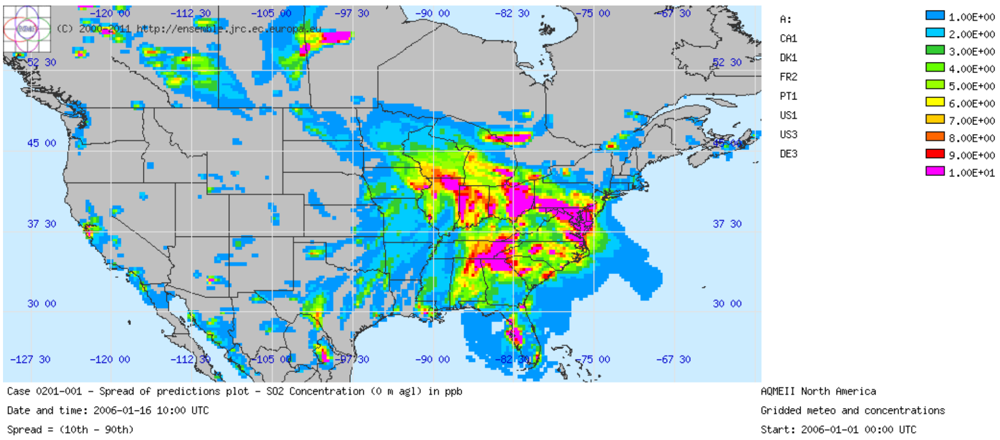


Figure 4~~Figure 4~~

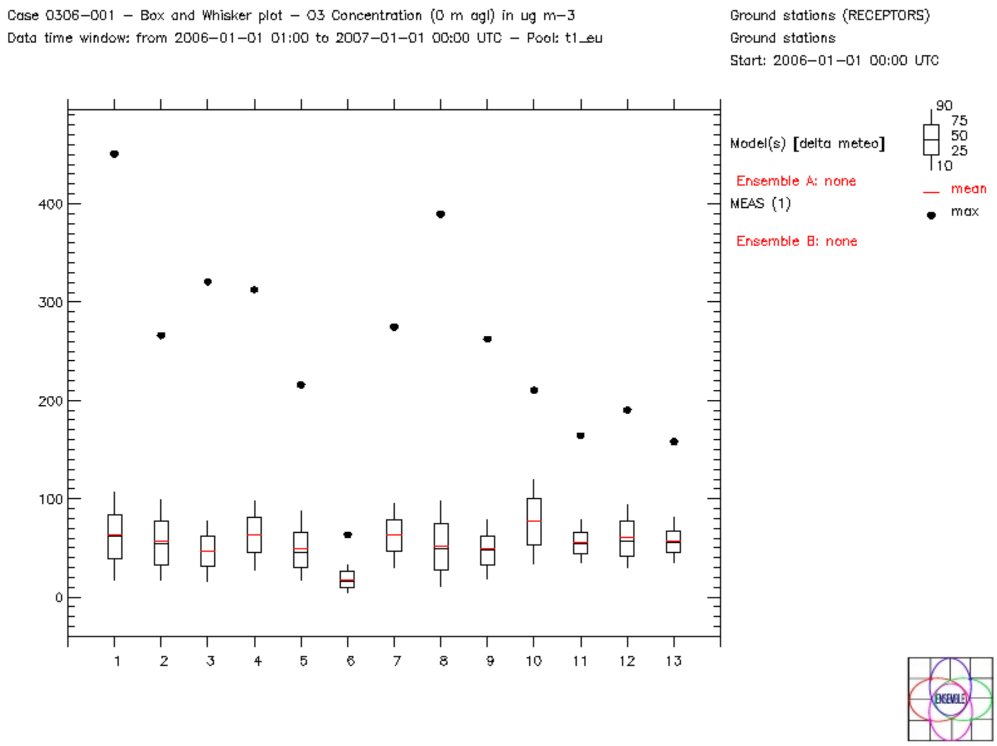


Figure 5~~Figure 5~~