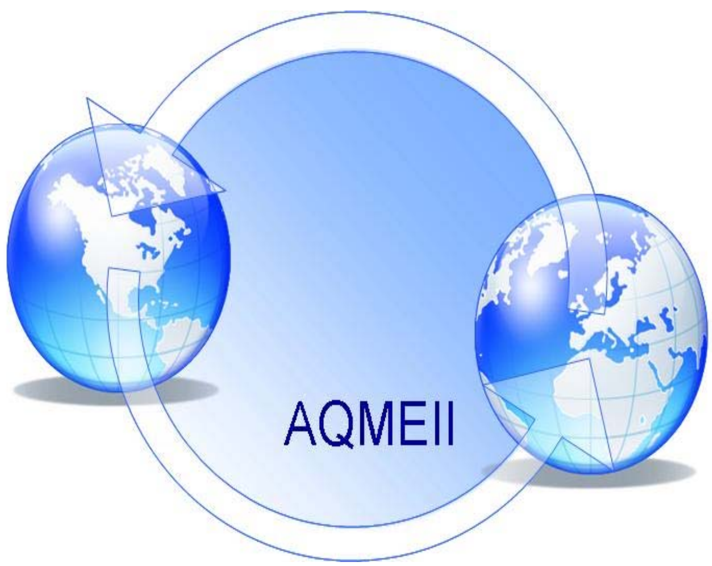
JRC Scientific and Technical Reports

Advancing the evaluation of regional-scale air quality models

Stefano Galmarini, Douw G. Steyn, Kenneth L. Schere[1], and Michael D. Moran[2]



Air
Quality
Model
Evaluation
International
Initiative

JRC
EUROPEAN COMMISSION

Institute for Environment and Sustainability

EUR 24245 EN - 2010

The mission of the JRC-IES is to provide scientific-technical support to the European Union's policies for the protection and sustainable development of the European and global environment.

European Commission
Joint Research Centre
Institute for Environment and Sustainability

Contact information:

Stefano Galmarini
European Commission
Joint Research Centre
Institute for Environment and Sustainability
Via E. Fermi, 2749 - Ispra (VA),
21020, Italy
E-mail: stefano.galmarini@jrc.ec.europa.eu
Tel.: + 39 0332 785382
Fax: + 39 0332 785466
http://ies.jrc.ec.europa.eu/
http://www.jrc.ec.europa.eu/

S.T. Rao
(1) Atmospheric Modeling and Analysis Division
National Exposure Research Laboratory
Office of Research and Development
U.S. Environmental Protection Agency (MD - E243-02)
109 T.W. Alexander Drive, Room E-240D
Research Triangle Park,
NC 27711, USA
E-Mail: rao.st@epa.gov
Tel: +1 919 541-4541
Fax: +1 919 541-1379
http://www.epa.gov/amad

Keith Puckett
(2) Air Quality Research Division
Environemnt Canada
4905 Dufferin Street
Toronto, Ontario
M3H 5T4, Canada
E-Mail: keith.puckett@ec.gc.ca
Telephone : + 1 416-739-4836
Fax : +1 416-739-4224
http://www.ec.gc.ca/

Douw G. Steyn
Address: Department of Earth and
Ocean Sciences, University of British
Columbia, Vancouver, Canada
E-mail: dsteyn@eos.ubc.ca
Tel.: +1 604 827 5517
Fax: +1 604 822 6088
http://www.eos.ubc.ca

JRC56481

EUR 24245 EN
ISBN 978-92-79-15007-4
ISSN 1018-5593
DOI 10.2788/61186

Luxembourg: Office for Official Publications of the European Communities

Printed in Italy

Introduction

This document is the culminating product of a series of ad-hoc workshops[1] led by a group of European and North American air quality scientists interested in instigating a significant advance in the way regional-scale air quality modeling systems[2] are evaluated. The initiative arising out of these workshops has come to be known as the Air Quality Model Evaluation International Initiative (AQMEII). The motivation for the workshops was a growing sense that current model evaluation practices have stagnated, and as a result are not serving research, operational and policy sectors as well as they should. While technical (often statistical) approaches have advanced in the past two decades, it was felt that present approaches are too often uncritically applied, without due consideration of the foundations upon which the techniques are based.

This document contains a set of research recommendations for advancing model evaluation. These recommendations are based on ideas generated during workshop presentations, group discussions and informal discussions that occurred at and around the workshops. The overall structure of the recommendations is drawn from a manuscript that arose out of the first (Raleigh, North Carolina, USA, August 2007) model evaluation workshop (Dennis et al., 2010). This structure identifies four modes of model evaluation:

Operational Model Evaluation: Operational evaluation involves the direct comparison of model output with analogous observations in an overall sense. It utilizes routine observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables.

Diagnostic Model Evaluation: Diagnostic evaluation examines the ability of a model to predict pollutant concentrations by correctly capturing physical and chemical processes, and their relative importance as incorporated in the model. This type of model evaluation generally requires detailed atmospheric measurements that are not routinely available.

Dynamic Model Evaluation: Dynamic evaluation focuses on the model's ability to predict changes in air quality concentrations in response to changes in either source emissions or meteorological conditions. This exercise requires historical case studies where known emission changes or meteorological changes occurred that could be confidently estimated.

Probabilistic Model Evaluation: Probabilistic evaluation attempts to capture statistical properties, including uncertainty or level of confidence in the model results for air quality management or forecasting applications. This approach is necessarily based on knowledge of uncertainty imbedded in both model predictions and observations.

---
[1] Held at Raleigh, North Carolina, USA (August 2007); Utrecht, The Netherlands (March 2008); Stresa, Italy (April 2009).
[2] In this document, "model evaluation" will be used to mean "evaluation of regional- scale air quality modeling systems".

"Operational, diagnostic, and dynamic evaluation approaches complement one another by not only characterizing how well the model simulated the air quality levels at that time, but how well the model captures the role and contributions of individual inputs and processes and the ability of the air quality model to respond correctly to changes in these factors." (Dennis et al., 2010). While it is true that all evaluation approaches use a statistical formalism, and statistical techniques, this framework conceives probabilistic evaluation as a rather more comprehensive approach than the mere application of statistical tools.

The overall purpose of this document is to organize and compile the ideas arising out of the workshops into a blueprint that will serve to advance the practice of model evaluation. It is anticipated that the ideas will inspire a broad re-thinking of the way models are evaluated, and at the same time provide specific recommendations for the improvement of model evaluation practices. An intended outcome of AQMEII is a set of international collaborations involving model evaluation exercises carried out on shared data sets using different state-of-science regional air quality models. It is intended that this initiative be open to all interested researchers.

1. Operational evaluation

A growing number of national and international organizations in North America (NA) and Europe (EU) are putting considerable effort into continuously running regional-scale air quality models for various purposes. The output from these models is being used for air quality analysis and to prepare air quality forecasts, often in the form of indices that are presented in print and electronic media and used to prepare public health advisories/warnings. As with operational weather forecast models, output from these models is continuously compared with routinely collected air quality data from monitoring networks. This constitutes a very particular type of air quality model evaluation because measurements and corresponding model output are available with extensive space and time coverage and under all weather conditions, not just during episodes of degraded air quality. In practice, operational models are run on the regional scale: 10-50 km horizontal grid spacing on continental-scale domains with hourly outputs of ozone (O_3) and fine particulate matter ($PM_{2.5}$).

Much can be learned about model performance, and the phenomenon of air pollution by performing informed operational model evaluations. Because of the particular nature of operational evaluation, a number of aspects need special attention:

1) Recognizing that models must be fit for a particular purpose, operational air quality models should be evaluated for their suitability for operational use. This is more than evaluation of model output.
2) Measurements as well as models should be fit for purpose. This means that models and measurements should be matched – on appropriate time and space scales and same chemical species.
3) Air pollution is an essentially 3-D phenomenon. Operational evaluation should be 3-D if at all possible.
4) Air pollution fields are essentially coupled space-time fields, and we should employ metrics that consider variability in space, time, and space-time together.
5) The form of model evaluation metrics (e.g., air quality index) should be determined by the form of the impact – health / ecosystem / climate.
6) Careful consideration must be given to the selection of data used for operational model evaluation. Data sets to consider are: air quality, meteorological data, surface data, profiles, remote sensing data, measurements from movable platforms – vehicles/ bicycles/aircrafts/ships – when available, super sites, 3-D data lidar, vertical profiles of O_3, satellite data. Attention must be paid to the matching of model results and measurements (this includes the idea of representativeness).
7) Operational model evaluation must be designed so as to answer specific scientific hypotheses, rather than being the mere accumulation and presentation of model-data comparison statistics. The evaluation should be based on selected metrics from the

available set of statistical and graphical analyses: time series and scatter plots including RMSE, bias, correlation coefficient both in time, space, and globally, frequency distributions of parameters, spectral analysis, spatial concentration plots and spatial difference plots (Kriged observations), spatial plots of selected statistical metrics, vertical profiles, Figure of Merit in Space of concentrations exceeding thresholds (cf. ETEX), Taylor plots/soccer-goal plots for multiple models.

All air quality models require meteorology and emissions inputs. Since model performance is relative to model input, it is evident that emissions and meteorological inputs should also be evaluated. The evaluation of meteorological input is often no more than the evaluation of the output of a meteorological model, with relatively well-established procedures. However, it must be recognized that the present meteorological network was defined in the 1950s for synoptic scale weather, and might not be optimal for present day air quality model operational evaluation as the density of stations is hundreds of km. Far more complex, and just as important is the evaluation of emission inventory input to air quality models. While this topic is of general relevance to air quality model evaluation, it has particular importance in the context of operational air quality model evaluation. There exist a few approaches to the evaluation of emissions. Some are:

1) The estimation of emissions by the use of different surrogates for actual emissions, and a comparison of the different (often only two) resultant emissions estimates. An example is the use of fuel sales and vehicle kilometers travelled, to estimate vehicular emissions.
2) The evaluation of emissions, by measurements and “inverse modelling” with short distances between emissions and measurements.
3) The evaluation of emissions using both top-down and bottom-up approaches,
4) Evaluation by direct measurement of pollutant emissions at stacks or exhaust pipes.

Development of further emissions evaluation techniques, beyond those listed is needed.

A problem facing large-domain operational air quality models is the artificial variations in the land-use parameters and most frequently and importantly the variation of emission factors associated to same activities used in statistical estimates of emissions because of different national classification schemes. These differences must be reconciled so as to reduce discontinuities in emissions inventories at national borders. Further to that, potential inconsistencies between the land-use employed for the flow generation by meteorological drivers and that used by the transport and chemical models must be eliminated or minimized

While air quality modellers and the air quality monitoring community can assist in evaluating and refining the emission inventories, the ultimate responsibility for development of emissions inventories should lie with agencies responsible for maintaining the inventories, and not with air quality modellers. Oftentimes, the spatial and temporal resolution requirements of modellers are different from those of policy analysts that assemble the emission inventories; policy analysts are interested in annualized emissions for tracking their emission control policies while air quality modellers must have high-resolution emissions data for simulating hourly concentrations of pollutants.

2. Diagnostic Evaluation

2.1 Introduction

Diagnostic evaluation is a procedure designed to identify model weaknesses or deficiencies arising from the way specific processes are represented in the model, or model sensitivity to specific input data. The approach to diagnostic evaluation should be to identify in which process/es the problem resides, estimate its/their impact on the final results, make use of updated information or modeling studies to improve that/those specific process/es that are the origin of the model deficiencies.

2.2 Identification of process model deficiencies and/or dependence on input data

The diagnostic evaluation goes through an identification phase and a correction phase. The former could result from systematic or sporadic anomalous model behaviors when compared with other models or experimental evidence. This phase could be followed by a thorough screening of the model performance for the identification of the processes responsible for the anomalies. As described in Dennis et al. (2010), **sensitivity analysis can be instrumental in this phase provided that it is performed in rigorous terms trying to cover a large portion of the parameter space governing model performance**. Techniques exist that can be readily used but they are still viewed with skepticism and are hardly applied in the field of air quality (AQ) modeling.. **Techniques and practices such those described in Saltelli et al. (2004, 2008), and Cullen and Frey (1999) are theoretically well framed and widely applied in many other sectors of environmental sciences**. **Efforts and collaborations should be established between atmospheric modeling communities and statistical/sensitivity analysis communities for the development of practices and tools for application of formal sensitivity analysis to air quality models.** The implementation of these practices could initially be time consuming, but would guarantee a precise and conclusive targeting of the model sensitivities, allowing the identification of processes or input data that need attention. Sensitivity analysis techniques also estimate the uncertainty related to the identified parameters thus quantifying their impact or transmission through the model variables.

Reference to past studies that have attempted to identify critical processes in models is an alternative that remains confined to the specific model used and the particular application, and does not guarantee an appropriate coverage of the parameter space of a particular model. In this respect, as reported by Dennis et al. (2010), these studies have addressed the issue of meteorological sensitivity: Seaman (2000) provided a comprehensive summary of the key meteorological issues most relevant for air quality modeling. Hanna and Yang (2001) evaluated the boundary layer outputs of several mesoscale meteorological models (e.g., MM5, RAMS, OMEGA), stressing meteorological variables used by air quality models (AQMs). Otte (2008) provides an example of a diagnostic study that demonstrates that assimilation of observations into the meteorological predictions can contribute to improved meteorological predictions, which, in turn, can lead to improved ozone predictions.

2.3 Quality of input data for Air quality models: meteorological fields and emission inventories

The quality of input data to AQMs is an important element in this context. As identified by Dennis et al. (2010): "Meteorological models have long been used to forecast weather, but AQM predictions are sensitive to a number of different meteorological variables that are not as critical to weather prediction. Evaluation of such models for the purpose of providing weather forecasting guidance may not be sufficient to assure their reliable use in air quality applications." **This points towards improving the communication between the AQ modeling community and the meteorological community for a better understanding of problems and needs.**

AQ model results are sensitive to emission inventory quality. **The assessment of the quality of that information and the uncertainty associated with the species provided by emission inventories is of paramount importance. While the main responsibility lies within the emission inventory community, communication should be established with the air quality community on the way uncertainty in emission inventories could be expressed. The major uncertainties are assumed to be related to biogenic VOCs, ammonia (NH_3), and dust/**fine particles.

AQ models are generally used in a "top-down" approach to evaluate emissions. Inverse modeling techniques have proven to be effective in many instances (e.g. Bergamaschi et al., 2000) however limitations in detailed information on concentrations of specific species make this task hard. For observationally-based methods such as receptor models, speciated observations are needed on shorter time scales than generally available in order to decipher the source signatures, and to distinguish between different source types. In many cases, the data are only available for limited time periods and specific locations.

For model evaluation purposes, as well as emissions evaluation, there is a need to collect monitoring data of VOC speciation and formaldehyde (in situ and remotely sensing). In both

NA and EU these data exist in scarce datasets mainly relating to episodes and ad hoc field campaigns. The lack of robust continuous data collection methods has been an obstacle to the growth in the measurement networks for VOCs. NH_3 is an important gas precursor for aerosol formation. Temporal variation in its emissions and issues such as soil compensation, as well as mean surface flux estimates are very uncertain. Additional NH_3 measurements as well as measurements at higher frequency are needed. Dust measurements need to discern between different types of sources and the amount of dust lifted high enough for transport out of the immediate source area. There are also key needs for PM chemical speciation, size distributions, and number data, as well as VOC/NOx, EC/NOx ratios. Comparison of air quality model results versus receptor model results will help evaluate the emissions inventories.

2.4 Diagnostic evaluation of chemical schemes

With respect to diagnostic evaluation of chemical schemes and mechanisms, fundamental work is still required. **Comparison of box models with a Master Chemical Mechanism (MCM) or possibly smog chambers, model responses to chemical schemes, pathway tracking, and radical closure experiments are all in need of investigation. From the experimental view, point data are needed on NOy speciation. Also needed is the development of diagnostic indicators like those used to diagnose the response of model predictions to emission changes.** For example, the indicator ratios of H_2O_2/HNO_3 (Sillman, 1995; Sillman et al., 1998; Kleinman, 1994; Kleinman et al., 1997), and O_3/NO_x (Tonnesen and Dennis, 2000 a,b; Arnold et al., 2003) are both response-surface probes that have been used to characterize how O_3 will change with NO_x and VOC levels in a given area. More recently, the potential for nitrate replacement and less reduction in total $PM_{2.5}$ than anticipated with SO_2 emission reductions has been studied using the Gas Ratio, which is a ratio of free ammonia to total nitrate (Ansari and Pandis, 1998; Pinder et al., 2009; Dennis et al., 2010) or an ionic mass balance (Blanchard et al., 2000). By comparing modeled results to observations from special field studies, these types of diagnostic probes help to extend diagnostic evaluation from assessment of predicted concentrations to evaluation of the model's ability to respond correctly to emission changes. **The use of specific case studies with a strong chemical signature could be of great help in this respect, such as:**

- **Mexico city used as a chemical reactor (Molina et al., 2008)**
- **Rapid formation of ozone in Houston, Texas**
- **Plume from megacities like Paris plume (problem addressed experimentally in the MEGAPOLI)**
- **Ohio power plants' NOx case (NOX SIP call) (Gilliland et al., 2008)**
- **ICARTT (Fehsenfeld et al., 2007)**

Data are required on nitrate formation that would allow establishment of the nitrate budget (fine/coarse, gas/particles); more accurate measurements of HNO_3, NH_3 as well as NH_4^+, NO_3^-, and $SO_4^=$. Furthermore, coarse and fine mode concentrations plus associated ions (sea salt [SS], dust). The functional relationship between NH_4NO_3 dissociation constant and temperature and relative humidity. **The latter could be obtained through the use of MARGA measurements in US and EU.**

Tracer measurements are required to differentiate contributors to OC like SOA/POA fraction, ^{14}C, levoglucosan and other organic PM tracers , and VOC. These can be available in Europe through the projects: CARBOSOL, EUCAARI, EMEP intensives, Paris-MEGAPOLI and North America through: North Carolina intensive (Kleindienst et al., 2007), Texas studies (Daum et al., 2004; Gilman et al., 2009) and Canadian studies in Alberta, great lakes or great plains.

Boundary layer processes are crucial elements with remaining uncertainties that require additional diagnostic attention. Examples include: planetary boundary layer (**PBL) height and its temporal evolution, turbulent transport and dispersion with specific reference to entrainment and detrainment process at boundary layer top, nocturnal transport aloft in the residual layer in regional-scale AQMs and verification of the correct transport of passive tracers like potential temperature through the boundary layer, and evaluation of dry deposition and wet deposition with ambient data**. The use of high resolution modeling techniques such as LES (Large Eddy Simulation) has produced important advancements in boundary-layer meteorology from which simple parameterizations should be considered. In general, upper air measurements of meteorological data should be exploited whenever available.

Diagnostic evaluation should not neglect numerical schemes used in models as a possible source of errors. Standard tests exist to evaluate the performance of a numerical model or one of its components. Mass conservation should be considered among the most important of those tests which in turn should involve operator splitting, chemical solvers and advection schemes as possible sources of numerical error.

Datasets readily available for diagnostic evaluation include:

- EMEP intensives 2006-2009
- MEGAPOLI Paris 2009
- CARBOSOL 2003/4
- EUCAARI Cabauw 2008
- ICARTT/INTEX 2004 (Fehsenfeld et al., 2007)
- TEXAQS 2006 (Gilman et al., 2009)
- Central California Air Quality Studies 2000; this major study (>$30M) included both the Central California Ozone Study (CCOS) and the

California Regional $PM_{2.5}/PM_{10}$ Air Quality Study (CRPAQS) for a full year of ozone and PM monitoring with several intensive field studies (CARB, 2009)
- North Carolina 2003 studies (Kleindienst et al., 2007)
- Alberta 2005

3. Dynamic evaluation

3.1 Introduction

Dynamic evaluation focuses on the capacity of AQMs to simulate emission changes resulting from specific policy implementations, or changing social/industrial contexts. The overall intent of this evaluation practice is to generate confidence in the models as tools for policy assessment and implementation. As described in Dennis et al. (2010): "This method is used in addition to traditional indicator ratios that focus on a model's potential response to a change in emissions through chemical relationships (e.g., O_3/NO_y)." We will refer to Dennis et al. (2010) for examples on past applications of this practice. Dynamic evaluation is considered a good example of policy-relevant science.

3.2 Concurrent changes in emissions and meteorology.

In reality, emissions changes as a result of growth or control over time occur simultaneously with meteorological changes. However we wish to determine the impact of the emissions changes on ambient concentrations in the absence of meteorological changes. This is important for assessing the efficacy of emissions control programs for air quality management. Since we cannot conduct such a real-world experiment where the meteorology stays constant, we can explore this concept through air quality modelling, where we can change emissions in the presence of the same meteorology. We can also explore the effect of changing meteorology in the presence of the same emissions, to ascertain the incremental impacts of these forcings on ambient air quality. Evaluating these model responses with real world data requires careful diagnostic analysis, including normalizing for meteorological trends in air quality data. These procedures are aspects of dynamical evaluation.

3.3 Dataset and case studies suitable for dynamic evaluation

There is a need to select or create datasets that would be useful for carrying out dynamic evaluation. Such data sets should be scientifically challenging, policy relevant, and extending over long time periods with comparable high-quality measurements. Emission changes should be larger than 15-20%; variability should be discernible in the observations; the variability should be regional or local scale; and it should be possible to define indicators based on observations.

The time and space scales at which dynamic evaluation would produce results are:

- Episodes: A few days. This concerns meteorological and emission situations, which allow the determination of AQ based on a statistical approach (e.g. number of exceedances). Here, the focus will be on extreme conditions, like situations with extreme emissions like forest fires, or black-out days from power faliures, or situations with extreme meteorological conditions like stagnant conditions with very low wind speeds and very stable boundary layers.
- Long term: Extended periods (years). Here the focus is on an evaluation which covers all meteorological conditions as well as all emissions over a large region, and on an hour-by- hour basis over at least a year, preferably longer. The evaluation should then focus on the determination of the situations in which the model performs well, and cases in which the model performs poorly.
- The time as well as the spatial scale over which the evaluation will take place will depend on whether the situation is an episodic (extreme) event, or whether the evaluation is focussed on an hour-by-hour full analysis over a whole year. In the latter case, the evaluation needs to focus on AQ standards that need to take their statistical definition into account, i.e. annual mean or higher percentiles.

Examples of episodes and situations that could be analysed in the context of dynamic evaluation and should potentially contain a strong emission signature are:

- **Traffic emissions:**
  - **August in Italy: different emissions due to tourist travel patterns**
  - **Public transport days, if visible in the data**
  - **Lead: emission reduction: simulation, but only coarse measurements.**
- **Aircraft emissions:**
  - **9.11: for one week no aircrafts in US – to analyse if impact on surface level concentrations.**
- **SO2:**
  - **Step change in SO2 emission from power plants (about 1995, US).**
- **NOx:**
  - **Step change in NOx emissions in summer 2004 (power plants in east US)**
- **SO2/NOx:**

  - **Significant emission changes (e.g. ~1990, Germany) – power plants / industry / traffic**
- **Multiple sources and tracers:**
  - **Weekday/weekend effect: NOx/VOC test case for US and Europe**
  - **Current (2008-2009) global emission reduction phase.**
  - **Emission reduction at Olympics. Mainly meteorologically driven.**
  - **US: black-out.**
- **local scale cases (emission / inversion phenomena; air trapped in city base → idea on emission; change in meteorology releases this trapped air mass for larger scale dispersion).**

Alternatively, these are potential cases for studying (mainly) impacts of meteorology changes (climate)

- It might be possible to group days on the basis of similar meteorological conditions (e.g., conditional sampling, statistical clustering methods). Possible case: N-W-Europe in year 2003 (very dry and hot summer) compared to year 2002 (very wet summer); eastern US in year 2004 (very wet summer) compared to year 2005 (dry summer)

As for other evaluation practices, it is always of paramount importance to provide uncertainty estimates on emissions.

3.4 Metrics for dynamic evaluation

Dynamic evaluation has a wide range of applications from forecast and climate impacts, to emission strategies and air quality and climate interactions. However for each of the specific fields of application, metrics need to be developed. Besides the traditional measures, we should define indicators which describe, e.g., concentration relationships, for measurements and model results. Such indicators should primarily be based on measurements. The indicators should be characteristic of emission data and/or meteorology and should serve as instruments for the identification of tendencies. **Specific research activities could be developed to devise these indicators in particular for the regional scale.** Conditional quantile plots can help to determine relations of meteorology/concentration data (Ries and Schlünzen, 2009).

Table 1. Key variables assumed to play a role in the dynamical evaluation of an AQ model and their impact on concentrations.

Meteorological Parameter	**Where**	**Relevance for concentration forecasts**	**Influence on**	**Averaging**
Boundary layer height and its diurnal variation		Very relevant	All concentration values	Per hour
Wind direction	At surface	Very relevant	chemical reactions	Overall, per hour
Wind direction	Upper air	Very relevant	Pollutant mixing, dispersion direction,	Overall, per hour
Atmospheric stability	At surface	Very relevant	All concentration values	Per hour
Atmospheric stability	Upper air	Very relevant	All concentration values	Per hour
Radiation	Upper air	Very relevant	Photochemical reactions	Overall, per hour
Precipitation	At Surface	Very relevant	Deposition	Overall
Cloud cover	Upper air	Relevant	Photochemical reactions	Overall

4. Probabilistic Evaluation of Air Pollution Models:

4.1 Introduction

Air pollution models are functionally deterministic in the sense that a fixed set of input, initial- and boundary conditions will always result in the same output. In reality, the output of an air pollution model must be treated as non-deterministic for the following complex set of linked reasons.

Model output is always subject to uncertainty due to: uncertainties in amounts, chemical speciation, location and timing of emissions; uncertainties in initial and boundary values of meteorological and chemical conditions; uncertainties due to parameterizations in meteorological and chemistry models; and uncertainties due to discretization and numerical solvers in meteorological and chemistry models. This means that the (apparently deterministic) output of an air pollution model is in reality just a single sample value from a stochastic variable that has an underlying probability distribution.

While nobody would doubt that measurements of air pollution are subject to uncertainty, it is important to realize that the processes underlying uncertainty in measurements (instrument error, malfunction, bias; environmental fluctuation) are quite different than those underlying model uncertainty. The two quantities (model output and measurement) are thus likely to have different underlying probability distributions, making a direct comparison of model output versus measurement (as is done in a myriad of published scatter plots of observed versus modeled quantities) of questionable statistical and logical utility.

A further difficulty with direct comparison of model output and measurements arises because model output is always a volume average, while measurements are usually point values. In the statistical literature, this is termed ”change of support”, and is a matter that deserves careful treatment. In far too many model evaluation studies, this difficulty is simply ignored. A related difficulty is that while measurements are point values, multi-scale advective processes influence measurements with the result that sub-grid scale effects due to land-use variability in the instrument’s near-field will always be present, and will be expressed in ways that are wind-direction dependent. These effects are not present in model output.

The recent adoption in air pollution modeling of ensemble methods pioneered in weather forecasting has introduced an explicitly probabilistic approach into the field. Potempski and Galmarini (2009) argues that ensemble modeling of air pollution has developed in an ad-hoc fashion and has accepted practices but no firm theoretical basis. Clearly this theoretical basis must be couched in probabilistic terms.

These considerations lead directly to the conclusion that probabilistic (or at least sophisticated statistical) approaches to air pollution model evaluation are called for. The following sections outline a few major technical approaches and strategic considerations that should be thoroughly addressed.

4.2 Estimating model uncertainty

As indicated in the introduction, a major difficulty lies in determining the uncertainty underlying a single, apparently deterministic model output. **There already exists a number of technical tools, including data assimilation that give direct or indirect ways to use a deterministic model in a probabilistic framework. These methods must be used, tested, and developed for use in model evaluation.**

The Bayesian paradigm (Savage 1954) provides the natural framework for this, since all uncertainty is represented by probabilities. In particular all the fixed but "known unknowns" have probability distributions. In absence of any information, these are called prior distributions but the "knowns", for example, measured values of pollutant concentrations alter these distributions in accordance with Bayes' rule to yield posterior probability distributions. This paradigm has a fundamental role in modern inductive inference since it embraces both "aleatory uncertainty" (that due to chance phenomena such as measurement error) as well as "epistemic uncertainty" (due to lack of knowledge).

However, characterizing those distributions in complex dynamic systems is challenging owing to their large numbers of unknowns. One useful approach uses the Bayesian hierarchical model (Gelman et al. 2003) in which all the unknowns are arranged in a sequence of clusters, the probability distribution for each cluster being conditional on (purely hypothetical) knowledge of all the previous clusters so that Bayes' rule may be applied in steps. This allows uncertainty to be characterized in a structured fashion, one cluster at a time, and simplifies the stochastic modeler's task. Nevertheless the resulting distribution can still be quite intractable.

Ultimately the great success of hierarchical models can be ascribed to the Monte Carlo Markov Chain (MCMC) method that enables the repeated drawing of random values from the posterior distribution (Gelfand and Smith 1990), the (Gibbs) algorithm being tailored to the hierarchial structure of these models. The large sample of values obtained in this way can be used to characterize even the most intractable of hierarchial distribution models.

Hierarchical models for dynamic systems represent the evolution of states of the system over time and have three components. The first, which is conditional on all the

unknowns in its two successors, represents measurements made on responses generated by that system whose states are not usually observable and hence unknown. Such models can reflect such things such as measurement error. The second component is the process model. It represents the deterministic components with their uncertainties due to lack of knowledge and unknown parameters. The third component involves a prior distribution for those unknown parameters.

Process models like those in air pollution modeling may be expressed in terms of differential equations (DEs), solved numerically via their difference equations. By adding stochastic perturbations to represent errors of approximation and lack of knowledge, the latter become a Kalman filter (Christakos and Raghu 1996; Kalenderski 2009). Finally the third component turns this filter into a dynamic state space model for which abundant theory exists (West and Harrison 1997). Nevertheless, running an MCMC for the posterior distribution of the vector of unknown states, conditional on the measurements can be impossible owing to excessive computational requirements. That has led to various simplifications and computational strategies. One, the ensemble Kalman filter (EnKF; Evensen, 1994) draws at time t, a small sample of state space vectors from the predictive posterior distribution given the data up to and including time t-1, which can be used to generate estimates of some covariance parameters needed for the filter that takes us to time t. Taking that "ensemble" of forecasts along with the parameter estimates as given, the filter and data for time t can be used to generate a new synthetic sample of state vectors for time t, this time drawn from the predictive posterior distribution conditional on data up to and including time t, thus providing a sample from the posterior distribution whose properties can then be inferred. Particle filtering (Doucet et al. 2001) provides another approach in the same spirit, but this time using sequential importance sampling to draw from a simplification of the state predictive distribution and ratio-adjusting this one so that in effect it is drawn from the distribution of interest. The deterministic model is now combined with measurements, and the resulting synergies can be exploited to better achieve the intended purpose of the deterministic model. At the same time, the uncertainty about the combined model outputs can be provided in measurement or model space as appropriate.

The methods just described works when a small number of differential equations (DEs) are involved in the process model but this approach is not practical for chemical transport models (CTMs) since they involve a large number of DEs. Another Bayesian hierarchical approach has been developed for this situation at least when only a spatial process is involved. This approach, Bayesian melding as it is called (Fuentes and Raftery 2005 for SO_2; Zhong et al 2007a for O_3), recognizes the difference between the meso-scale of the CTM and the micro-scale of the measurement. It introduces a latent process (the "truth") as the state of the system which generates measurements through the classical measurement error model. This process is represented in the computer model. That output is regarded as a linear transformation of the integral of the truth over the meso-scale grid cell plus a stochastic perturbation for the resulting discrepancy between the real process and the one in the computer model. The combined model can then be used for the intended purpose, again with uncertainties characterized. Note in particular, that this approach yields the recalibration scheme needed to go from the model to measurement space and vice-versa. Software is now available online (http://enviro.stat.ubc.ca).

A second alternative that, like the melding approach, incorporates the computer model (the simulator) output directly and is concerned with spatial processes alone, assumes the truth at any given point in space is "regressed" on a linear transformation of the computer output at that point plus a stochastic perturbation (Kennedy and O'Hagan 2001). The result is a method that can recalibrate the computer output to yield inferences in measurement space with regression parameters that recalibrate the simulators to move them from the meso- to micro-scale. The importance of this approach lies in the foundation it lays for the development of an "emulator" for the simulator. That tool, which is now available online along with tutorials is based on a Gaussian field approximation to the simulator (http://mucm.aston.ac.uk/MUCM/MUCMToolkit/index.php?page=MetaHomePage.html). It runs quickly, unlike the typical simulator, and can be used to assess properties of the simulator based on very large samples from the emulator

An alternative to the previous approach, which is also a Bayesian hierarchical model based on a regression approach, has been constructed to include time (Zhong et al 2007b). This model runs quickly, and like the emulator, calibrates the model output to align it with measurement space through the regression coefficients. In measurement space, it performs better than melding (no doubt because it can borrow strength over time).

The approaches listed above are generally based on analysis of model output as a space-time process, and through this analysis, can be used to provide estimates of model uncertainty.

A completely different approach to model uncertainty lies in sensitivity analysis methods already discussed in section 2.2 (Saltelli et al. 2008).

One very particular aspect of model uncertainty lies in assessing the uncertainty introduced into the model by the myriad of simplifications (including parameterizations, linearization and chemical simplification schemes) that are incorporated into the model system. These simplifications all propagate through a highly nonlinear modeling system to produce uncertainties in model output. These uncertainties are of enormous importance because they will be of primary utility in developing model improvements. **While brute force[3] methods do exist for quantifying the nature and magnitude of these uncertainties, a**

[3] By "brute force" we mean the estimation of uncertainties by performing multiple runs of the entire modeling system with small, but realistic changes of the independent variables in question.

comprehensive, theoretically based and computationally parsimonius framework remains to be defined. Closely related to model uncertainties arising out of simplifications are model uncertainties arising from uncertainties in input data that are propagated through the model and appear as uncertainties in model output. This class of uncertainties will presumably be treatable in the same framework as uncertainties arising out of model simplifications.

4.3 Uncertainties from Ensemble Modeling

The combination of several model results in what is normally defined as ensemble modelling has proven able to produce an improvement in the model results when compared with measurements and with the individual model ensemble members. While results are improving in all studies published so far regardless of the way in which the ensemble has been constituted and what the generic expression ensemble means, it is felt that the whole ensemble modelling technique has more the character of a practice than a theoretical framework as pointed out by Potempski and Galmarini (2009). The use of ensemble and in particular multi-model ensembles is felt to be an excellent opportunity for a collaborative model inter-comparison exercise, with the added value of improving individual model performance. This goes beyond an organized scientific and strategic selection of model results that are supposed to maximize the quality of the final product. It is acknowledged that multi-model ensemble is a good practice to produce policy consensus around the results originating from multiple sources and it therefore should have a prominent role as one of the techniques to be used in the future. Given the level of attention gathered around it (not only in regional air quality modelling also on global scale and climate modelling), **it becomes more and more urgent that research efforts are dedicated to a more rigorous theoretical framing of the discipline. There is a series of fundamental questions that need to be urgently addressed, namely: in what way should an ensemble of model results be assembled? what is the minimum number of model results necessary to define the group of an ensemble? how can we get around the model dependence issue? can we diagnose a-priori the ensemble properties based on model characteristics? how can we guarantee a priori maximum coverage of the measurements pdf by the ensemble and complement for missing portions?** Given the fact that ensembles are an operational practice in weather forecasting and in that context theories have been developed and tested, **it is suggested to look at weather forecasting as a possible source of inspiration for the development of the air quality ensemble theory. Within that context, parameters and indicators can be found to be instrumental to air quality applications.**

4.4 Comparing uncertain models and uncertain observations

As outlined in the introduction, **both model output and observations are subject to uncertainties, but those uncertainties are likely to have different statistical properties. This makes direct comparison of model output and observations a practise fraught with difficulties. In order to perform a statistically valid comparison, the two quantities and their (differing) probability distributions must be reconciled. It should be obvious that simple least squares regression analysis in which observations are assumed (by unspoken convention) to be error-free is generally inappropriate. We, thus, recommend a concerted attempt to find approaches to the reconciliation of different probability distribution between model results and observations.**

Comparisons between model and measurement must take into account the fact that model and measurement uncertainties are driven by different processes with different underlying PDFs. Among the few simple approaches that do exist are generalized least squares regression and maximum likelihood estimation. These and related approaches should be explored.

Implicit in the foregoing is the presumption that the underlying measurement PDF is known. This is generally not the case, **so we recommend that specific campaigns be established to explore the underlying PDF of air quality measurements? A major topic to be addressed in such field studies (in a quantitative way) is the representativeness of observations, and how many measurement points are needed to represent a volume?**

One interesting approach to this problem is to conceive of three different spaces: Exposure Space; Model Space and Measurement Space. The problem can then be approached as an exercise in mapping between spaces so that comparisons are only performed in one space. Some techniques for achieving this mapping do exist. They are:

1) **Model Space → Observation Space: Bayesian melding, variance modelling, other subgrid scale modelling methods, data massaging based on knowledge, other data assimilation techniques**
2) **Observation Space → Model Space: multi-point averaging, geostatistical techniques, entropy based network design (Le and Zidek, 2007)**
3) **Model space → Exposure Space: methods have yet to be developed**

We note that there exist a number of potentially applicable tools that are defined and employed in the meteorological context, for example: Brier scores; Talagrand Diagrams; Spread versus Skill; Relative Operating Characteristics Diagram; attribution methods in climate change research.

4.5 Representing and communicating model uncertainty

Much of the motivation for enhanced model evaluation approaches stems from the need to use model output to provide guidance in the policy-making realm. Communication within the scientific realm is a well-established practise, and makes use of particular media and language. By contrast, communication between scientific and policy realms is difficult because the two realms use different media and languages. If model results, as evaluated by scientists are to be transferred to the policy realm, it is clear that particular attention must be paid to the means and languages of communication. **For this reason, we recommend that scientists work with communications specialists, journalists and psychologists in order to develop communications strategies that will be effective in the policy realm. This work must include the development of methods for display and presentation of model output, including animations, spaghetti plots and other devices. Of particular difficulty will be the communication of the linked space-time nature of air pollution fields, and the conception and use of probabilities in making environmental decissions.**

While estimation of uncertainties in model output is a relatively unexplored topic, estimation of uncertainties in measurements is a rich and detailed subject. Some outstanding issues that require attention are:

1) Comparison is generally made between a relatively between small number of surface point measurements and a selected set of modeled values. A vast majority of modeled values are not subjected to evaluation, and are assumed to behave in the same way as those that are evaluated. This assumption must be examined.
2) We should ignore stations that are not regionally representative for certain species. For example, urban stations are likely representative for SO_4 but not for most other species. This idea should be investigated in detail.
3) Representativeness of measurements is generally assumed to be pollutant specific. This assumption must be examined.
4) It may be possible that pollutant values in a sub-region are coherent, thus allowing aggregation of the information (model output and measurements). This possibility must be examined. This is in effect exploitation of autocorrelation of measurements as a possible way of evaluating representativeness. This idea must be further examined.

5. Conclusions and Summary of Recommendations

This guidance document has been prepared within the context of the Air Quality Model Evaluation International Initiative (AQMEII), whose overall purpose is to inspire an advance in air quality model evaluation approaches and practices. This will be achieved by initiating an international collaborative research effort designed to explore and elaborate on the four modes of model evaluation outlined by Dennis et al. (2010), namely, Operational, Diagnostic, Dynamic and Probabilistic. The elaboration presents specific guidance on technical approaches appropriate for each mode, and indicates research directions designed to initiate further international investigation into the area of model evaluation for regional photochemical models being used in a policy development context.

The table below gives a summary of what are considered to be areas most urgently in need of investigation. It is anticipated that participants in the AQMEII collaboration will address most, if not all of these research questions.

Operational	A major problem facing large-domain operational air quality models is the artificial variation of land use parameters that are used in statistical estimates of emissions because of different national land use classification schemes. These differences must be reconciled so as to reduce discontinuities in emissions inventories at national borders.
	Development of further emissions evaluation techniques, beyond those listed is called for.
Diagnostic	Efforts and collaborations should be established between atmospheric modeling communities and statistical/sensitivity analysis communities for the development of ad-hoc practices and tools for air quality applications of sensitivity analysis.
	towards improving the communication between AQ modeling community and the meteorological community for a better understanding of problems and needs
	The assessment of the quality of emission inventory information and the uncertainty associate to the species provided by emission inventories are felt as of paramount importance but that should find its reason of existence within the emission inventory community. However communication should be established with the air quality community on the way uncertainty in emission inventories could be expressed. The major uncertainties assumed to be related to biogenic VOCs, ammonia, dust and anthropogenic emission inventories.
	comparison of box models with MCM or possibly smoke chambers, address model responses to schemes, pathway tracking, and radical closure experiments. From the experimental view point data are needed on the NOy speciation and the development of diagnostic indicators like those used to diagnose the potential response of model predictions to emission changes
	PBL height and its time evolution, turbulent transport and dispersion with specific reference to entrainment and detrainment process at boundary layer top. Verification on the correct transport of passive tracers like potential temperature through the boundary layer. Evaluation of dry deposition and wet deposition with data
Dynamic	There is a need to select or create datasets that would be useful for carrying out dynamic evaluation with the characteristic of being scientifically challenging, policy relevant, extending to long period of times with comparable quality standard in the measurements.
	Examples of episodes and situations that could be analyzed in the context of dynamic evaluation and that should potentially contain a strong emission signature
	Specific research activities could be finalized at devising these indicators in particular for the regional scale.
Probabilistic	While brute force methods do exist for quantifying the nature and magnitude of model uncertainties, a comprehensive, theoretically based and computationally parsimonious framework remains to be defined.
	It becomes more and more urgent that research effort are dedicate to a more rigorous theoretical framing of the discipline. There are a series of fundamental questions that need to be urgently addressed, namely: in what way should an ensemble of model results be assembled, what is the minimum number of model results necessary to define the group an ensemble, how to get around the model dependence issue, can we diagnose a-priori the ensemble properties based on model characteristics, how can we guaranty a priori maximum coverage of the measurements pdf by the ensemble and complement for missing portions.
	both model output and observations are subject to uncertainties, but those uncertainties are likely to have different statistical properties. This makes direct comparison of model output and observations a practise fraught with difficulties. In order to perform a statistically valid comparison, the two quantities and their (differing) probability distributions must be reconciled. It should be obvious that simple least squares regression analysis in which observations are assumed (by unspoken convention) to be error free is generally inappropriate. We thus recommend a concerted attempt to find approaches to the reconciliation of different probability distribution between model results and observations.
	We recommend that specific campaigns be established to explore the underlying PDF of air quality measurements? A major topic to be addressed in such field studies (in a quantitative way) is the representativeness of observations, and how many measurement points are needed to represent a volume?
	One interesting approach to this problem is to conceive of three different spaces: Exposure Space; Model Space and Measurement Space. The problem can then be approached as an exercise in mapping between spaces so that comparisons are only performed in one space.
	This work must include the development of methods for display and presentation of model output, including animations, spaghetti plots and other devices. Of particular difficulty will be the communication of the linked space-time nature of air pollution fields, and the conception of probabilities.

Contributors to the workshop

AQMEII co-Chairs

- ST Rao, US-EPA & S. Galmarini, JRC/IES

Key note speakers and presentation's titles

- ST Rao (US EPA): *State of the art of air quality model evaluation practices in the US: Regional-Scale Air Quality Model Evaluation: Establishing Model's Credibility*
- P. Builtjes (TNO): *State of the art of air quality model evaluation practices in EUROPE*
- C. Seigneur (CEREA, Ecole des Ponts ParisTech/EDF R&D Universite' Paris-Est): *Diagnostic evaluation*
- L. Rouil (INERIS): *Operational evaluation of air quality models*
- C. Hogrefe (State University of New York at Albany): *Dynamic evaluation*
- J. Zidek (UBC): *Probabilistic modeling and assessment*
- A. Saltelli (EC-JRC): *Sensitivity analysis*
- S. Galmarini (EC-JRC): *Ensemble dispersion modeling: the parable of the blind leading the blinds?*
- G. Foley (US EPA): *Air quality modeling in support to US policy*
- A. Kobe (EC- DGENV): *Air quality modeling in support to EU policy*

Break out session Chairs

- P. Builtjes, TNO, Netherland (Dynamic evaluation)
- M. Beekmann, LISA / IPSL, France. (Diagnostic evaluation)
- M. Moran, Environment Canada, Canada (Operational evaluation)
- D. Steyn, University of British Columbia, Canada (Probabilistic evaluation)

Rapporteurs

- K. Heinke Schlünzen, University of Hamburg, Germany (Dynamic evaluation)
- M. Schaap, TNO, Netherland, (Diagnostic evaluation)
- J. Brandt, Univ. of Aarhus, Denmark (Operational evaluation)
- R. Vautard, CEA/CNRS/UVSQ, France, (Probabilistic evaluation)

Provocateurs

- C. Hogrefe, New York State Dep. of Environ. Cons., USA (Dynamic evaluation)
- K. Schere, US-EPA, USA, (Diagnostic evaluation)
- G. Kallos, University of Athens, Greece, (Operational evaluation)
- S. Galmarini, JRC/IES (Probabilistic evaluation)

Participants in break out sessions

- B. Bessagnet, INERIS, France
- N. Brown, Standford University, USA
- P. Cassanelli, DEFRA, UK
- I. Coll, LISA / IPSL, France
- R. Derwent, Consultant, UK
- H. Elbern, University of Koeln, Germany
- G. Foley, US-EPA, USA
- J. Jansen, Southern Company, USA
- N. Kumar, EPRI, USA
- L. Lobocki, University of Warsaw, Poland
- D. McCabe, US-EPA, USA
- A. Martilli, CIEMAT, Spain
- R. Mathur. US-EPA, USA
- V. Matthias, GKSS, Germany
- A.-I. Miranda, University of Aveiro, Portugal
- V.-H. Peuch, MeteoFrance, France
- T. Pierce, US-EPA, USA
- L. Robertson, SMHI, Sweden
- C. Seigneur, CEREA, France
- D. Simpson, EMEP, Norway
- M. Sofiev, FMI, Finland
- P. Suppan, FZK, Germany
- M. Tombrou, University of Athens, Greece
- A. Trukenmüller, UBA, Germany
- R. Wolke, Uiversity of Leipzig, Germany
- J. Zidek, University of British Columbia, Canada
- M. Antonelli, JRC/IES
- E. Gueorguieva, JRC/IES

References

Arnold, J.R., R.L. Dennis, and G.S. Tonnesson, 2003: Diagnostic evaluation of numerical air quality models with specialized ambient observations: Testing the Community Multiscale Air Quality modeling system (CMAQ) at selected SOS 95 ground sites, *Atmos. Environ.*, **37**, 1185-1198.

Ansari, A.S., S.N. Pandis, 1998: Response of inorganic PM to precursor concentrations. *Environmental Science & Technology,* **32**, 2706-2714.

Bengtsson, T, Snyder, C and Nychka, D (2003). Toward a nonlinear ensemble filter for high-dimensional systems. J Geophysical Research, v108 (D24), 8775, doi:10.1029/2002JD002900.

Bergamaschi, P.; Hein, R.; Heimann, M.; Crutzen, P.J., 2000: Inverse Modeling of the Global CO Cycle: 1. Inversion of CO Mixing Ratios", *Journal of Geophysical Research*, 105, D2, 1909-1927.

Berliner, LM, (2003). Physical-statistical modeling in geophysics. J Geophysical Research, vol. 108, no. d24, 8776, doi:10.1029/2002jd002865.

Blanchard, C.L., P.M. Roth, S.J. Tannenbaum, S.D. Ziman, J.H. Seinfeld (2000). The use of ambient measurements to identify which precursor species limit aerosol nitrate formation. *J. Air Waste Manage. Assoc*., **50**, 2073-2084.

CARB (2009). http://www.arb.ca.gov/airways/

Christakos, G and Raghu, VR, (1996). Dynamic stochastic estimation of physical variables. Math Geology, 28, 341-365

Cullen, A.C., and H.C. Frey (1999). *Probabilistic Techniques in Exposure Assessment.* Plenum Press: New York.

Daum, P. H., L. I. Kleinman, S. R. Springston, L. J. Nunnermacker, Y. N. Lee, J. Weinstein-Lloyd, J. Zheng, and C. M. Berkowitz (2004), Origin and properties of plumes of high ozone observed during the Texas 2000 Air Quality Study (TexAQS 2000), *J. Geophys. Res.*, *109*, D17306, doi:10.1029/2003JD004311.[

Dennis, R. , T. Fox, M. Fuentes, A. Gilliland, S. Hanna, C. Hogrefe, J. Irwin, S.T. Rao, R. Scheffe, K. Schere, D. Steyn, A. Venkatram (2010), A framework for evaluation of regional-scale numerical photochemical modeling systems . *Env. Fl. Mech.,* In Press.

Doucet, A, de Freitas, JFG and Gordon, NJ, Eds. (2001). Sequential Monte Carlo Methods in Practice. New York: Spinger-Verlag.

Evensen, G (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res. 99 , 10 143–10 162.

Fehsenfeld F.C. et al. (2007): International Consortium for Atmospheric Research on Transport and Transformation (ICARTT): North America to Europe – Overview of the 2004 summer field study. *J. Geophys. Res*., **111**, D23S01, doi: 10.1029/2006JD007829.

Fuentes M, and Raftery A (2005). Model evaluation and spatial interpolation by bayesian combination of observations with outputs from numerical models, Biometrics, 61, 36-45.

Gelfand, AE and Smith, AFM. Sampling-Based Approaches to Calculating Marginal Densities. *J. American Statistical Association*, 85:398-409, 1990

Gelman, A, Carlin, JB, Stern, HS and Rubin, DB (2003). Bayesian data analysis, Second Edition. London: Chapman and Hall/CRC Press.

Gilliland, A.B., Hogrefe, C., Pinder, R.W., Godowitch, J.M., Foley, K.L., and Rao, S.T., 2008. Dynamic evaluation of regional air quality models: Assessing changes in O3 stemming from changes in emissions and meteorology. *Atmos. Environ*., **42**, 5110-5123.

Gilman et al., 2009: Measurements of volatile organic compounds during the 2006 TexAQS/GoMACCS campaign: Industrial influences, regional characteristics, and diurnal dependencies of the OH reactivity. *J. Geophys. Res*., **114**, D00F06, doi: 10.1029/2008JD011525.

Hanna, S. R., Yang, R. 2001: Evaluation of mesoscale numerical models' simulations of near surface wind, Temperature Gradients and Mixing depth. *Journal of Applied Meteorology,* Vol, **40**, 1095-1194.

Kalenderski, S (2009). Stochastic modeling of space-time processes. An air pollution problem. PhD Dissertation, University of British Columbia 123 p.

Kennedy, MC and O'Hagan, A (2001). Bayesian calibration of computer models. J Roy Statist Soc, Series B, 63,. 425-464.

Kleindienst et al., 2007: Estimates of the contributions of biogenic and anthropogenic hydrocarbon precursors to secondary organic aerosol at a southeastern U.S. location. *Atmos. Environ.*, **47**, 8288-8300.

Kleinman, L.I., P. Daum, J.H. Lee, Y.-N. Lee, L. Nunnermacker, S. Springston, J. Weinstein-Lloyd, L. Newman, and S. Sillman, 1997. Dependence of ozone production on NO and hydrocarbons in the troposphere, *Geophysical Research Letters*, **24**, 2299-2302.

Kleinman, L.I., 1994. Low- and high-NO_x tropospheric photochemistry, *Journal of Geophysical Research*, **99**, 16,831-16,838.

Le, N.D. and Zidek, J.V., 2007: *Statistical Analysis of Environmental Space-Time Processes* Series: Springer Series in Statistics, Springer-Verlag, New York.

Molina, L. T., Madronich, S., Gaffney, J. S., and Singh, H. B.: Overview of MILAGRO/INTEX-B Campaign, IGACtivities Newsletter, 38, 2–15, 2008.

Otte, T.L., 2008: The impact of nudging in the meteorological model for retrospective air quality simulations. Part II: Evaluating collocated meteorological and air quality observations, *J. Applied Met. and Clim.*, 47, 1868-1887.

Pinder, R.W., R.C. Gilliam, K.W. Appel, S.L. Napelenok, K. Foley, and A.B. Gilliland, 2009: Efficient probabilistic estimates of surface ozone concentration using an ensemble of model configurations and direct sensitivity calculations, *Environmental Science & Technology*, 43, 2388-2393.

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

Ries, H., Schlünzen K.H. (2009): Evaluation of a mesoscale model with different parameterisations and resolutions for the bay of Valencia. *Monthly Weather Review*, doi: 10.1175/2009MWR2836.1\

Saltelli A. Tarantola S., Campolongo, F. and Ratto, M., 2004, Sensitivity Analysis in Practice. A Guide to Assessing Scientific Models, John Wiley & Sons publishers.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D. Saisana, M., Tarantola, S., 2008, *Global Sensitivity Analysis. The Primer*, John Wiley & Sons publishers.

Savage, L.J., 1954: The Foundations of Statistics (New York, Wiley, 1954)

Seaman, N.L., 2000: Meteorological Modeling for Air-Quality Assessments. *Atmospheric Environment*, 34, 2231-2259.

Savage, Leonard J. 1954. *The Foundations of Statistics*. New York, Wiley. West, M. and Harrison, P J (1997). Bayesian Forecasting and Dynamic Models, Second Edition. Springer: New York.

Sillman, S., D. He, M.R. Pippin, P.H. Daum, D.G. Imre, L.I Kleinman, J.H. Lee, and J. Weinstein-Lloyd, 1998: Model correlations for ozone, reactive nitrogen and peroxides for Nashville in comparison with measurements: Implications for O_3-NO_x-hydrocarbon chemistry, *Journal of Geophysical Research*, **103**, 22,629-22,644.

Sillman, S., 1995. The use of NO_y, H_2O_2, and HNO_3 as indicators for ozone-NO_x-hydrocarbon sensitivity in urban locations, *Journal of Geophysical Research*, **100**, 14175-14188.

Tonnesen, G.S. and R.L. Dennis, 2000a: Analysis of Radical Propagation Efficiency to Assess Ozone Sensitivity to Hydrocarbons and NO_x. Part 1: Local Indicators of Instantaneous Odd Oxygen Production Sensitivity, *Journal of Geophysical Research*, **105**, 9213-9225.

Tonnesen, G.S. and R.L. Dennis, 2000b: Analysis of Radical Propagation Efficiency to Assess Ozone Sensitivity to Hydrocarbons and NO_x. Part 2: Long-Lived Species as Indicators of Ozone Concentration Sensitivity, *Journal of Geophysical Research*, **105**, 9227-9241.

West, M. and P.J. Harrison, 1997: *Bayesian Forecasting and Dynamic Models.* Springer-Verlag, New York.

Wikle, CK, Berliner, M and Cressie, N (1998), Hierarchical Bayesian space-time models. J. Environ. Ecol. Statist.*, 5, 117–154.*

Zhong L, Le, ND and Zidek, JV (2007a), Calibrating Deterministic Modeling Output With Application To Ozone Fields. TR 232, Department of Statistics, University of British Columbia.

Zhong L, Le, ND and Zidek, JV (2007b) An Appraisal of Bayesian Melding for Physical-Statistical Modeling. TR 233, Department of Statistics, University of British Columbia

DISCLAIMER
The views expressed here are those of the authors and do not necessarily reflect the views and policies of the European Commission, U.S. Environmental Protection Agency, and Environment Canada.

European Commission

EUR 24245 EN – Joint Research Centre – Institute for Environment and Sustainability
Title: **Advancing the evaluation of regional-scale air quality models**
Author(s): S. Galmarini, D. Steyn, K. Schere, M. Moran
Luxembourg: Office for Official Publications of the European Communities
2010 – 31 pp. – 29 x 20.5 cm
EUR – Scientific and Technical Research series – ISSN 1018-5593
ISBN 978-92-79-15007-4
DOI 10.2788/61186

Abstract
This document is the culminating product of a series of ad-hoc workshops led by a group of European and North American air quality scientists interested in instigating a significant advance in the way regional-scale air quality modelling systems are evaluated. The initiative arising out of these workshops has come to be known as the Air Quality Model Evaluation International Initiative (AQMEII). The motivation for the workshops was a growing sense that current model evaluation practices have stagnated, and as a result are not serving research, operational and policy sectors as well as they should. While technical (often statistical) approaches have advanced in the past two decades, it was felt that present approaches are too often uncritically applied, without due consideration of the foundations upon which the techniques are based. This document contains a set of research recommendations for advancing model evaluation. These recommendations are based on ideas generated during workshop presentations, group discussions and informal discussions that occurred at and around the workshops.

How to obtain EU publications

Our priced publications are available from EU Bookshop (http://bookshop.europa.eu), where you can place an order with the sales agent of your choice.

The Publications Office has a worldwide network of sales agents. You can obtain their contact details by sending a fax to (352) 29 29-42758.

The mission of the JRC is to provide customer-driven scientific and technical support for the conception, development, implementation and monitoring of EU policies. As a service of the European Commission, the JRC functions as a reference centre of science and technology for the Union. Close to the policy-making process, it serves the common interest of the Member States, while being independent of special interests, whether private or national.

LB-NA-24245-EN-C

EUROPEAN COMMISSION

Publications Office
Publications.europa.eu

ISBN 978-92-79-15007-4

9 789279 150074