Appraisal project

Air Pollution Policies
foR Assessment
of Integrated Strategies
At regional and Local scales

Grant Agreement number 303895

WP 2 Review and gaps identification in AQ and HA methodologies at regional and local scale
Aristotle University of Thessaloniki
D2.5 Uncertainty and robustness

Reference: APPRAISAL / AUTH / WP 2 / D2.5 / VERSION 1.1

Category: Coordination

Author(s): John Douros (AUTH), Lia Fragkou (AUTH)

Contributions from: Philippe Thunis (JRC), Claudio Belis (JRC), Enrico Pisoni (UNIBS), Claudio Carnevale (UNIBS), Marialuisa Volta (UNIBS), Zbigniew Nahorski (SRI)

Verification: Nadège Blond (CNRS), Alain Clappier (UDS), Jean-Luc Ponche (UDS).

Date: 1/5/2013

Status: Version 1.2

Availability: Public

Summary

In the Deliverable D2.5 on “Uncertainty and Robustness” current state of the art approaches in model validation and uncertainty estimation are reviewed and their limitations are briefly described. The focus of the report is on model use for regulatory purposes and therefore, the different uncertainty approaches in Air Quality Assessment, Health Impact Assessment and Integrated Assessment Modelling are considered, in view of the EU legislation requirements. Information for this review was derived from published scientific papers and from the answers received in response to the questionnaire distributed within the framework of APPRAISAL activities. Model quality assessment and evaluation methods are examined separately for model use in relation to Air Quality Planning and for model use in relation to other purposes, e.g. Air Quality Assessment or research projects.

Version History

Version	Status	Date	Author(s)
0.1	First Draft	15/4/2013	John Douros (AUTH), Lia Fragkou (AUTH) *Contributions from*: Philippe Thunis (JRC), Claudio Belis (JRC), Enrico Pisoni (UNIBS), Claudio Carnevale (UNIBS), Marialuisa Volta (UNIBS)
0.2	Second Draft	29/3/2013	John Douros (AUTH), Lia Fragkou (AUTH) *Contributions from*: Philippe Thunis (JRC), Claudio Belis (JRC), Enrico Pisoni (UNIBS), Claudio Carnevale (UNIBS), Marialuisa Volta (UNIBS) Nadège Blond (CNRS), Alain Clappier, Jean-Luc Ponche.
1.0	Final	12/4/2013	John Douros (AUTH), Lia Fragkou (AUTH)

Contents

1 INTRODUCTION

Air quality model evaluation and estimation of model uncertainty have received increasing interest from local authorities and decision makers (Borrego et al., 2003) since the first model applications for air quality management purposes. As the role of modelling in understanding the influence of physical and chemical processes on the dispersion and transformation of pollutants is increasingly being recognised, the current European Directive 2008/50/EC on ambient air quality and cleaner air for Europe (AQD) encourages the use of air quality modelling, in combination with monitoring, as scientifically relevant tools for a range of policy applications. Models may be used to assess and predict exceedances and high-pollution areas, to identify the main polluting sources, to develop air quality plans and mitigation strategies and to perform risk assessment in the case of accidental atmospheric releases. As policy makers consult models for strategic decisions with health and economical consequences, it is important that their results are quantifiably accurate, precise and realistic. Model results are subject to limitations and uncertainties and, therefore, model performance evaluation is necessary in order to use model results with confidence for policy purposes. Thus, the need to incorporate uncertainty estimation in air quality modelling is also recognised by policy makers and is required by the AQD. Quantification of modelling precision is reflected in the modelling quality objectives described in Annex I of the AQD, which are given as a relative uncertainty (%). However, no particular methodology for estimating uncertainty is prescribed in the AQD and the wording of the text related to uncertainty is ambiguous.

It is worth noting that there has been a long standing ambiguity as to the exact meaning of the term “uncertainty”. In the literature, the term has been associated both with the evaluation process as well as to represent the stochastic character of natural variables inside air quality models. The two are of course closely linked, but the methods used for the quantification and study of each of the two can differ substantially. This deliverable, as well as the questionnaire that is formulated and analysed here, attempts to tackle both aspects of uncertainty by inquiring both about the evaluation process and the uncertainty quantification and propagation methodologies. The latter could, in lack of a better term, be named “indefiniteness”. The current legislative framework associates uncertainty principally with the model evaluation process while the propagation of uncertainties remains largely a scientific endeavour, focused mostly on scenario analyses, issues of model development and better understanding of the implementation in models of various physical/chemical processes. An attempt to highlight the differences between the two, is presented in table 1.

Table 1: The notion of uncertainty

Term	Time	Assessment type	Methodology	Quantitative Indices	Legislative provision
“Uncertainty”	Past / Present	Past / Current state assessment	Evaluation / Validation	Statistical indices (error, bias, RDE)	Yes (2008/50/EC)
“Indefiniteness”	Future	Scenario analyses	Model propagation	Confidence intervals, error bars	No

In view of promoting a harmonised and standardised approach for air quality model use for regulatory purposes, several scientific attempts have been recently made to develop a methodology for the quantification of uncertainty in model results. These attempts mainly focus on the identification of uncertainty sources in air quality modelling, on the setting of uncertainty indices and on the development of a consistent procedure for the uncertainty quantification of air quality model results that is reconcilable with the requirements of EU policy. The “Guidance on the use of models for the European Air Quality Directive” is an ETC/ACC report, summarising the efforts of the “Forum for air quality modelling in Europe” (FAIRMODE, http://fairmode.ew.eea.europa.eu/) to provide detailed instructions on the use of models for regulatory purposes, including quantification of uncertainty, according to the requirements of the Directive (see Ch.6).

In order to rely on model results for air quality decision making, both model performance evaluation as well as uncertainty estimation are of imperative importance. Model evaluation suggests validation of model results against measured data (DEFRA, 2010). In the case model results are found to compare well with observed values, the model is considered to accurately represent physical reality. Model evaluation includes the issues of model spatial and temporal resolution, which has to be consistent with or representative of the measured data, in order for the model to perform well in a point-by-point comparison between predictions and observations. In terms of comparing model results to measurements, the “fit-for-purpose” of the model has to be taken into account. For example, the model may not produce realistic results if applied to physical situations that are different to those used to derive the model. This consideration is relevant both to simple air quality models (e.g. empirical models), but also to more complex modelling tools. An empirical model based on near ground data over a flat agricultural field would not perform well if applied to dispersion from a tall stack in a domain of complex terrain (Hanna, 1988). In the case of complex Eulerian models, the physical assumptions and Planetary Boundary Layer parameterisation schemes used in the model will influence its performance under different meteorological conditions. It is therefore necessary to perform model evaluation under different atmospheric conditions, using different spatial and temporal resolution and for both homogenous and complex terrain cases, in order to identify the models limitations. The model can then be calibrated on the basis of a sensitivity analysis that will identify the conditions and input data for optimal performance. The sensitivity analysis is part of a diagnostic model evaluation and is connected to uncertainty analysis, as the effect of different parameters on the “error” or “uncertainty” of the model output is examined and quantified.

The deviation between model calculation and measurements is correlated to model uncertainty, assuming that measured data are a representation of reality or selected to be representative of the model temporal and spatial resolution. In this context, uncertainty is a measure of the reliability of the model results. The total uncertainty involved in air quality modelling simulations can be calculated as the sum of three components (Hanna, 1988): (a) the uncertainty due to errors in the model physics or deliberate simplifications to reduce computing time, (b) the uncertainty due to input data errors (meteorology, emissions data, boundary & initial conditions), (c) the uncertainty due to the stochastic nature of atmospheric processes (e.g. turbulence). The first component of model uncertainty may be reduced by introducing more physically realistic and computationally efficient algorithms. Some of the

effects of input data errors may also be reduced through the development and use of more accurate monitoring instruments. However, the stochastic fluctuations are a natural characteristic of the atmosphere that cannot be avoided. Stochastic fluctuations as some other uncertainty sources are of unknown magnitude and can not be realistically quantified. For example, in the case of model application for the prediction of air quality in relation to future emission scenarios, future atmospheric conditions (e.g. relating to climate change) or future technological advances (e.g. new technologies in vehicles) represent an important source of uncertainty (Colvile et al., 2002).

A number of statistical parameters are commonly used for the assessment of model uncertainty and for evaluating model performance against measurements (Chang and Hanna, 2004; Yu et al., 2006). Some statistical indices, such as bias, may be calculated in both cases. Other statistical parameters are more often associated either with uncertainty quantification (e.g. standard deviation and probability distribution functions) or with model performance evaluation (e.g. correlation coefficient and index of agreement). The statistical results of an evaluation analysis must be considered as the basis of a comprehensive methodology to assess model performance which will provide the explanation for model deviation. A suggestion for a standardised methodology of model evaluation and uncertainty analysis in support of European legislation requirements still remains a challenge for the scientific community. The current-state approaches performed by regulatory bodies and scientists in several EU countries have to be assessed in order to identify problems and needs. A recent survey was undertaken within FAIRMODE SG2 activities. The survey consisted of distributing questionnaires to EIONET NFPs representatives (representing 40 European countries) and 49 experts and regulators (Fragkou et al., 2012). Questions on the methodology used for the evaluation and uncertainty estimation of model results were addressed in the questionnaire. Although the questionnaire focused on the application of models only for the purpose of source apportionment, a few important results were obtained. The majority of the reported studies have applied some form of evaluation methodology, which most commonly involved comparison of model results with measured data and model intercomparison. However, limited information was contained in the returned questionnaires on the estimation of uncertainty and several researchers have commented on the need for a guidance including common criteria, indicators and performance measures to facilitate the procedure. In view of these needs, another survey was conducted within the APPRAISAL project in order to obtain more detailed information on the uncertainty estimation and evaluation methodologies used in model applications for different regulatory purposes by EU member states. The novel element in this questionnaire distributed to interested stakeholders and modellers within the frame of the APPRAISAL project was that not only air quality models were included but other types of models used for source apportionment or Health Impact Assessment studies were represented. Also the application of evaluation and uncertainty estimation procedures used in Integrated Assessment Modelling was examined.

Before discussing the results of the APPRAISAL questionnaire (Ch.4), it is useful to present state-of-the-art practices of uncertainty estimation and model evaluation related to the use of models for policy purposes.

2 STATE-OF-THE-ART

2.1 Uncertainty in regard to Source Apportionment applications

Understanding the factors that contribute to the uncertainty in Source Apportionment (SA) studies is quite complex, since the actual contribution of pollution sources to the level of pollutants observed using measurement instruments is unknown. As in every model, the uncertainty in source apportionment models' outputs depends largely on the quality of the input data. In addition, the noise in the input data may be amplified by the one introduced by the model to the output.
In receptor models the uncertainty derives from both inaccuracy in the input data and model assumptions and ambiguities. Interpreting the results of a source apportionment study and comparing results from studies in different sites or in the same site with different models requires proper uncertainty estimation (Karagulian & Belis, 2012). Considering that receptor models rely on the mass conservation principle between source and receptor, substantial departures from this assumption due to evaporation, condensation or degradation of species, constitute a source of uncertainty. When quantitative information about the processes that precursor species undergo after emission is available, it is possible to introduce empirical coefficients (e.g. Fractional Aerosol Coefficients, Grosjean and Seinfeld, 1989) to estimate the expected amounts of products at the receptor.

In multivariate models, the number of relevant factors and their correspondence with emissions is unknown and represents another source of uncertainty. Estimating the number of factors is often performed with an iterative procedure by checking the influence of the number of factors on the model performance. Another contribution to the overall uncertainty in factor analysis is the lack of a unique solution due to the large number of unknown variables, so called rotational uncertainty (Paatero and Hopke, 2009). This limitation is partially removed by the introduction of non-negativity constraints. In addition, a number of tests (e.g. FPEAK tests and analysis of residuals) can be used to identify the best rotation (Paatero et al., 2002). Moreover, introducing additional information about the sources and other constraints contributes to reduce or eliminate the rotational uncertainty.

In receptor model studies there is a distinct trend towards tools that estimate the uncertainty of their outputs. In fact, while one third of the studies published before 2010 reported source contribution uncertainty, this value has raised to two thirds for the studies published since 2010 (Belis et al., 2013). Mostly, the reported uncertainties vary between 2% and 60%. Uncertainties derived for most Positive Matrix Factorization (PMF) and Chemical Mass Balance (CMB) studies include the measurement error and the quality of the fit to the mass balance equation. To account for the uncertainty associated with the selection of fitting species and source profiles, some authors have statistically evaluated the results for a number of different solutions (Subramanian et al., 2007, Gelencsér et al., 2007, Gilardoni et al., 2011). In addition, new tools with improved uncertainty evaluation are at an advanced development stage (Paatero, pers. comm.).

As an alternative, overall model uncertainty may be assessed by comparing models or

specific implementations of models. Different approaches have been used to compare the performance of different models on the same dataset ranging from simple visual comparison of models' source contribution estimate (SCE) mean and standard deviation for each source type to regression analysis between the SCE obtained with different models. More recently, a methodology to evaluate intercomparison results on the basis of international standards for proficiency testing exercises has been used (Karagulian & Belis, 2012). This kind of intercomparison exercises consists of comparing the results of source apportionment analyses performed by independent practitioners using the same or different RMs on the same dataset. The main objectives of an intercomparison exercise are: a) to gather information about the reproducibility between different approaches and scientific backgrounds and b) to assess whether the uncertainty of the model output (SCE) meets given quality criteria. At present, almost 400 source contributions estimated by 38 participants have been evaluated in two European exercises (Karagulian et al., 2012 ; Belis et al., 2013 in preparation). The results indicate a good quantitative agreement between the source contribution estimation of the reported solutions. More than 80% of the solutions meet the quality criteria corresponding to a 50% standard uncertainty. Nevertheless, the number of identified factors may vary among participants.

In terms of evaluation approaches for dispersion models used for SA applications, several reviews have revealed that in most EU SA studies reported in the literature, evaluation of results is indirectly accounted for, and efforts to systematically evaluate the performance of alternative methodologies and estimate their intrinsic uncertainties have been scarce (Favez et al., 2010; Viana et al., 2008). A current review was performed within the framework of FAIRMODE SG2 activities (Fragkou et al., 2012). In this FAIRMODE review, it was encouraging to note that a high percentage (88%) of reported SA studies have evaluated their results. The most frequently used SA evaluation method was by comparing model results to data obtained from dedicated measurement campaigns (59% of reported studies corresponding to 55% of EU countries). For specific pollutants, such as Polycyclic Aromatic Hydrocarbons, correlating calculated levels with other pollutants measured at the receptor site during sampling campaigns can be used for evaluating SA results. This method is only feasible if the ratio between the pollutant of interest and the measured pollutant is characteristic for a specific source (Larsen and Baker, 2003).

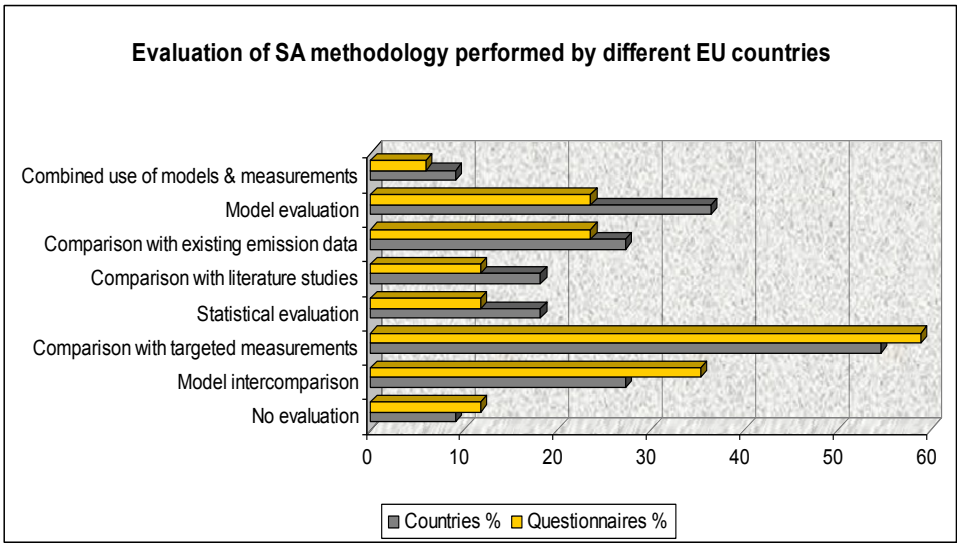

Figure 1: Evaluation method of SA in the responses of SG2 FAIRMODE review.

The next most common SA evaluation method used by EU member countries participating in the FAIRMODE review was model validation (36% of countries, 24% of questionnaires). This approach can involve the comparison and statistical evaluation of calculated pollutant concentrations against measured values, in order to test the performance of the dispersion model. Another SA evaluation method related to model validation is the sensitivity method, which is highly represented in the literature, particularly when dispersion models are applied for the identification and attribution of sources. SA modules incorporated into dispersion models can be evaluated by comparing SA results with results from model runs in which emissions from a particular source are greatly reduced (Brute Force Method - BFM) or set to zero (zero-out method). In the FAIRMODE review, information reported in a questionnaire returned by researchers from Spain, indicates the use of the zero-out sensitivity method to evaluate SA results for NO_x and O_3, based on dispersion model calculations. The use of the brute force sensitivity method has been reported by researchers in Italy for SA evaluation of NO_x. An added advantage of the sensitivity methods for SA evaluation, especially in terms of regulatory needs, is that the relative importance of each source category and the potential implications on source-oriented emission control strategies can be examined. Also, they can be applied with a limited computational cost as the runs for SA evaluation need only cover limited time periods, ideally for which measurement data are also available. However, the applicability of this method is pollutant-specific and depends on the linearity of the chemical reactions of the examined pollutant (Yarwood et al., 2005). For example, due to non-linearity of nitrate chemistry reactions, zero-out results have potential deficiencies as source apportionments for the case of SOA and NO_x.

Model intercomparison as the preferred SA evaluation method was reported in a considerable number of responses (27% of countries) of the FAIRMODE review. In some cases different receptor models were applied for SA and their results were compared. The combined use of different types of receptor models could solve the limitations of the individual models (Viana et al., 2008) and is therefore a method used frequently for SA evaluation. In SA studies reported in responses by other countries, results from different

dispersion model types were compared to evaluate NO_2 and O_3 SA results.

Existing emission data and emission inventories were used for SA evaluation in questionnaires returned by Italy, Finland and Spain, representing 24% of the questionnaires and 27% of the countries. Other SA evaluation methods that were less frequently reported by FAIRMODE review participants include statistical evaluation (18% of the countries and 12% of the questionnaires), comparison with literature studies for the area of interest (18% of the countries and 12% of the questionnaires) and the combined use of model results and meteorological observations to verify the validity of the SA results (9% of the countries corresponding to 6% of the questionnaires). Meteorological variables such as wind direction and seasonal circulation phenomena can be correlated to pollutant transport, thus indicating possible pollutant sources (Chakraborty and Gupta, 2010).

Regarding the estimation of uncertainties, no particular approach for calculating uncertainties was reported in the questionnaires returned by the member states in the FAIRMODE review. Receptor modelling tools, such as CMB, provide uncertainty estimates corresponding to the calculated values for contributions from each source as their standard output. However, source profile species and receptor concentrations, each with uncertainty estimates, should be provided as input data to the CMB model in order to calculate uncertainties of SA results (Fujita et al., 2007). Routine PMF analysis provides output uncertainty estimations based on input data uncertainty and bootstrapping.

2.2 Uncertainty in regard to Health Impact Assessment

Health impact analysis relies on two main processes, namely exposure assessment and epidemiological analysis relating exposure to the health outcome. These two processes include a number of basic steps, finally leading to the quantification of the expected atmospheric pollution induced health burden in the target population, most commonly expressed in terms of years of life lost attributable to the exposure to the atmospheric pollutant (s) under study (Krzyzanowski, Cohen and Anderson, 2002). Assumptions and uncertainties related to each process may significantly influence the result of the analysis. The main sources of uncertainty in HIA studies can be summarised as follows:

1. Uncertainties related to the results of the epidemiological studies or to their generalisation: It is therefore important that the selected health outcomes to be assessed are represented in reliable epidemiological studies, based on which reliable concentration-response relationships have been characterised. In terms of generalisation of epidemiological results, another important issue is the validity of extrapolating results from epidemiological studies carried out on a population to other populations for HIA. Although the biological processes linking exposure to susceptibility may not significantly differ between populations, a number of other factors could introduce bias and result in different exposure patterns for the same ambient concentration (e.g. differences in daily pattern of activity, climatic conditions, urban structure) or in different importance of confounding factors (Martuzzi, Krzyzanowski and Bertollini, 2003).
2. Uncertainties in estimating the impact for each health outcome: This uncertainty is mainly related to the health-outcome frequencies data. Mortality may be considered generally

accurate, but frequency measures of morbidity and data on health-care systems contain uncertainties (Künzli et al.; 2000). Furthermore, in contrast to directly countable events listed in national health statistics (e.g., deaths or injuries due to traffic accidents), it is not possible to directly identify the victims of mixtures with cumulative toxicity, such as smoking or air pollutants. Also, the health outcomes may not be specifically linked to air pollution due to synergistic effects with other factors.

3. Uncertainties in exposure assessment: Poor exposure assessment is an important source of uncertainty in HIA (Martuzzi, Krzyzanowski and Bertollini, 2003) and can result from errors and biases in either air quality models or in exposure models (Fuentes, 2009). Exposure models are mostly probabilistic models accounting for the numerous sources of variability, including human activity data (that is often neglected, considering an immobile population/ static population distribution). The different sources of error and uncertainties in the exposure models result from variability not modelled or incorrectly modelled, inaccurate inputs, errors in coding, simplifications of physical, chemical and biological processes to form the conceptual models, and flaws in the conceptual model. Emission and meteorological input data accuracy and physical/chemistry assumptions and parameterisations in the air quality model largely affect the reliability on its results, the spatial distribution of ambient pollutant concentrations. Furthermore, statistical methods (e.g. kriging) used to produce higher resolved air pollution fields starting from air quality model results and other inputs (local observations, emissions etc) may also introduce uncertainties at specific locations far away from the observations. Evaluation of the air quality and exposure models is therefore highly recommended in HIA, preferably on the basis of probabilistic methods (e.g. Bayesian analysis). In cases of limited data availability, presenting results from a small number of model scenarios could provide an adequate uncertainty analysis for the air quality and exposure models.
4. Uncertainties related to the concentration-response functions, estimated by epidemiological models: Some of the formal approaches for uncertainty analysis in epidemiological concentration-response models include Bayesian analysis, Monte Carlo analysis and model intercomparison (Fuentes, 2009).
5. Uncertainties concerning the temporal scale of effects, i.e. the latency times from exposure to adverse event. This is an uncertainty mainly associated with long-term exposure studies, as acute effects follow exposure by a few days (Martuzzi, Krzyzanowski and Bertollini, 2003).
6. Uncertainties related to the exposure reference value: In order to estimate attributable risks and attributable number of cases, as a function of concentration-response coefficients, a reference value for “no exposure” has to be defined (Martuzzi, Krzyzanowski and Bertollini, 2003). As assuming zero ambient pollutant concentration is not realistic, other assumptions may be used, such as using the estimated “natural” background, or using different reference levels associated to changes in pollutant concentration, in order to illustrate the potential benefits associated with different reduction policy scenarios. In the study by Künzli et al. (2000), the significant influence of the exposure reference value on the results of a HIA study was demonstrated. The health impact estimates would be ~54% higher if the exposure reference value was reduced from 7.5 $\mu g/m^3$ to zero.

2.3 Uncertainty in regard to Integrated Assessment Modelling

In the field of Integrated Assessment Modelling (IAM), uncertainty can be related to (Carnevale et al., 2012):

- the decision model approaches. In the literature different Integrated Assessment Modelling (IAM) methodologies are presented: scenario analysis (Thunis et al., 2007), cost-benefit analysis (Vlachokostas et al. 2009), cost-effectiveness analysis (Carslon, Haurie, Vial, & Zachary, 2004), multi-objective analysis (Pisoni, Carnevale and Volta, 2009; Guariso, Pirovano, & Volta, 2004). How does the approach impact effective planning design?
- the optimization algorithms. The decision problem is solved by means of optimization algorithms. How does the optimization algorithm bias the determination of effective policies?
- the planning indicators for human, ecosystems and materials exposure. The decision problem determines the abatement measures or other actions that optimize the objectives, and that have to comply with physical, economical and environmental constraints. Objectives and environmental constraints are typically indicators of human, ecosystems and material exposure. How do different sets of indicators impact on policies design?
- the source-receptor relationships. Due to the large computational resources needed to run deterministic 3D air quality modelling systems, it is not possible to fully integrate them into an optimization problem. For this reason source-receptor relationships, that present a simplified relation between emissions and pollutant concentration, need to be derived. What is the uncertainty of these source-receptor relationships (Pisoni et al.,2009)? Which is the sensitivity of the decision problem solutions to different source-receptor relationships?
- the baseline and projection emission scenarios and the emission reduction strategies
- the spatial scales. A decision problem can be defined for different scales and resolutions. Which are the approaches suitable for different scales?
- the meteorology. Source-receptor relationships are identified processing CTM simulations for different reference years. How do the meteorological conditions of reference years influence the design of policies?

From the point of view of uncertainty in policy making, it is also important to keep in mind the results of the “UNECE workshop on uncertainty treatment in integrated assessment modelling” (UNECE, 2002), in which it was concluded that policy makers are mainly interested in robust strategies. Robustness implies that optimal policies do not significantly change due to changes in the uncertain model elements. Robust strategies should avoid regret investments (no-regret approach) and/or the risk of serious damage (precautionary approach) (Amann et al., 2011). This issue is also linked to the need of defining a set of indexes and a methodology to measure the sensitivity of the decision problem solutions. It is in fact worth underlining that, while for air quality models the sensitivity can be measured by referring in one way or the other to field data, for IAMs this is not possible, since an absolute “optimal” policy is not known and most of the times does not even exist. The traditional

concept of model accuracy must thus be replaced by notions such as risk of a certain decision or regret of choosing one policy instead of another.

3 QUESTIONNAIRE STRUCTURE

The questionnaire that was distributed by APPRAISAL addressed five topics relevant to the use of models, including Topic 5 on uncertainty and robustness. The questionnaires were specifically addressed to national contact points in EU member states and stakeholders involved in the development of Air Quality Plans, but also to model users applying models in the frame of research projects. Both multiple choice questions as well as open questions were represented in this section of the questionnaire. Multiple choice questions were proposed as they are more straightforward to answer and process in a database afterwards, so they were carefully phrased in order to provide as much information as possible. However, as the problems and needs of the stakeholders could not be confined to the multiple choice questions, a number of open questions were also considered necessary.

3.1 Explanation of close-ended questions

Regarding the close-ended (multiple choice) questions of the questionnaire, the first two questions concerned model performance evaluation while the questions 3-6 addressed the issue of uncertainty estimation. In terms of performance evaluation, the aim was to obtain information on whether an evaluation methodology was applied for a particular model application (air quality modelling, source-receptor relationships, source apportionment, health impact assessment and integrated assessment) and on the type of the evaluation methodology performed. The choices suggested within the questionnaire for the types of evaluation methodology follow the framework proposed by Dennis et al. (2010) which distinguish four components of model performance evaluation as follows:

1. operational evaluation involves assessment of model results compared with monitored data, which may include routine or field campaign observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables.
2. diagnostic evaluation is a process-oriented analysis to determine whether the individual physical and chemical processes are correctly represented in the model.
3. dynamic model evaluation is the analysis of model responses to changes in model input data, such as source emissions or meteorological conditions.
4. probabilistic model evaluation is performed on the basis of methods such as model inter-comparison and ensemble modelling, and attempts to capture statistical properties, including uncertainty or level of confidence in the model results, for regulatory model applications. This approach requires knowledge of uncertainty imbedded in both model predictions and observations. Probabilistic model evaluation is particularly helpful for predicting the accuracy of model results for future emission changes, and it is therefore considered essential for future planning purposes (Hogrefe and Rao, 2001). In some cases, however, where there is a lack of sufficient data for evaluation, the dependence on expert judgment is required for decision making. Experts must rely on their scientific theoretical knowledge and their experience from similar applications of the examined model (Rao, 2005).

The questionnaire also requests information on the specific model use that was evaluated (air quality modelling, source-receptor relationships, source apportionment, health impact assessment and integrated assessment) and on the application of specific software for evaluation. The development of specific software tools for model evaluation is mainly related to operational model evaluation, as the tools provide a platform for statistic analysis of model results compared to measurements. Statistical evaluation tools have been developed since the first model applications for regulatory purposes, as for example the Statistical Analysis System (SAS) (Baldridge and Cox, 1986). Examples of validation tools include the BOOT model evaluation software package (e.g. Chang and Hanna, 2004) and the Atmospheric Model Evaluation Tool (AMET) for evaluating meteorological and air quality models (Appel et al., 2011). A comprehensive tool (DELTA tool) including benchmarking service, where templates for reporting model performance according to EU legislation are also automatically produced, has been developed within the frame of FAIRMODE activities (Thunis, Georgieva and Pederzoli, 2012). It is important to gain information on the extent of use of such tools for regulatory applications in EU member states.

Considering uncertainty analysis, the questionnaire includes questions on the uncertainty quantification methodologies used (global or local methods) and on the model components that were individually assessed. Meteorological parameters have long been recognised to affect air quality model results (Seaman, 2007) and are a widely examined uncertainty source. Moreover, much attention has been given by scientists to the importance of accurate emission input data for realistic air quality model results. For example, in a study by Digar et al. (2011), methods for estimating the likelihood that a given level of emission reductions will achieve a targeted improvement in air quality, in light of parametric uncertainties in the photochemical model used, were suggested. Emissions are a source of uncertainty that can be improved, and, therefore, policy-makers have pursued continuous improvement of the reliability of national emission inventory data (DEFRA, 2010). Uncertainty analysis of the model algorithms, physics, assumptions and codes is also an integral part of scientific evaluation which examines the accuracy, efficiency and sensitivity of model formulation. However, such an analysis is usually not performed when a model is applied for regulatory purposes, but should ideally precede model application. In many legislative applications, reliance on model results is based on the scientific evaluation of the model in previous studies.

3.2 Explanation of open-ended questions

Apart from the multiple choice questions, six open-ended questions were also proposed in the questionnaire. The first question aimed at receiving more in depth and detailed information on the performance indicators considered in the uncertainty/evaluation analysis. In the next two questions, model users were asked to explain in their own words how they judge that their model application results for both assessment and planning purposes have reached a sufficient level of quality. Furthermore, the open questions provide the opportunity for model users to comment on the problems and difficulties they have encountered in performing uncertainty analysis and model evaluation and to report the reasons for not

undertaking these quality control procedures. This question was added in the questionnaire in order to examine the needs and limitations of model users. The scientific community of APPRAISAL will then proceed to identify and suggest a methodology to respond to these needs, thus facilitating the introduction of evaluation and uncertainty assessment procedures in air quality model applications for regulatory purposes in the EU. It is important to note that all questions and answers refer only to the spatial scales addressed by the APPRAISAL project, i.e. the regional and local scales.

One of the open questions was related to the quality control methodology used in case of applying an Integrated Assessment Model (IAM). In terms of the treatment of uncertainty in IAMs, the most common approach is to separately consider the uncertainty of the different model components, for example the uncertainties in the meteorological model, the uncertainties in the air quality model and the uncertainties in the cost-benefit model. In this approach, the main aim is to accurately quantify the existing uncertainties of the IAM separately. Another aspect in the assessment of IAMs is uncertainty prioritisation, which aims to identify the weakest components of the system. These are the constituents whose individual lack of quality contributes the most to the overall lack of quality in model results.

4 ANSWERS ANALYSIS

4.1 Close-ended questions

The responses to the APPRAISAL questionnaires were collected and stored in a database which was developed within the frame of the project. Out of the 53 questionnaires received at the time of this report, 39 included responses to the topic on “uncertainty and robustness” (one of the five topics addressed by the questionnaire). The responses reported the current practise in quality control procedures when applying integrated assessment modelling for air quality related studies (including Research Projects and Other studies) and Air Quality Plans. As the main component in air quality integrated assessment, air quality modelling was the most commonly evaluated component (Figure 2).

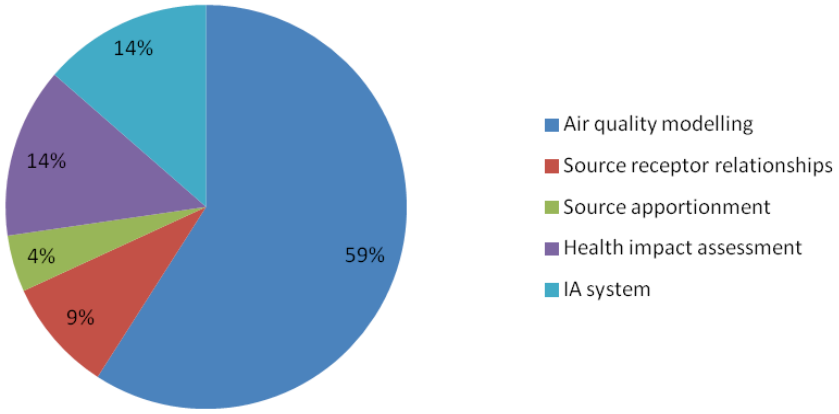

Figure 2: Evaluation frequency for different components of air quality integrated assessment.

In terms of the evaluation methodology used, operational and diagnostic methods were applied with higher frequency in comparison to the other methods, while expert judgement was also reported in a significant number of responses. Evaluation methods of higher complexity, such as dynamic and probabilistic approaches, were only applied in very few cases (Figure 3).

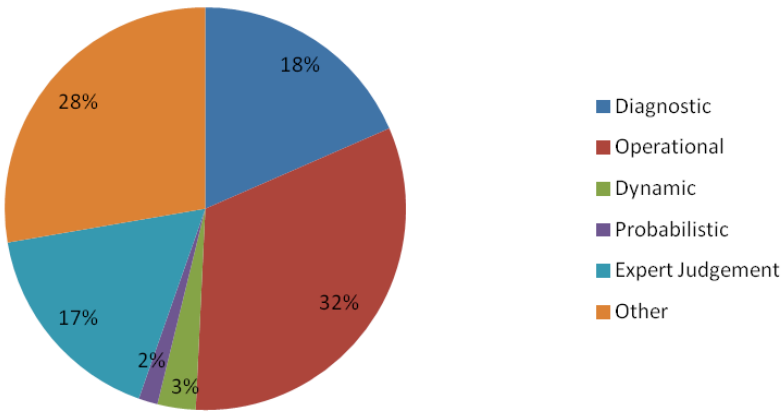

Figure 3: Frequency of use of different evaluation methodologies, as reported in the APPRAISAL questionnaires.

Out of these 39 responses on the topic of uncertainty estimation and model evaluation, 20 were regarding air quality plans (AQPs) while 15 were research projects (RPs) and 4 represented Other purposes. In the following sections, the analysis of the answers on model evaluation and uncertainty estimation will also be correlated to the purpose of model application, in particular whether the model was used within the frame of a RP or of an AQP.

As it would be expected, the majority of model users rely on the operational evaluation technique (comparison with measurements) to assess the quality of the model results both in AQPs and RPs (Figure 4). The other evaluation methods were also represented in the returned questionnaires, although not so commonly applied. In the case of RPs, the percentage of responses indicating the use of a probabilistic or diagnostic method increases, whereas the number relying on expert judgement is relatively low. It can be therefore concluded, that a more comprehensive model evaluation process is performed in European member states in the frame of RPs than for AQP, with the operational evaluation dominating but complemented by other techniques. This can be attributed to the fact that these additional evaluation techniques require intensive personnel, infrastructure and time resources.

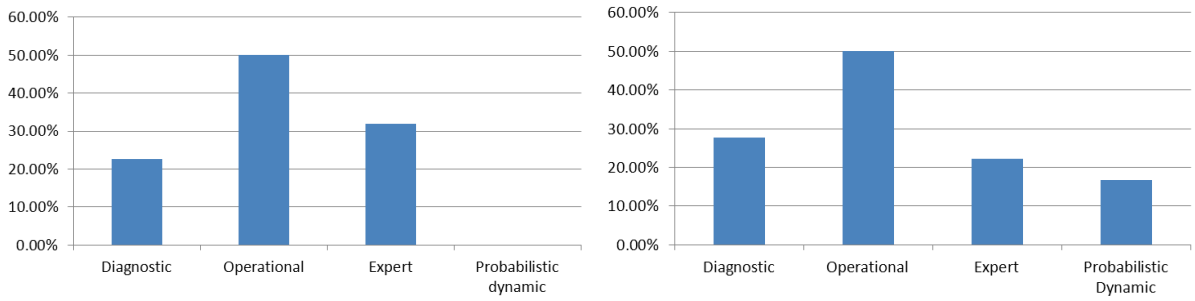

Figure 4: Overview of evaluation methodologies used for the assessment of Air quality plans (AQP) and Research projects (RP). Note the total can exceed 100% as more than one methodology can be used at the same time.

In terms of uncertainty analysis, it becomes obvious from the results of the questionnaire responses that uncertainty was mainly considered in the air quality modelling part of the IAM applications (Figure 5).

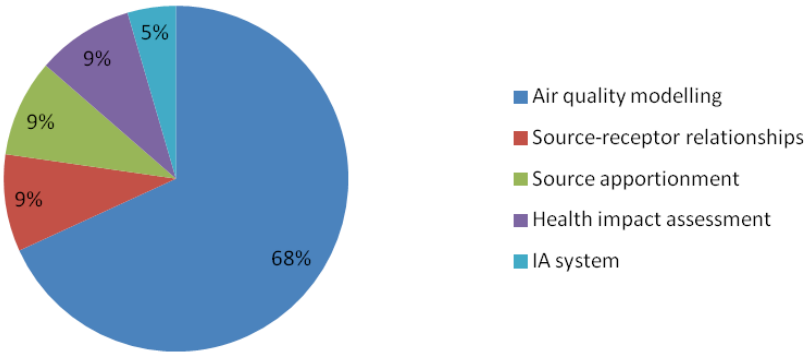

Figure 5: Uncertainty estimation separately considered in different model applications within IAM.

9% of the responses reported that uncertainty estimation was performed for source-receptor relationships, source apportionment and health impact assessment, while uncertainty quantification for the IA system as a whole was represented only in 5% of the responses.

Regarding the uncertainty analysis that was reported in the questionnaires, in the case of AQPs (Figure 6, left), 12 plans quantified uncertainty for the plan under study, 8 plans used uncertainty analysis from previous studies, while for 8 plans no response was obtained on whether uncertainty analysis was performed in the specific study or in previous studies. In the case of RPs (Figure 6, right), the majority (13 questionnaires) of the replied questionnaires reported uncertainty quantification specifically performed for the examined study, only 2 questionnaires have used previous studies and 3 out of 18 have not applied uncertainty quantification techniques.

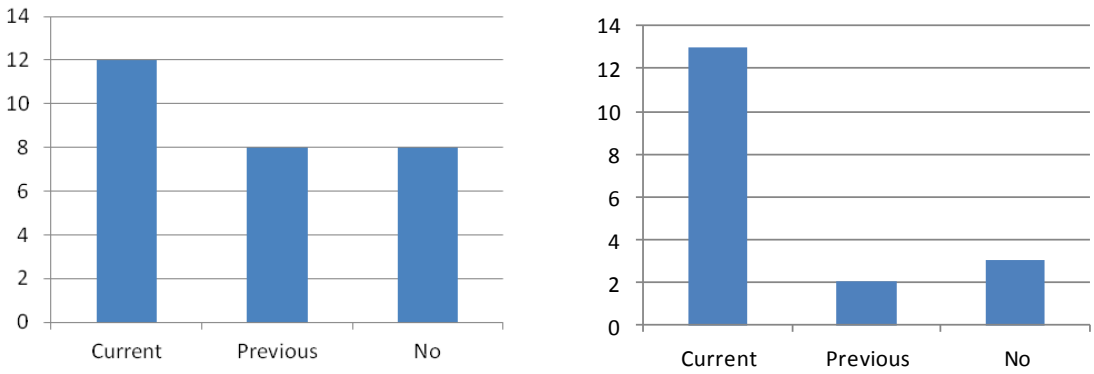

Figure 6: Uncertainty estimation performed in AQPs (left) and RPs (right). “Current” refers to uncertainty analysis applied to the reported IAM study, while “Previous” refers to uncertainty analysis performed in previous studies but for the same models.

Global uncertainty analysis methods (e.g. Monte Carlo analysis) have been used in more studies compared to local uncertainty analysis methods, both in AQPs (Figure 7, left) and, more significantly, in RPs (Figure 7, right). It should be taken into account that the results in Figure 6 include the use of global or local methods for uncertainty analysis performed both in the specific (current) study as well as in previous studies. Also, in some of the questionnaires, no answer was provided for the methodology used (local or global), particularly in the case of AQPs.

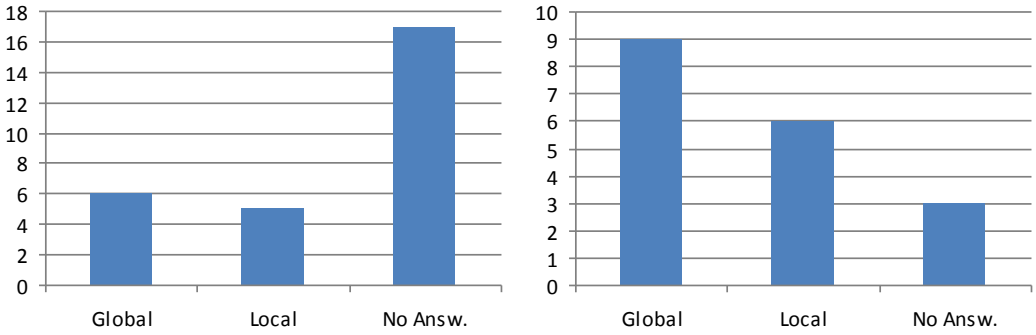

Figure 7: Uncertainty analysis approaches in AQPs (left) and RPs (right).

Variance-based uncertainty estimation methods are the most commonly used of the global uncertainty assessment methods. However, local uncertainty analysis methods (sensitivity methods, OaT) are also significantly represented in the responses, particularly in the case of RPs (Figure 8).

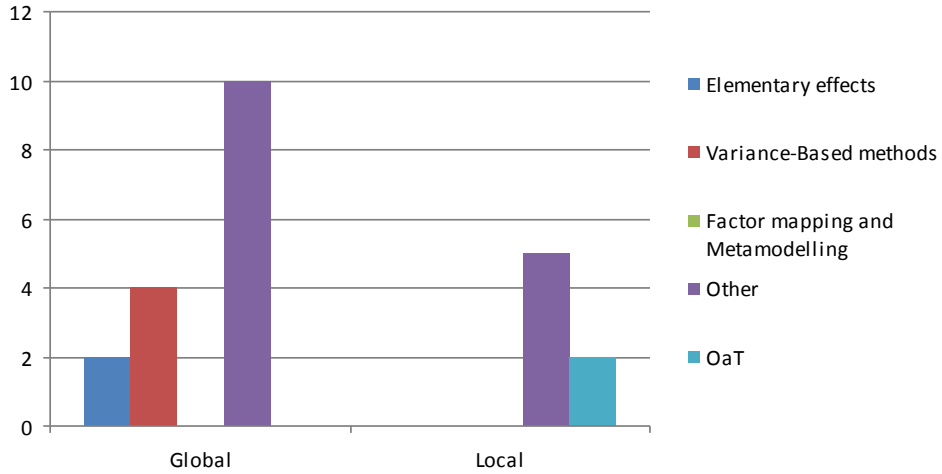

Figure 8: Local and Global analysis methods in the questionnaire replies.

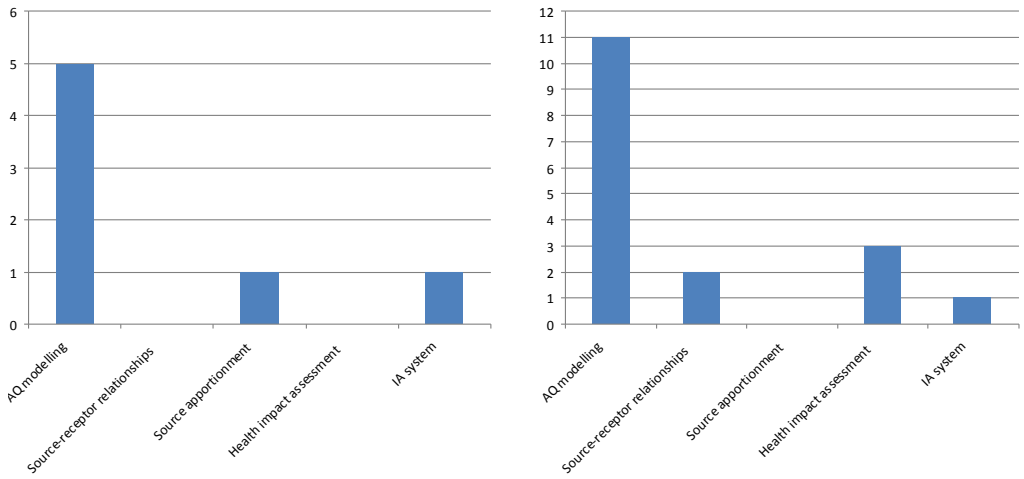

Figure 9: Uncertainty estimation in different IAM components in AQPs (left) and RPs (right).

AQ modelling is the IAM component for which uncertainty analysis is most commonly considered in the questionnaire responses, both in the case of AQPs (Figure 9, left) as well as for RPs (Figure 9, right).

Figure 10 provides information on the AQ modelling elements for which uncertainty estimation was specifically carried out. As expected, model formulation was not one of the priority aspects examined in the case of AQPs (Figure 9, left), it was however considered in a significant number of RPs (Figure 9, right). In the frame of AQPs, uncertainties were mostly analysed for meteorology, emissions and boundary conditions (BC). Regarding RPs, it is interesting to note that uncertainties related to boundary conditions received less attention. For both AQPs and RPs, emissions related uncertainties are identified to significantly contribute to the total AQ modelling uncertainties. The problem of emissions related uncertainty was also often commented in the “open questions” replies of the questionnaire.

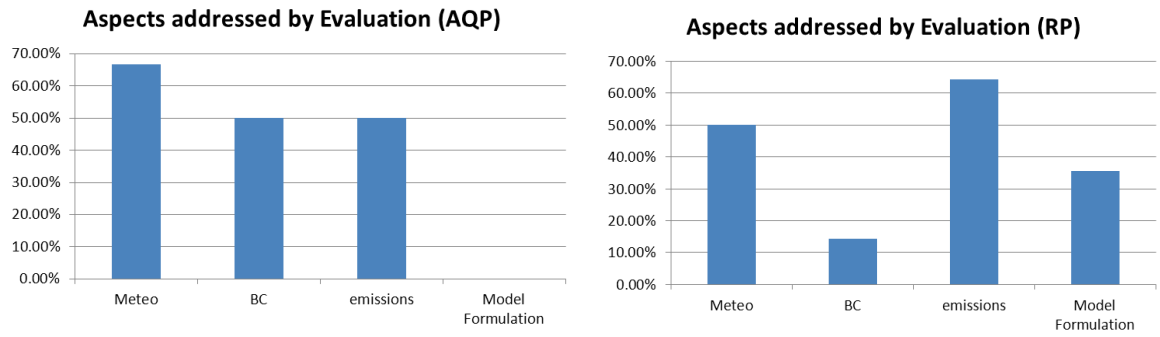

Figure 10: Uncertainty estimation of different components of AQ modelling in the case of AQPs (left) and RPs (right). Note that the total can exceed 100%, as more than one methodology can be used at the same time.

4.2 Open-ended questions

From the analysis of the answers received to the open questions of the questionnaire, the main approaches in the quality control of model results used for assessment purposes are summarised below. The following four levels must be seen as progressive (from the simplest to the most complex).

1. The model is assumed of sufficient quality and fit for assessment purposes because it is peer-reviewed and has been tested in many configurations. Past experiences are here the main quality justification.

2. The model is assumed of sufficient quality by performing a comparison of model results with observations using basic statistical indicators (e.g. correlation, Root Mean Square Error, bias). Expert judgement or comparison with other studies is then used to assess the quality of the AQ model results. Continuous evaluation against measurements in forecast mode is also seen as a method to maintain a high quality level.

3. The model is assumed to be of sufficient quality by comparing values obtained for selected statistical indicators (e.g. Bias, Index of Agreement) with values from literature (e.g. Boylan and Russel 2006, Gilliam et al. 2006) or with reference values, e.g. values set in the EU Directive 2008/50/EC which requires a model result uncertainty less or equal to 50%.

4. In addition to the previous approaches, the model is assumed of sufficient quality after its evaluation in model inter-comparison studies. We also added in this step the application of the updated recommendations for model evaluation of the FAIRMODE network.

Figure 11 illustrates the distribution of the responses organised around these four levels/categories. The majority of answers corresponds to level 3, indicating the comparison of statistical indicators with reference values (literature, Air Quality Directive). It must be noted that all answers corresponding to level 4 (more time/cost demanding one) refer to RPs exclusively, and are not used to assess the quality of the model results for applications regarding AQPs.

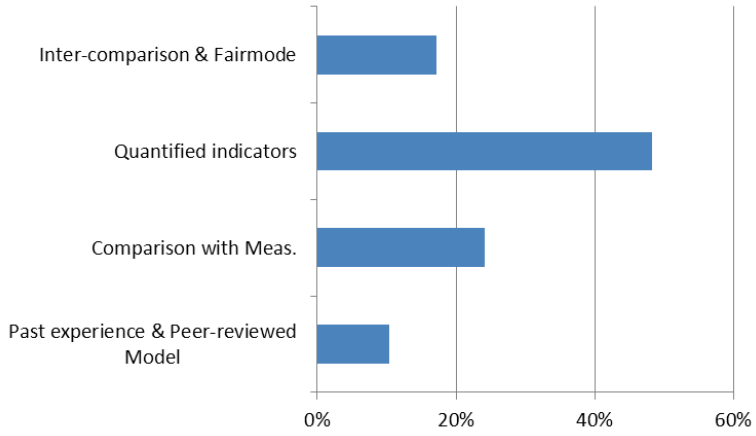

Figure 11: Overview of referred methodologies to assess the quality of model assessment applications (ranked as mentioned in text above). Both AQPs and RPs are considered.

In terms of quality control of model results for planning applications, the following approaches were represented in the replies to the questionnaires.

1. The model is assumed to be adequate for planning when it behaves correctly for assessment applications.

2. The reliability of the model is based on model intercomparison and ensemble approaches.

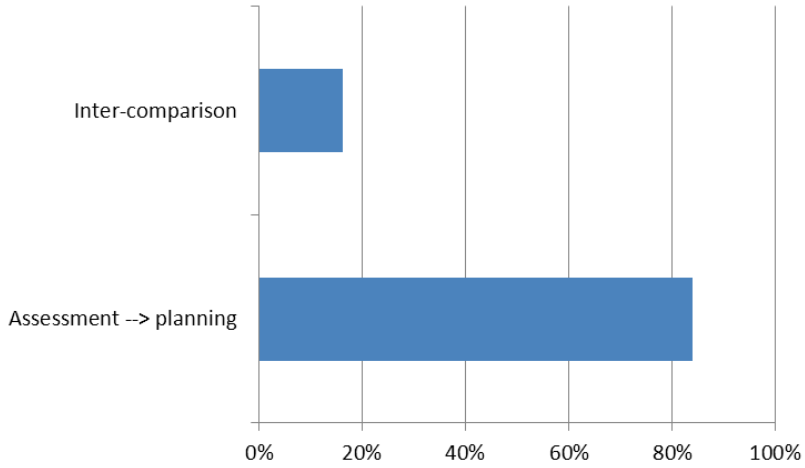

Figure 12: Overview of reported methodologies to assess the quality of model results for planning applications.

Most of the answers (Figure 12) indicate that the quality of the model results for planning applications (typically to investigate the impact of an emission reduction scenario) is not explicitly assessed, but relies on evaluated model performance for assessment purposes. This is probably related to the fact that as planning applications refer to future time, no reference observations exist to compare model results with. Furthermore, the quality control methodologies which could be used (such as model inter-comparison, sensitivity simulations) are relatively time-consuming and often require qualified personnel and infrastructure resources, which are usually only available within a research project. It is interesting to note, that no reference technique is proposed so far to check the quality of the models used to quantify the impact of emission reduction scenarios in AQPs.

Although not many replies were received in the open questions, a number of issues and needs can be identified in current-state estimation of model uncertainties, as reported by model users and regulators. These are summarised below and relate both to AQPs and RPs:

- Reduce uncertainties in the model input data, particularly emissions (for specific compounds, e.g. NH_3, or for specific sectors, e.g. transport and residential heating).

- Need for more experimental data to validate models, especially for small scale models.

- In terms of improving modelling tools and introducing best practices in air quality modelling, refining of model resolution to address modelling at smaller scales is essential. Similarly, higher resolution input data are needed. Extension of modelling periods to a full year or longer is also required in many applications. Finally, improvement of model quality during periods of specific weather conditions is recommended, e.g. for winter time especially for NO_2 and PM_{10}.

- Need for an alternative evaluation technique apart from operational evaluation (diagnostic, dynamic, sensitivity studies) and for a relevant protocol.

5 LIMITATIONS OF THE CURRENT ASSESSMENT AND PLANNING TOOLS AND KEY AREAS FOR FUTURE RESEARCH AND INNOVATIONS

Uncertainty estimates are an essential element of air quality assessment. Uncertainty, information is not intended to directly dispute the validity of the assessment estimates, but to help prioritise efforts to improve the accuracy of those assessments in the future and guide decisions on methodological choice as regards the tools that are being used. Every type of model evaluation faces different types of limitations, therefore it is easier to consider these according to the classification proposed by Dennis (2010), as described in section 2.1.

Operational

Measurements, as well as models are not always fit for purpose. For example, the issues of representativity of measurements as well as the matching of temporal and spatial scales between measurements and models are largely unresolved and do thus present challenges in the process of operational evaluation. In addition, operational model evaluation often comprises large amounts of statistical indices and diagrams which require a "second-order" analysis in order to provide useable information.

Diagnostic

This kind of evaluation is usually performed in order to identify the processes which present implementation problems and to estimate their impact on the final results. Techniques and practices for diagnostic evaluation have been described in various works (e.g. Saltelli et al 2004, Saltelli et al. 2008, Cullen and Frey, 1999) and are now theoretically sound but are only rarely used in the field of air quality modelling (Galmarini et al. 2010). Several past studies have been based on sensitivity analyses of meteorological input to air quality models (e.g. Hanna and Yang, 2001), however a methodological framework of how to do this for the purposes of air quality assessment and not weather prediction is still lacking (Dennis et al, 2010). Finally, ad hoc or specialised measurements are often needed in order to perform detailed diagnostic evaluation (eg. speciated PM, VOCs), which are not always available.

Dynamic

This type of evaluation is considered a good example of policy-relevant science. The main challenge in applying dynamic evaluation lies in the ability to distinguish the impact of changes in emissions in the absence of meteorological changes. This need calls for cases with specific characteristics: emission changes should be larger than 15-20%, the variability of concentrations should be discernible in the observations and the variability should be regional or local scale. There are quite a few good examples of such cases, however the data needs are often prohibitive for a thorough examination (Galmarini et al., 2010).

Probabilistic

While brute force methods to evaluate model application from a probabilistic point of view do exist for quantifying the nature and magnitude of model uncertainties, a comprehensive, theoretically based and computationally affordable framework remains to be defined. Various advances have taken place towards this goal, moving us from Monte Carlo methods (Moore

and Londergan, 2001; Beekman and Derognat, 2003 ; Werner et al., 2005 ; Werner, 2009) to methods such as the Direct Decoupled Method, the use of Green functions or Stochastic Response Surface Methods (Isukapalli et al., 1998). It is worth noting here that 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 very difficult. In order to perform a statistically valid comparison, the differing probability distributions for the two quantities must be taken into account (Galmarini et al. 2010).

6 CONTRIBUTION TO THE AIR QUALITY DIRECTIVE

In order to assess the total uncertainty and evaluate the performance of an IAM system, the uncertainty related to the different modelling components of the system (meteorological modelling, air quality modelling, exposure modelling, cost-benefit modelling) has to be separately quantified. However, it remains a scientific challenge to interconnect all the individual uncertainties of IAMs, as the chemical and physical processes involved are not linear and, also, some uncertainties may compensate each other. Combining all uncertainties to calculate a total uncertainty would require a great number of simulations to take into account all possible combinations. This complexity does not allow for setting straightforward quality criteria in terms of IAMs, even though IAM is now considered an important policy tool. In terms of AQ policy, AQ modelling is the IAM component explicitly mentioned in EU legislation. In particular, the 2008/50/EC Framework Directive places more emphasis on, and encourages, the use of models in combination with monitoring in a range of regulatory applications, in comparison to previous Directives, which have based AQ assessment and reporting almost exclusively on measurement data.

In contrast to measurements, no reference methodology for modelling is defined in the AQD, but, as with measurements, model results have to meet certain accuracy standards (Stern and Flemming, 2004). These standards are set with regard to the calculated annual, daily and hourly pollutant values, as air quality modelling is primarily applied for AQ assessment, in order to assess compliance with limit values of similar temporal resolution. However, as the directive does not provide guidelines on how to carry out model evaluation to achieve the quality requirements imposed, the development of relevant guidelines is necessary for modellers and authorities. Several attempts have been made for the establishment of uncertainty assessment guidelines within a number of projects, including AIR4EU (Denby et al., 2011) and FAIRMODE. The Guidance Document that was elaborated within FAIRMODE is the current reference point for model users and regulators to ensure that their air quality model meets the quality criteria required by EU legislation.

6.1 Minimum requirements and methods to achieve them

AQ modelling may be applied to a range of applications relevant to the AQD, including assessment of the existing air quality and compliance with limit values, management (including mitigation and planning for future air quality) and source apportionment. However, in the AQD, modelling is only explicitly mentioned in regard to the application of assessment and, therefore, the model quality objectives defined in Annex I of the AQD apply only to air quality assessment applications when reporting exceedances.

The modelling quality objectives are described in Annex I of the AQ Directive and are given as a relative uncertainty (%). Uncertainty is then further defined as: "*The uncertainty for modelling is defined as the maximum deviation of the measured and calculated concentration levels for 90 % of individual monitoring points, over the period considered, by the limit value (or target value in the case of ozone), without taking into account the timing of the events. The uncertainty for modelling shall be interpreted as being applicable in the*

region of the appropriate limit value (or target value in the case of ozone). The fixed measurements that have to be selected for comparison with modelling results shall be representative of the scale covered by the model."

Two mathematical formulations have been proposed to estimate model uncertainty for air quality assessment applications, in view of the previous definition of uncertainty according to the AQD. The formulation in the FAIRMODE Guidance Document (Denby, 2010) calculates the Relative Directive Error (RDE), while the formulation proposed by Stern and Flemming (2004) calculates the Relative Percentile Error (RPE). Both these formulations are based on the divergence between model results and available measurements at a particular station. Therefore, it is important that spatial and temporal model resolution is considered for using any of the two formulations. When the observed and modelled concentrations are well below the limit value, the RDE formulation of uncertainty is recommended.
Other factors to consider when applying AQ modelling for air quality assessment are the following:

- The "90% of stations requirement", according to which the AQ Directive states that the uncertainty will be determined from the maximum of 90% of the available monitoring stations, in order to exclude outliers from the uncertainty calculation. However, this does not apply if less than 10 monitoring stations correspond to the same scale as the model, in which case all stations have to be considered.
- In order to use model results with confidence for compliance purposes, it is important that the model has been adequately validated for the particular application and well documented and that it contains the relevant physical and chemical processes suitable for the type of application, the scale and the pollutant for which it is applied.
- Finally, the quality of required input data has to be sufficient, e.g. the relevant emission sources for the application need to be adequately represented and suitable meteorological data must be available.

6.2 Standardisation and harmonisation

In response to the need for a standardised methodology to perform uncertainty estimation when relying on the results of an AQ model for air quality assessment, the DELTA tool has been developed within the frame of the FAIRMODE activities. The DELTA Tool is a model evaluation software which provides summary statistics (i.e. BIAS, RMSE, correlation coefficient) as well as scatter-plots, time series plots, Taylor, Target and other diagrams providing an overview of the quality of model results against available observations (Thunis, Georgieva and Pederzoli, 2012). In particular, the statistical indicators calculated by the DELTA tool are presented in ANNEX III. A number of customised diagrams (Timeseries, bar plots, scatter diagrams, Taylor/Target diagrams, etc.) can be automatically produced on the basis of the statistical results. A benchmarking service is also implemented in the DELTA tool, which automatically produces standardised summary reports containing performance indicators related to a given model application according to AQD requirements. These indicators provide an overview of the strengths and weaknesses of each model. Different

performance criteria are suggested per statistical indicator for each pollutant and spatial scale. The DELTA tool facilitates the evaluation of the model performance and provides several functions, for example the possibility of switching from one diagram to another while keeping the same data selection and the possibility of identifying differences in model behaviour in terms of location, station type, etc. The DELTA tool software along with detailed instructions for use is currently restricted to people actively involved in FAIRMODE activities (http://aqm.jrc.ec.europa.eu/DELTA/disclaimer.htm). The tool has already been tested in various applications, including the evaluation of the WRF model (Miglietta et al., 2012) and of the Transport Chemical Aerosol Model (TCAM ; Carnevale et al., 2008).

In conclusion, the current AQD (2008/50/EC) dictates the need for Member States to report uncertainties of both the monitoring and modelling assessments. However, the Directive does not state what these uncertainties actually imply, i.e. that there is some probability for an exceedance, rather than a definite, yes or no. For this reason a major challenge for the scientific community in the coming years but also during the drafting of the revised AQ Directive is to provide a robust methodology for the uncertainty assessment in exceedances reporting that effectively deals with this aspect. This can be achieved by taking into account other types of legislation that use methods for dealing with probability and uncertainty.

7 CONCLUSIONS AND SUMMARY

In the present report of the Deliverable D2.5 on “Uncertainty and Robustness”, current state of the art approaches in model validation and uncertainty estimation are reviewed and their limitations are briefly described. The focus of the report is on model use for regulatory purposes and therefore, the different uncertainty approaches in Air Quality Assessment, Health Impact Assessment and Integrated Assessment Modelling are considered, in view of the EU legislation requirements. Information for this review was derived from published scientific papers and from the answers received in response to the questionnaire distributed within the framework of APPRAISAL activities. Model quality assessment and evaluation methods are examined separately for model use in relation to Air Quality Planning and for model use in relation to other purposes, e.g. Air Quality Assessment or research projects.

The main outcome from the analysis of the questionnaire replies indicates that model evaluation and uncertainty estimation is more regularly performed in air quality modelling, while it is not often applied in other IAM components such as for example in the case of HIA applications. Operational and diagnostic evaluation are the evaluation methods preferred both in the case of modelling for the purpose of air quality planning as well as for research projects. For the purpose of Air Quality Plans, expert judgement is also frequently used. Uncertainty propagation methodologies are also used, although not so often, to quantify confidence levels of Air Quality model results. The needs that emerged from the replies were related to the quality and quantity of input and validation data and to the improvement of modelling tools and the use of best modelling practices. Many replies reported the need for the establishment of an evaluation protocol in order to standardise and harmonise validation and uncertainty estimation methods in EU countries.

REFERENCES

1. Amann M. et al. (2011) Cost-effective control of air quality and greenhouse gases in Europe: Modeling and policy applications, Environmental Modelling & Software, 26, 1489-150.
2. Appel K.W, Gilliam R.C., Davis N., Zubrow A., Howard S.C. (2011) Overview of the atmospheric model evaluation tool (AMET) v1.1 for evaluating meteorological and air quality models, Environmental Modelling & Software, 26, 434-443.
3. Baldridge K.W. and Cox W. M. (1986) Evaluating air quality model performance, Environmental Software, 1 (3), 182-187.
4. Beekmann M., Derognat C. (2003) Monte Carlo uncertainty analysis of a regional-scale transport chemistry model constrained by measurements from the atmospheric pollution over Paris area (ESQUIF) campaign, Journal of Geophysical Research, 108 (D17), 8559.
5. Belis C. A. and Karagulian F. (2013) Results of the European Intercomparison exercise for Receptor Models 2012-2013. Part II. JRC Scientific and Policy Reports. (in preparation).
6. Belis C.A., Karagulian F., Larsen B.R. and Hopke P. K. (2013) Critical review and meta-analysis of ambient particulate matter source apportionment using receptor models in Europe, Atmospheric Environment, 69, 94-108.
7. Campolongo F., Cariboni J., Saltelli A. (2007) An effective screening design for sensitivity analysis of large models, Environmental Modelling and Software, 22, 1509-1518.
8. Carnevale C., Finzi G., Pisoni E., Volta M., Wagner F. (2012) Defining a nonlinear control problem to reduce particulate matter population exposure, Atmospheric Environment, 55, 410-416.
9. Carslon D. et al. (2004) Large-scale convex optimization methods for air quality policy assessment. Automatica, 40, 385–395.
10. Chakraborty A. and Gupta T. (2010) Chemical Characterization and Source Apportionment of Submicron (PM1) Aerosol in Kanpur Region, India, Aerosol and Air Quality Research, 10, 433–445.
11. Chang J.C., Hanna S.R. (2004) Air quality model performance evaluation, Meteorology and Atmospheric Physics, 87, 167–196.
12. Colvile R.N., Woodfield N.K., Carruthers D.J., Fisher B.E.A., Rickard A., Neville S., Hughes A. (2002) Uncertainty in dispersion modelling and urban air quality mapping, Environmental Science & Policy, 5, 207–220.
13. Cullen, A.C. and H.C. Frey (1999) Probabilistic Techniques in Exposure Assessment. Plenum Press: New York.
14. DEFRA (2010) Evaluating the performance of air quality models. Issue 3, June 2010.
15. Dennis, R., and Coauthors, 2010: A framework for evaluating of regional-scale numerical photochemical modeling systems. Environ. Fluid Mech., in press, doi:10.1007/s10652-009-9163-2.
16. Digar A., Cohan D.S., Cox D.D., Kim B.-U. and Boylan J.W. (2011) Likelihood of Achieving Air Quality Targets under Model Uncertainties, Environ. Sci. Technol., 45 (1), 189–196.
17. Favez O., El Haddad I., Plot C., Boréave A., Abidi E., Marchand N., Jaffrezo J.-L., Besombes J.-L., Personnaz M.-B., Sciare J., Wortham H., George C. and D'Anna B. (2010) Inter-comparison of source apportionment models for the estimation of wood

burning aerosols during wintertime in an Alpine city (Grenoble, France), Atmospheric Chemistry and Physics Discussions, 10, 559-613.
18. Fragkou E., Douros I., Moussiopoulos N. and Belis C. A. (2012). Current Trends in the use of Models for Source Apportionment of Air Pollutants in Europe, International Journal of Environment and Pollution 50 (1-4): 363-375.
19. Fuentes M. (2009) Statistical issues in health impact assessment at the state and local levels, Air Qual Atmos Health, 2(1), 47–55.
20. Fujita E. M., Campbell D. E., Arnott W. P., Chow J. C. and Zielinska B. (2007) Evaluations of the Chemical Mass Balance Method for Determining Contributions of Gasoline and Diesel Exhaust to Ambient Carbonaceous Aerosols, Journal of the Air & Waste Management Association, 57, 721–740.
21. Galmarini S., Douw G.S., Schere K.L., Moran M.D. (2010) Advancing the evaluation of regional-scale air quality models, JRC Scientific and Technical Reports, http://publications.jrc.ec.europa.eu/repository/bitstream/111111111/13563/1/report_galmarini.pdf
22. Gelencsér A., May B., Simpson D., Sánchez-Ochoa A., Kasper-Giebl A., Puxbaum H., Caseiro A., Pio C. A. and Legrand M. (2007) Source apportionment of PM2.5 organic aerosol over Europe: Primary/secondary, natural/anthropogenic, and fossil/biogenic origin, Journal of Geophysical Research D, Atmospheres, 112.
23. Gilardoni S., Vignati E., Cavalli F., Putaud J. P., Larsen B. R., Karl M., StenstrÃm K., Genberg J., Henne S. and Dentener F. (2011) Better constraints on sources of carbonaceous aerosols using a combined 14C-macro tracer analysis in a European rural background site, Atmospheric Chemistry and Physics, 11, 5685-5700.
24. Grosjean D. and Seinfeld J. H. (1989). Parameterization of the formation potential of secondary organic aerosols, Atmospheric Environment, 23(8), 1733-1747.
25. Guariso G., Pirovano G, Volta M. (2004) Multi-objective analysis of ground-level ozone concentration control, Journal of Environmental Management, 71, 25–33.
26. Hanna S.R. (1988) Air Quality Model Evaluation and Uncertainty, JAPCA, 38:4, 406-412.
27. Helton J., Johnson C., Sallaberry C. and Storlie C. (2006) Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliability Engineering and System Safety, 91, 1175–1209.
28. Hogrefe C., Rao S.T. (2001) Demonstrating attainment of the air quality standards: integration of observations and model predictions into the probabilistic framework, J Air Waste Manag Assoc., 51(7), 1060-72.
29. Isukapalli S.S., Roy A., Georgopoulos P.G. (1998) Stochastic Response Surface Methods (SRSMs) for Uncertainty Propagation: Application to Environmental and Biological Systems, Risk Analysis, 18, 351–363.
30. Karagulian F. and Belis C. A. (2012) Enhancing Source Apportionment with Receptor Models to Foster the Air Quality Directive Implementation, International Journal of Environmental Pollution, 50, 190-199.
31. Karagulian F., Belis C. A. and Borowiak A. (2012) Results of the European Intercomparison exercise for Receptor Models 2011-2012. Part I. JRC Scientific and Policy Reports. 94 pp. ISBN 9789279281303.
32. Krzyzanowski M., Cohen A., Anderson R. and the WHO Working Group (2002) Quantification of health effects of exposure to air pollution, Occupational & Environmental Medicine, 59(12), 791-793.

33. Künzli N., Kaiser R., Medina S., Studnicka M., Chanel O., Filliger P., Herry M., Horak Jr F., Puybonnieux-Texier V., Quénel P., Schneider J., Seethaler R., Vergnaud J.-C., Sommer H. (2000) Public-health impact of outdoor and traffic-related air pollution: a European assessment, THE LANCET, 356, 795-801.
34. Larsen R. K. III and Baker J. E. (2003) Source Apportionment of Polycyclic Aromatic Hydrocarbons in the Urban Atmosphere: A Comparison of Three Methods, Environmental Science & Technology, 37 (9), 1873–1881.
35. Martuzzi M., Krzyzanowski M. and Bertollini R. (2003) Health impact assessment of air pollution: providing further evidence for public health action, European Respiratory Journal, 21, 86–91.
36. Moore G.E. and Londergan R.J. (2001) Sampled Monte Carlo uncertainty analysis for photochemical grid models, Atmospheric Environment, 35, 4863–4876.
37. Paatero P. and Hopke P. K. (2009). Rotational tools for factors analytic models, Journal of Chemometrics, 23, 91-100.
38. Paatero P., Hopke P. K., Song X. H. and Ramadan Z. (2002). Understanding and controlling rotations in factor analytic models, Chemometrics and Intelligent Laboratory Systems, 60(1-2), 253-264.
39. Pisoni E., Carnevale C., Volta M. (2009) Multi-criteria analysis for PM10 planning, Atmospheric Environment, 43, 4833-4842.
40. Rao K.S. (2005) Uncertainty Analysis in Atmospheric Dispersion Modeling, Pure appl. geophys., 162, 1893–1917.
41. Saltelli A., Annoni P., Azzini J., Campolongo F., Ratto M., Tarantola S. (2010) Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index, Computer Physics Communications, 181, 259-270.
42. 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.
43. Saltelli A., Tarantola S., Campolongo F. and Ratto M. (2004) Sensitivity Analysis in Practice.
44. A Guide to Assessing Scientific Models, John Wiley & Sons publishers.
45. Seaman, N.L. (2007) Evaluating the performance of meteorological processes within air quality modelling systems. Workshop on the Evaluation of Regional-scale Air Quality Modeling Systems, Research Triangle Park, North Carolina, USA.
46. Stern J. and Flemming R. (2004) Formulation of criteria to be used for the determination of the accuracy of model calculations according to the requirements of the EU Directives for air quality – Examples using the chemical transport model REM-CALGRID, Freie Universität Berlin, Institut für Meteorologie.
47. Subramanian R., Donahue N. M., Bernardo-Bricker A., Rogge W. F. and Robinson A. L. (2007) Insights into the primary-secondary and regional-local contributions to organic aerosol and PM2.5 mass in Pittsburgh, Pennsylvania, Atmospheric Environment, 41, 7414-7433.
48. Thunis P. et al. (2007) Analysis of model responses to emission-reduction scenarios within the CityDelta project, Atmospheric Environment, 41, 208–220.
49. Thunis P., Georgieva E., Pederzoli A. (2012) A tool to evaluate air quality model performances in regulatory applications, Environmental Modelling & Software, 38, 220–230.

50. UNECE (2002) Progress Report Prepared by the Chairman of the Task Force on Integrated Assessment Modelling, United Nations Economic Commission for Europe, Geneva, Switzerland.
51. Viana M., Kuhlbusch T.A.J., Querol X., Alastuey A., Harrison R.M., Hopke P.K., Winiwarter W., Vallius M., Szidat S., Prévôt A.S.H., Hueglin C., Bloemen H., Wåhlin P., Vecchi R., Miranda A.I., Kasper-Giebl A., Maenhaut W. and Hitzenberger R. (2008) Source apportionment of particulate matter in Europe: A review of methods and results, Journal of Aerosol Science, 39, 827-849.
52. Vlachokostas C. et al. (2009) Decision support system for the evaluation of urban air pollution control options: Application for particulate pollution in Thessaloniki, Greece, Science of the Total Environment, 407, 5937–5948.
53. Werner S., Samaali M. and J.-L. Ponche ; 2005. Determination of uncertainties in emissions inventories at regional scale : application to the ESCOMPTE emission inventory. Marseilles Final ESCOMPTE Workshop Feb. 2-4, 2005 , Marseille France.
54. Werner, S. (2009) Optimisation des cadastres d'émissions : estimation des incertitudes, détermination des facteurs d'émissions du « black carbon » issu du trafic routier et estimation de l'influence de l'incertitude des cadastres d'émissions sur la modélisation : application aux cadastres ESCOMPTE et Nord-Pas-de-Calais. Thèses de doctorat, Université de Strasbourg.
55. Yarwood G., Wilson G. and Morris R. (2005) Development of the CAMx Particulate Source Apportionment Technology (PSAT), ENVIRON Final Report. Available at: http://www.ladco.org/reports/rpo/modeling/camx-psat.pdf

ANNEX I : THE QUESTIONNAIRE REGARDING UNCERTAINTY AND ROBUSTNESS (QA/QC)

TOPIC 5: Uncertainty and robustness, including QA / QC

1. **What type of evaluation methodology was applied?**

Diagnostic
Operational
Dynamic
Probabilistic
Expert judgment
Other

please, provide a reference

2. **Did you use already available software for your evaluation?**

2.1. Air quality modelling
- ◯ No
- ◯ Yes

2.2. Source receptor relationships
- ◯ No
- ◯ Yes

2.3. Source apportionment
- ◯ No
- ◯ Yes

2.4. Health impact assessment
- ◯ No
- ◯ Yes

2.5. IA system
- ◯ No
- ◯ Yes

3. **Did you explicitly address uncertainty in the current activity or is the uncertainty evaluation based on previous works?**
- ◯ This project
- ◯ Previous

4. **What type of uncertainty quantification was applied?**
- ◯ Global
  - ◯ Elementary Effects
  - ◯ Variance-Based methods
  - ◯ Factor Mapping and Metamodelling
  - ◯ Other
- ◯ Local

- ○ OaT
- ○ Other

5. Was model uncertainty assessed in its different components?

- ○ Yes
- ○ No

6. If yes, what did you study specifically?

6.1. AQ modelling

- ☐ Emissions
- ☐ Meteorology
- ☐ AQ Boundary conditions
- ☐ AQ Model formulation (atmospheric dynamics and chemistry, numerical solutions, choice of modelling domain and grid structure)
- ☐ none

6.2. Source-receptor relationships

- ☐ Emissions
- ☐ Meteorology
- ☐ Formulation
- ☐ none

6.3. Source apportionment

- ○ Yes
- ○ No

6.4. Health impact assessment

- ○ Yes
- ○ No

6.5. IA system

- ○ Optimization algorithms
- ○ Planning indicators
- ○ None

7. Open questions

7.1. What were the main performance/quality indicators:

7.2. How did you judge that your model is good enough for air quality planning?

7.3. How did you judge that your model is good enough for air quality assessment?

7.4. What would be the point that you would like to improve with respect to this project?

7.5. How do you interconnect the different uncertainties of IAM?

7.6. If you are not applying any evaluation/uncertainty assessment techniques, why?

ANNEX II : GLOSSARY OF TERMS

Diagnostic Evaluation refers to a process-oriented analysis to determine whether the individual physical and chemical processes are correctly represented in the model (Dennis et al., 2010).

Dynamic Evaluation involves the analysis of model responses to changes in model input data, such as source emissions or meteorological conditions (Dennis et al., 2010). Sensitivity analysis is most commonly applied within the frame of the dynamic evaluation.

Probabilistic Evaluation: This type of evaluation is performed on the basis of methods such as model inter-comparison and ensemble modelling, and attempts to capture statistical properties, including uncertainty or level of confidence in the model results, for regulatory model applications (Dennis et al., 2010). This approach requires knowledge of uncertainty imbedded in both model predictions and observations.

Operational Evaluation involves assessment of model results compared with monitored data, which may include routine or field campaign observations of ambient pollutant concentrations, emissions, meteorology, and other relevant variables (Dennis et al., 2010).

Global uncertainty quantification methods: these uncertainty methods are based on exploring the space of the input factor, according to the consideration that it is possible to select a set of data points that are more informative and robust than derivative values estimated at a single data point at the centre of the space. A number of multiple choices are given for global uncertainty methods.

- Elementary Effects: A sensitivity analysis method based on calculating for each input a number of incremental ratios, called Elementary Effects (EE), from which basic statistics are then computed to derive sensitivity information (Campolongo et al. 2007).
- Variance-Based methods: A sensitivity analysis method that involves the decomposition of the total output variance into the contributions of the input factors. The aim is to compute “first and “total order sensitivity indexes” (Saltelli et al., 2010).
- Factor Mapping and Metamodelling: These methods may be used in analyzing when a particular model provides results of sufficient quality in certain ranges (Helton et al., 2006).
- Other: other global uncertainty analysis methods not fitting into the above categories, or a combination of methods

Local uncertainty quantification methods: these methods require a limited number of simulations and are less accurate than global one. One example of such methods is:

- OaT (One-at-a-Time) is one of the simplest and most common approaches as it involves changing one factor in each model simulation, to see what effect this produces on the output.

Optimization algorithms: are tools to automatically determine the best alternative(s) when one or more performance criteria have been formulated in mathematical terms.

Planning indicators for human, ecosystems and materials exposure. The decision problem in Integrated Assessment Modelling determines the abatement measures or other actions that improve the objectives, and comply with the physical, economical and environmental constraints. Objectives and environmental constraints are typically indicators of human, ecosystems and material exposure.

ANNEX III: Statistical performance indicators calculated by DELTA tool

Mean	$\overline{M} = \frac{1}{N}\sum_{i=1}^{N} M_i \;,\; \overline{O} = \frac{1}{N}\sum_{i=1}^{N} O_i$
Standard Deviation	$\sigma_M = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - \overline{M}\right)^2} \;,\; \sigma_O = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}$
Mean Bias	$MBias = \frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right)$
Mean Fractional Bias	$MFB = \frac{1}{N}\sum_{i=1}^{N}\frac{M_i - O_i}{(M_i + O_i)/2}$
Mean Fractional Error	$MFE = \frac{1}{N}\sum_{i=1}^{N}\frac{\lvert M_i - O_i\rvert}{(M_i + O_i)/2}$
RootMeanSquare Error	$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right)^2}$
Ratio of Systematic and unsystematic RMSE	$RMSE_S / RMSE_U = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\hat{M}_i - O_i\right)^2} \Big/ \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - \hat{M}_i\right)^2}$ *where* $\hat{M}_i = a + bO_i$ *are the regressed model values, estimated from a least square fit to observations;* $RMSE^2 = RMSE_S^2 + RMSE_U^2$.
Target	$RMSE / \sigma_O = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(M_i - O_i\right)^2} \Big/ \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}$
Pearson Correlation Coefficient	$R = \sum_{i=1}^{N}\left(M_i - \overline{M}\right)\cdot\left(O_i - \overline{O}\right) \Big/ \sqrt{\sum_{i=1}^{N}\left(M_i - \overline{M}\right)^2}\cdot\sqrt{\sum_{i=1}^{N}\left(O_i - \overline{O}\right)^2}$
Index of Agreement	$IOA = 1 - N\cdot RMSE^2 \Big/ \sum_{i=1}^{N}\left(\lvert M_i - \overline{O}\rvert + \lvert O_i - \overline{O}\rvert\right)^2$
Relative Directive Error and its maximum	$RDE = \frac{\lvert O_{LV} - M_{LV}\rvert}{LV}$ *where* O_{LV} *is the closest observed concentration to the limit value concentration (LV) and* M_{LV} *is the correspondingly ranked modelled concentration.* *MRDE=Max (RDE over 90% of stations)*
Relative Percentile Error and its maximum	$RPE = \frac{\lvert O_p - M_p\rvert}{O_p}$ *where* p *is the percentile corresponding to the allowed number of exceedances of the limit value* *MRPE=Max (RPE over 90% of stations)*
Factor of modelled values within a factor of two of observations	$FAC2 = \frac{1}{N}\sum n_i \quad with \quad n_i = \begin{cases} 1 \text{ for } 0.5 \le \lvert M_i/O_i\rvert \le 2 \\ 0 \text{ else} \end{cases}$
Centred Root Mean Square error	$CRMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[\left(M_i - \overline{M}\right) - \left(O_i - \overline{O}\right)\right]^2}$
Model Efficiency Score	$MEF = 1 - RMSE^2$