IDŐJÁRÁS
Quarterly Journal of the Hungarian Meteorological Service
Vol. 119, No. 3, July – September, 2015, pp. 355–378

Regional air quality models and the regulation of atmospheric emissions

Bernard E.A. Fisher *[1], **Charles Chemel** [2], **Ranjeet S. Sokhi** [3], **Xavier V. Francis** [4], **Keith J. Vincent** [5], **Anthony J. Dore** [6], **Stephen Griffiths** [7], **Paul Sutton** [8], and **Raymond D. Wright** [9]

[1] *Retired from Environment Agency, Reading, UK beafisher@cantab.net*
Permanent address: Little Beeches, Headley Road, Leatherhead KT22 8PT, UK

[2] *National Centre for Atmospheric Science, Centre for Atmospheric and Instrumentation Research, University of Hertfordshire, Hatfield, UK c.chemel@herts.ac.uk*

[3] *Centre for Atmospheric and Instrumentation Research, University of Hertfordshire, Hatfield, UK r.s.sokhi@herts.ac.uk*

[4] *Centre for Atmospheric and Instrumentation Research, University of Hertfordshire, Hatfield, UK X.Francis@herts.ac.uk*

[5] *Ricardo-AEA, Harwell, UK Keith.Vincent@ricardo-aea.co.uk*

[6] *Centre for Ecology and Hydrology, Penicuik, UK tony.dore@ceh.ac.uk*

[7] *E.ON Technologies, Ratcliffe-on-Soar, UK Stephen.Griffiths@eon.com*

[8] *RWE npower, Swindon, UK paul.sutton@rwenpower.com*

[9] *Spacefleet Ltd, UK rdw@spacefleet.co.uk*

**Corresponding author*

(Manuscript received in final form October 13, 2014)

Abstract—This paper investigates regional air quality model performance and the regulation of atmospheric emissions. Although evaluation of regional models cannot be reduced to a set of rules, the paper shows ways of developing better understanding of model performance. It draws on studies in recent years by the Environment Agency to quantify the uncertainty in predictions of regional air quality models. It is argued that a decision by a regulator on how to use a regional air quality model should be based on both operational evaluation (involving comparison with observation) and diagnostic evaluation (for developing understanding of the model), using operational and diagnostic metrics. Operational and diagnostic evaluations were undertaken, using a 'constructor' (CMAQ) and a 'seer' type (TRACK-ADMS) regional air quality model, for the secondary pollutants PM_{10} , $PM_{2.5}$ and ozone, though for episodic ozone it was not possible to define an appropriate performance metric.

Neither type of model showed clearly better performance when applied to long-term average concentrations. There was not enough information to set a minimum margin of error in operational evaluations but margins of 20% or more are to be expected. Unlike operational metrics there is no obvious way of deriving diagnostic metrics. However a footprint diagnostic metric was shown to be a way to reveal the behaviour of PM_{10} and $PM_{2.5}$ in both types of model. It is therefore suggested that seer models are used to reveal the structure of a model's underlying mathematical equations from which diagnostic metrics can be formed.

In the absence of an objective basis for setting acceptance criteria for models, it is proposed that the underlying pragmatic principle should be to use whatever has comparable accuracy with the best existing international practice. For regulatory applications, the error expected in current types of air quality models should be a consideration in any decision made on the basis of models.

Key-words: Inter-comparison, regional air quality, model, footprint, metric, diagnostic, operational, evaluation, seer, constructor

1. Introduction

This paper addresses the question "When is a regional air quality model good enough to be the basis for making a decision about emission reductions to meet limit values?", accepting that the use of an environmental model may only be part of the decision making process. Model evaluation studies involve selecting appropriate metrics or diagnostics (parameters summarising key aspects of the behaviour of a model), and showing that the model can predict the metrics with appropriate accuracy compared with observations.

Regional model development has made considerable progress in recent decades and complex air quality models are essential for assessing secondary pollutants, such as wet deposition, ozone and particulate matter. However just because regional models are the only tools for assessing secondary pollutants, such as ozone and particulate matter, this does not ensure that they are adequate to make decisions about emission reductions.

The paper draws on studies in recent years by the Environment Agency to quantify the uncertainty in predictions of regional air quality models, the latest of which is the CREMO, Comparison of Regional Models, project. The results of the project are described in a number of reports (*Derwent*, 2013; *Fisher*, 2013; *Hayman et al.*, 2013a, 2013b, 2013c, and 2013d). These are used to draw conclusions about whether regional air quality models describing atmospheric concentrations over some 100's to a few 1000 kilometres, are an adequate basis for making decisions about emission reductions to meet environmental criteria.

2. General discussion of environmental models

The way environmental models have been developed and used has been discussed in the literature (*Edwards*, 2010). One can refer to attempts to ensure that models are used properly (*Pilkey-Jarvis* and *Pilkey*, 2008). Air quality models include parameters which are assigned values for the problem in hand, to distinguish them from dependent variables, such as pollutant concentration or deposition. Parameters are usually mathematical functions of coefficients and dependent variables, representing a physical process.

It is rare for there to be sufficient observational data to be able to test exhaustively the behaviour of an environmental model. More often than not for policy applications, one is interested in the behaviour of an environmental system under conditions which have never occurred in nature and one is therefore investigating model scenarios which cannot be explicitly tested. In such situations the qualitative behaviour of the system described by the mathematical model is the only realistic goal. The environmental model can generally be approximated by a set of time dependent non-linear ordinary differential equations. In regional air quality models the synoptic meteorological conditions describe the motion of air masses crossing the main source regions. Under steady synoptic conditions regional pollutant concentrations may build up or decay. Assuming steady meteorological conditions persist for long enough, the solution of the set of mathematical equations tends in time towards a stationary point, though in nature the long time limits are not necessarily reached as meteorological conditions are never steady. However the behavior at stationary points under such idealistic conditions reveals something about the structure of the mathematical system, even if such conditions do not occur in reality. Of special interest are the specific parameter values at points where the qualitative behaviour changes, say from a tendency to decay from initial values to the growth in concentration (see later for a suspected example for ozone of a bifurcation). For this reason two broad categories of model can be distinguished: (1) those which reveal the underlying structure of the mathematical system, and (2) those which try to emulate the full complexity of the environmental system.

Edwards (2010) describes these two broad modelling approaches in relation to climate modelling: in type (1) *model seers*[1] use models to understand and explore the climate system with emphasis on its sensitivity to changing variables and processes. In type (2) *model constructors* seek to capture the full complexity of the climate system, which can then be used for various applications, promoted by the power of modern computing. Constructors seek to include more realism, including all known physical processes that influence the climate. Seers tend to focus on modelling the most fundamental and understood processes and to use a variety of models. The 'state of the art' for seers depends

[1] Definition. A *seer* is one who has insight. The common implication that a seer can predict the future is not relevant in the context of this paper.

on the model application. For constructors a single ‘state of the art’ model exists, which uses the most up to date observations as input and evaluation, and contains the most detailed and physically realistic parameterisations. This does not separate good from bad modelling practice but illustrates two different strategies for extending knowledge. Seers are generally interested in simple models, which promote understanding, while constructors ignore simple models and focus on complex comprehensive models. Regulators prefer a single model which can be regarded as ‘fit for purpose’ for making a policy decision, but one conclusion from this paper is not to rely on such a narrow approach to regional modelling.

3. Advances in air pollution assessment

Regulators, such as the Environment Agency in the U.K., need to know whether a model can be used for decision making and this discussion has prompted model comparison exercises, starting with the comparison of dispersion models (*Hall et al.*, 2000a, 2000b) describing concentrations in the near-field out to 20 km from a source, and relatively simple acid deposition assessment models (*Abbott et al.*, 2003), where the model inter-comparison was used to gain understanding of model uncertainty. This paper focuses on the secondary pollutants, particulate matter (PM)[2], both PM_{10} and $PM_{2.5}$ and ozone (O_3). The last decade or so has seen an enormous increase in the sophistication of computer programs, making calculations of secondary pollutants more accessible. The practical implementation of the US Environmental Protection Agency (USEPA) supported CMAQ (Community Multi-scale Air Quality) modelling system, which follows the constructor approach, has been notable. Documentation on CMAQ is available from the official CMAQ website (http://www.cmaq-model.org) [accessed 21 May 2013]. Regulators therefore need to be able to assess the capabilities of different types of air quality models with different levels of sophistication. This has promoted the constructors' approach. However, just because regional models are the only way of assessing secondary pollutants, this does not mean that they are good enough for regulation (*Pilkey-Jarvis* and *Pilkey*, 2008).

Secondary $PM_{2.5}$ is generated on a regional scale, so that the $PM_{2.5}$ concentration is a mixture of local and regional components. Regional models are the only way of assessing the impact of individual sources on the regional component of the $PM_{2.5}$ concentration and of estimating the population exposure. The contribution from local sources is not included within the regional component, because only concentration variations over regional scales, typically

[2] PM_{10} particulate matter less than or equal to 10 microns (50% cut off);
$PM_{2.5}$ particulate matter less than or equal to 2.5 microns (50% cut off).

5 km or more, are treated in regional models such as CMAQ. A correction factor to take account of local sources needs to be included.

Regulators face a problem when applying complex, constructor models, such as CMAQ. In principle it is desirable that the results of a calculation can be replicated. However this becomes increasingly difficult as models become more complex. The detailed configuration may be difficult to set up on different computer platforms and decisions about which input data sets and options to apply become complicated to document.

Comparison of model predictions against observations is known as 'operational' evaluation. 'Diagnostic' assessment involves understanding the behaviour and response of a model system (*Dennis et al.*, 2010). Following the broad model categorisation described in Section 2, seers tend to consider 'diagnostic' evaluation, while constructors tend to focus mainly on the 'operational' evaluation of models. The operational performance of different types of regional models has been compared in the Model Evaluation Exercise for the UK Department for Environment, Food and Rural Affairs (Defra) (*Carslaw* 2011a, 2011b, and 2011c) and the Air Quality Modelling Evaluation International Initiative, AQMEII, (http://aqmeii.jrc.ec.europa.eu/ [accessed 21 May, 2013]). *Dore et al.* (2013) presented results of the Defra Model Evaluation Exercise which encompassed a wider set of models than comparisons described in this paper. Phase one of the AQMEII intercomparison (*Solazzo et al.*, 2012) involved 10 regional models. Four models were applied to both the European and North American domains; five models were applied to just the European domain and one model was applied to the North American domain. In the second phase of AQMEII (the Air Quality Model Evaluation International Inter-comparison) the earlier regional model inter-comparison assessment was extended to on-line air quality models in which the air quality and meteorological models were coupled together. The PM_{10} and $PM_{2.5}$ concentrations simulated by eight on-line coupled models, run by seventeen independent groups from Europe and North America, were compared with each other and with observations (*Im et al.*, 2014).

In HTAP, the taskforce on the Hemispheric Transport of Air Pollutants, modelling inter-comparison, the predictions of the change in surface ozone as a result of continental scale changes in emissions were compared in 14 global chemical transport models. In this case *Wild et al.* (2012) did not use a seer model to understand the result, but rather used a simple scaling model to represent the source-receptor relationships found from the results in the full models. One global transport model was used to investigate the range of emission changes over which this simple, linearised model is accurate enough for practical, policy applications.

Generally the constructionist models used in the CREMO, AQMEII and HTAP inter-comparisons are too complex for it to be possible to diagnose which factors have greatest influence on a model's performance. For particulate matter, model prediction was disappointing suggesting it is difficult to include all factors affecting concentrations.

Seer and constructor models have been used to evaluate policy decisions in the areas of air quality and acid deposition. Examples of a seer model in the UK include the semi-empirical TRACK-ADMS model, which is much simpler than CMAQ (*Abbott et al*. 2003; *Abbott et al.,* 2007) and was used to show that the average PM_{10} concentration would have been 3 μg m^{-3} higher in 2005 without the industrial policy intervention of the Environment Agency between 1990 and 2005 (*Vincent* and *Abbott,* 2008). Using the more complex, constructor model CMAQ, *Chemel et al*. (2014) estimated that the major industrial sources in the UK made up between 10 to 20% of the $PM_{2.5}$ concentration depending on location in 2006.

The type of a model used at the policy stage may not be the same as that used in the scientific stages of model evaluation. Instead it could involve model emulation.[3] As an example of the emulation methodology, a statistical emulation of the moderately complex constructor model, The Air Pollution Model (TAPM)[4] was used to estimate the consequences on human health, expressed as the number of life years lost, for one year's emission from a coal-fired power station (*Fisher et al.,* 2010). A range of emission conditions were evaluated without re-running TAPM for every set of conditions.

Simple, order of magnitude estimates, sometimes called 'back of envelope' calculations, can be considered to be a type of seer model. Provided they predict approximately similar responses to input changes as more complex, constructor models there is justification for using them in integrated assessments of air quality policy. This does not mean that the constructor models should be neglected by regulators and left to specialists. The constructor model provides the essential test bed for exploring understanding of the response of a model to changes in input over a wide range of conditions. Thus for regulators both the model *seer* and the model *constructor* approaches have important roles.

---

[3] For an explanation of an emulator, see http://mucm.aston.ac.uk/MUCM/MUCMToolkit/index.php?page=MetaOverviewEmulators.html [accessed 21 May, 2013]. If a computer simulation is computationally expensive, so that evaluating a constructor model for a choice of inputs takes a significant amount of computing time, one may be limited to evaluating the model at a small number of different input data sets. However one may want to know model predictions at a large number of different input values. One can deal with this problem by building an emulator: a statistical model of the model, constructed from a fairly small number of runs of the constructor model. The emulator will predict both output values, and report uncertainty in any prediction. In the Managing Uncertainty in Computer Models (MUCM) project, a toolkit for constructing Gaussian process emulators is described in which all parameter values and their interactions have uncertainty described by Gaussian functions. Gaussian functions possess convenient properties, making emulator formulation easy.

[4] The moderately complex meso-scale model, TAPM, was applied with UK emissions and meteorology in 2003 to illustrate the approach. The use of TAPM was an efficient way to generate a set of results distributed through parameter space. It would not have been practical at the time to have generated enough CMAQ results to build an emulator.

4. Model evaluation protocols and metrics

In the CREMO project, different versions of the constructor model CMAQ were compared with three air quality models developed in the UK which may be regarded as seer models: the Fine Resolution Atmosphere Multi-Pollutant Exchange (FRAME) model for acid deposition (*Dore et al*. 2007), the Trajectory model with Atmospheric Chemical Kinetics–Atmospheric Dispersion Modelling System (TRACK-ADMS) model for annual audits (*Abbott et al.,* 2007) and the Ozone Source–Receptor Model (OSRM) for ozone (*Hayman et al*., 2010). An Environment Agency report (*Hayman et al*., 2013a) lists the sites, measurements and methods available in the UK for evaluating models against observations, as well as summarising the models. The aim was to establish whether the models met acceptance criteria as part of an evaluation protocol and not to establish which model had the best performance judged against observations.

Defra (*Derwent et al*., 2010), and informally the USEPA (*Dennis et al*., 2010), have published protocols for operational model evaluation using familiar statistical measures. Benchmarking procedures (*Thunis et al*., 2011a) have also been produced. The optimal set-up (the set-up requiring the lowest computer resource and least preparation) for running constructor models efficiently over a year or more, to answer a policy question or a regulatory issue, cannot be specified precisely. The process involves setting up (1) emissions, (2) initial and boundary conditions, (3) running a meso-model to determine the meteorological fields, and then (4) running a chemical transport model, such as CMAQ. Although a model complies with an operational evaluation protocol, the model should not be used under emissions or meteorological conditions which have not occurred during the testing of the model.

The acceptance criteria set in the Ambient Air Quality Directive (*EC,* 2008) also set operational performance limits and are of great importance, because they impose compliance requirements on EU countries. The Directive allows models which satisfy the criteria to be used for air quality assessment to reduce the number of sampling sites, to prepare plans and abatement measures and to determine where the pollution is coming from. The guidance published by *FAIRMODE* (2011) and the guidance on NO_2 (*Denby,* 2011) interpret the Directive, elaborating on text in the Directive, such as 'relative directive error', but also describing its limitations.

In the Defra Model Evaluation Protocol (*Derwent et al*., 2010) predictions of the model should be accepted if the percentage of model predictions within a factor of two (*FAC2*) of the observations is greater than 50 per cent. The normalised mean bias (*NMB*) is defined as:

$$NMB = \frac{\sum_{i=1}^{N} M_i - O_i}{\sum_{i=1}^{N} O_i}, \tag{1}$$

where N is the number of observations, M_i are the calculated values, O_i are the observed values. The *NMB* should satisfy $-0.2 \leq NMB \leq 0.2$ in the Defra Model Evaluation Protocol. The *NMB* puts a higher weighting on model performance at higher concentrations, a distortion, but one which might be reasonable, given the greater concern over the occurrence of high concentrations. It would also be important to consider carefully the quality of the observations and the size of the sample N, the number of observed–calculated pairs.

A comparison of seer regional transport models (*Abbott et al.,* 2003) suggested that these simpler types of models could meet the *FAC2* criterion when calculating acid deposition. The set of possible statistical measures used to evaluate the constructor model CMAQ (*Chemel et al.,* 2010) included the percentage within a factor of two (*FAC2*) and the normalised mean bias (*NMB*). As part of the model intercomparison within CREMO (*Chemel et al*., 2011), all the models considered (CMAQ v4.6, CMAQ v4.7, TRACK-ADMS and FRAME) were shown to satisfy the *FAC2* criterion, that 50 per cent of the modelled results should be within a factor of two of the annual mean concentrations for *all* the species considered. In contrast, none of the models satisfied the criterion that the normalised mean bias *NMB* should be in the range –20 to 20 per cent, for *all* the species considered, which included PM_{10}.

An *acceptance* criterion for ozone could refer to the annual average ozone concentration or to the peak ozone level during episodes. The annual average ozone depends largely on domain boundary conditions and removal within the domain, while episodic ozone concentrations rely on regional generation within the domain. It is therefore necessary to distinguish between an acceptance criterion for shorter periods, such as episodes, and an acceptance criterion for annual average ozone. There is no commonly accepted agreement over the best choice of metric in the operational evaluation of regional ozone models. There are examples of diagnostic ozone metrics in the review by *Middleton et al.* (2007), while *Rao et al*. (2011) argue in favour of the seasonal average as the most suitable operational metric.

5. Diagnostic evaluation

5.1. Footprint metric

A systematic procedure for setting a diagnostic evaluation protocol cannot be defined as this involves understanding model behaviour. A regulator is primarily interested in attributing concentrations to emissions. Thus diagnostic evaluation in the CREMO project was focused on the differences in the footprint of sources calculated by different regional models used for regulation. A 'footprint' metric is a response function, showing how concentrations or deposition are influenced spatially by emissions from a single specified source, such as a power station. Footprints are obtained from the difference between the concentration or deposition, when all sources are included, and the concentration or deposition, when all sources, except for the specified source under consideration, are included. In the seer air quality models discussed in this paper, the individual footprints can be calculated directly. The footprint of secondary pollutants, such as particulate matter, depend directly on model parameters determining the formation by chemical reactions and removal by wet and dry deposition, which eventually lead to complete removal from the atmosphere after a number of days. In constructor models, such as CMAQ, the footprint determined after much computation appears to be distantly related to the original equations although the same removal processes influence behaviour. Consideration of the footprint metric can therefore link the performance of a seer type of regional model to the constructor type of model.

A footprint metric has two purposes: (a) it is a diagnostic of a seer or constructor model type, showing the change in concentration from the emission at a point source leading to understanding of the overall system behavior; and (b) for regulatory purposes, it shows how emission reductions may change concentrations.

A quantitative approach to evaluating footprints is to consider the distance dependent structure of the footprint of a single source (*Fisher et al.*, 2011). The weighted average concentration given by the average concentration along a typical trajectory, excluding dilution arising from horizontal dispersion, is obtained by multiplying the concentration by the distance from the source. This footprint metric is defined by:

$$\frac{r}{2\pi}\int_0^{2\pi} C(r,\theta)d\theta \tag{2}$$

where $C(r, \theta)$ is the concentration at a distance r from the specified source in a direction θ.

In CREMO the footprint profiles were normalised by the value of the near-field concentration, 30 km from the source, and this is illustrated in the profiles in the smaller figures in *Fig. 1*. The dependence of the PM_{10} and $PM_{2.5}$ concentrations along a radial trajectory does not decrease rapidly with distance because of the gradual production of secondary aerosol in the atmosphere. A numerical diagnostic can be defined as the distance between the point source and the point at which the radial average secondary PM_{10} or $PM_{2.5}$ concentration takes its maximum value. This diagnostic summarises the influence of the source on secondary aerosol formation.

The average dependence of the concentration with distance shown in *Fig. 1* (Eq. (2) without the distance r in the numerator) does decay rapidly with distance, because the average concentration contains a factor proportional to the inverse of the distance from the source ($1/r$) arising from the spread of air mass trajectories.

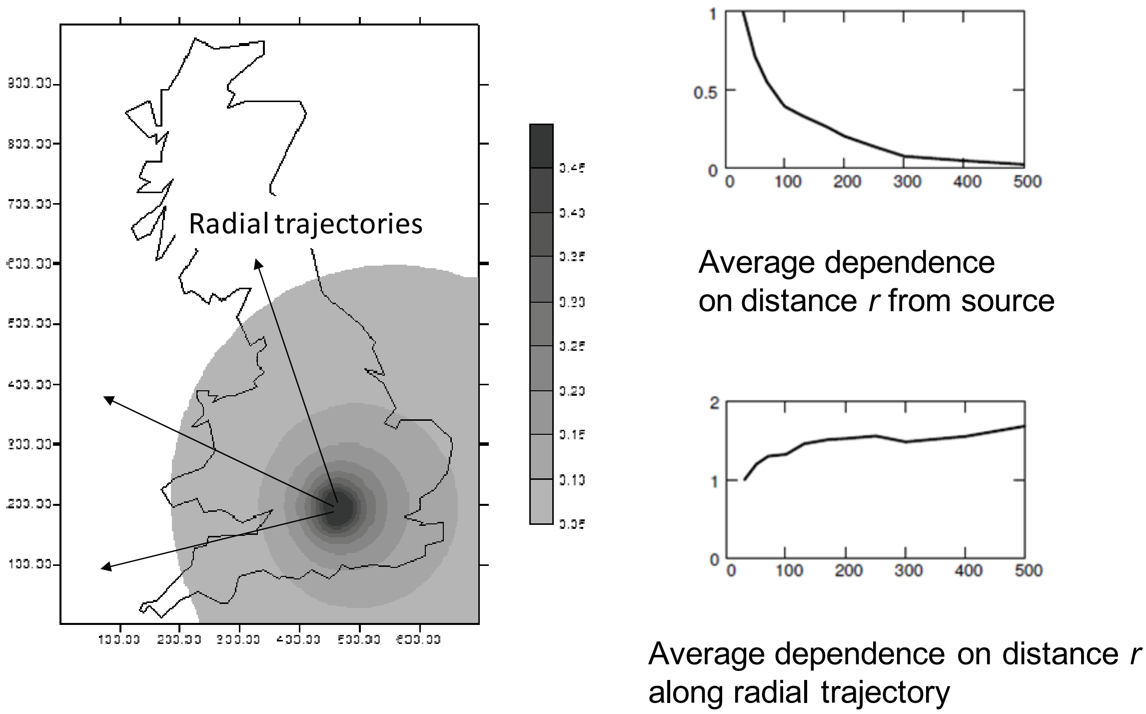

Fig. 1. Schematic illustration of the annual average PM_{10} concentration footprint in μgm^{-3} from a major stationary point source such as a power station. The main picture shows contours of the concentrations of PM_{10} in μgm^{-3}, with distances along the axes in kilometres. The *top* small right hand figure shows the concentration along a typical, radial trajectory as a function of distance in kilometres, normalised by the value of the near-field concentration 30km from the source. The *bottom* small right hand figure shows the average weighted concentration along a radial trajectory, excluding dilution arising from horizontal dispersion by multiplying the concentration by the distance from the source (Eq. (2)).

When making policy decisions, a regulator should understand the underlying structure of the model used in the assessment, for which the first step is to understand how the model responds to emission changes. As the next step in diagnostic evaluation using footprints, a regulator needs to know whether footprints can be aggregated to assess a source control strategy. This means that if Δq_i is the reduction in the source strength of the ith source under the control strategy, the reduction in concentration under the control strategy is approximately equal to ΔC, where:

$$\Delta C = \sum_i \Delta q_i G_i \tag{3}$$

and G_i is the footprint of the ith source of unit source strength. If Eq. (3) is a reasonable approximation, a reduction strategy can follow a sequential change in emissions, tracking in order which emission reductions are most effective.

5.2. Episode metric

The occurrence of photochemical episodes, when high ozone concentrations are generated, suggests the need to be aware of a variety of possible behaviours. Under steady meteorological conditions, ozone would be expected to tend towards a limit, which is the long-term solution of a system of ordinary differential equations for a given choice of VOC emission strengths. Under some conditions, corresponding to lower VOC emission rates under strong winds and high atmospheric dilution, the ozone limit value would be expected to be close to the initial ozone concentration close to the background ozone concentration. This represents an approximate balance between the production and destruction of ozone over the domain. Under other conditions, corresponding to higher VOC emission rates under light winds and low dilution, the limit ozone value may be far from the initial state, corresponding to the build-up of ozone in an anticyclonic episode. The large difference in the limiting ozone concentration between these two emission and meteorological situations corresponds to a bifurcation in the steady-state behaviour of the differential equations describing the ozone system. This does not imply any discontinuity in ozone concentration, but differences in the limit values of the solution of the underlying differential equations. A large change in the limiting ozone concentration occurs for a small change in emissions near a bifurcation, so a sensitivity analysis would break down near the bifurcation point.

The occurrence of ozone episodes will to some extent influence the long-term average ozone concentration, although the occurrence of episodes may vary considerably from year to year depending on weather conditions. Ozone formation can be attributed to specific sources during an episode using constructor models. An ozone footprint during an episode has been calculated using the CMAQ model (*Yu et al.*, 2008) by investigating changes in ozone formation when the emissions from a specified source are altered. The 'integrated downwind ozone production' (IDOP) described by *Derwent* and *Nelson* (2003) is a footprint metric during an episode. IDOP describes how much ozone is produced in the downwind environment under ideal ozone producing conditions by each VOC species of a specified stationary source from selected runs of the Photochemical Trajectory Model (*Derwent et al.*, 2009), which one could class as a seer model. Using IDOP as a metric is a convenient, precautionary approach to regulation when screening VOC emissions. In cases when the IDOP approach suggests that an emission is not acceptable, more complex models of a seer type, such as OSRM, or of a constructor type, such as CMAQ, need to be applied to test the conclusion. An alternative metric to IDOP is whether a 30% reduction in emissions of VOC or NOx suppresses the formation of an ozone episode.

5.3. Diagnostic evaluation of individual processes

Another kind of diagnostic evaluation involves the behaviour of individual processes within a model, although it is not always possible to compare individual processes exactly when they are embedded within a complex model. When *Derwent* (2013) compared six chemical mechanisms (CRI, CB-05, CBM-4, SAPRC-99, SAPRC-07 and OSRM) within the Photochemical Trajectory Model (PTM), he found that the differences in the details of the chemical reaction scheme are not the most important factor in explaining differences between the predictions of ozone.

The Integrated Process Rate analysis of the CMAQ model undertaken by *Francis et al*. (2011) was used to understand the causes of an episode of high ozone over south-east England in 2003. The contributions of cloud processes, chemical processes, advection, diffusion, vertical advection, vertical diffusion, horizontal advection and dry deposition, at different model heights were considered. For this episode, in a south-east England domain, meteorological processes were shown to have the greatest influence.

6. Examples of operational evaluation

In this section some examples of the operational evaluation of constructor and seer models are described. The seer model TRACK-ADMS was developed to enable the Environment Agency to assess contributions from major industrial sources (*Abbott* and *Vincent* 2007; *Abbott et al*., 2003; *Vincent* and *Abbott,* 2008). TRACK-ADMS assumes simplified meteorology to calculate the long-term average atmospheric concentration. It contains some degree of data assimilation which improves predictions (by applying a bias correction). The model is suited to calculating the contributions from large industrial sources and has been subject to uncertainty analysis by undertaking a Monte Carlo analysis of the variation in output over plausible ranges of input parameters. Its limitation is that one cannot be sure that the choice of parameter input values used to evaluate the model, is appropriate in future emission scenarios when emissions and boundary conditions over the model domain may be very different from the ones used to test the model.

Table 1 from *Chemel et al.* (2011) shows an example from the CREMO project of calculating PM_{10} concentrations across the UK, comparing the performance of the CMAQ model, with a resolution set of 5 km, against the model TRACK-ADMS, with a resolution of 1 km. Observations from some 40 rural and background sites in the UK Automatic Urban and Rural Network

(AURN) are used in the evaluation, for which no correction for local sources is applied. Many more parameters need to be specified in the CMAQ model compared with TRACK-ADMS leading to more opportunities for errors to arise in the input data.

Table 1. Comparison of performance in predicting annual average PM_{10} concentration at rural AURN sites in the UK in 2003, for which the local contribution should be minimal, for two versions of a constructor model, and a seer model

Model metric PM_{10} for 2003	**CMAQ v4.6**	**CMAQ v4.7**	**TRACK-ADMS**
FAC2 (%)	88.2	100.0	100.0
*r (*correlation coefficient)	0.09	0.0	0.45
NMB	–0.33	–0.09	–0.20
Single power station contribution (%)	0.34	0.28	0.28

Based on the *NMB* metric, no systematic difference can be seen between the performance of the two versions of the CMAQ model, of the constructor type, and that of the simpler, seer model TRACK-ADMS. A negative *NMB* bias implies a model under-prediction and so a margin of error should be included in model predictions. Results in *Table 1* suggest that the margin of error for PM_{10} is about 20%.

There is not sufficient observational data in 2006 in the UK on which to test the performance of regional model predictions of $PM_{2.5}$. However the CMAQ model has also been used to simulate air quality over North America and Europe for the year 2006 (*Appel et al.,* 2012) as part of the AQMEII project (*Galmarini* and *Rao,* 2011). *Table 2* shows the seasonal, domain averaged, normalised mean biases (*NMB*) of daily average $PM_{2.5}$ concentrations from the CMAQ model for the North American Air Quality System network and the European AirBase network in the year 2006. As far as possible, the CMAQ model configurations were similar for North America and Europe, with both simulations utilising version 4.7.1 of CMAQ. The North American simulation used 34 vertical layers and a 12 km horizontal grid spacing, while the European simulation used 34 vertical layers and an 18 km horizontal grid spacing covering most of Europe. The overestimate of $PM_{2.5}$ at North American sites is thought to arise from an overestimate in the unspeciated $PM_{2.5}$ mass (*Appel et al.,* 2012), which makes up a significant proportion of the $PM_{2.5}$ mass in version 4.7.1 of the CMAQ model. Improvements to the way this component is treated are incorporated in later versions of CMAQ. If one considers all the models for which predictions were available in AQMEII, then generally there appears to be significant under-prediction of both PM_{10} and $PM_{2.5}$ (*Schere et al.,* 2012).

Table 3 shows the seasonal, domain-wide normalised mean biases of daily average PM_{10} concentrations for the North American Air Quality System and European AirBase networks from the CMAQ model. The model performance for $PM_{2.5}$ and PM_{10}, especially for the European domain, shows large under-prediction and occurs in other constructor models where the finest grid resolution is some kilometres or more, not just in CMAQ (*Solazzo et al.,* 2012).

Table 2. NMB of daily average $PM_{2.5}$ comparisons between predictions and observations in different seasons of 2006 over North America and Europe made under AQMEII (*Appel et al.* 2012) using the CMAQ model

Season and domain	**Approximate number of sites**	***NMB***
Winter, North America	958	0.304
Winter, Europe	160	–0.550
Spring, North America	958	0.189
Spring, Europe	160	–0.369
Summer, North America	958	–0.046
Summer, Europe	160	–0.372
Autumn, North America	958	0.363
Autumn, Europe	160	–0.242

Table 3. NMB of daily average PM_{10} comparisons between predictions and observations for different seasons in 2006 over North America and Europe made under AQMEII (*Appel et al.* 2012) using the CMAQ model

Season and domain	**Approximate number of sites**	***NMB***
Winter, North America	956	–0.479
Winter, Europe	1000	–0.648
Spring, North America	956	–0.565
Spring, Europe	1000	–0.562
Summer, North America	956	–0.574
Summer, Europe	1000	–0.612
Autumn, North America	956	–0.465
Autumn, Europe	1000	–0.468

The general tendency for constructor models with realistic chemistry and transport to under-predict particulate concentrations can be readily interpreted as either (1) due to the omission of local combustion source contributions in models with grid resolution of 5 km or more, or (2) due to the neglect or to the inaccurate estimate of the emissions of non-combustion, windblown or re-suspended dust (especially for the coarse fraction of particulate, the difference between PM_{10} and $PM_{2.5}$) or (3) inaccuracies in the instrumentation or the site description (rural, background *etc.*) in the observational network.

In constructor models, such as CMAQ, short-term average concentrations, such as daily concentrations, can be calculated. In *Table 4* from *Chemel et al.* (2010), the performance of CMAQ v4.6 is shown for the daily variation of the two main secondary species of interest, ozone and PM_{10}[5]. The model performance for daytime ozone concentrations over a year is seen to be superior to that of PM_{10}.

Table 4. Comparison of performance in predicting the daily maximum ozone and daily mean PM_{10} concentrations at AURN sites in the UK in 2003 for CMAQ v4.6

CMAQ v4.6 metric	**Maximum daily running eight-hour mean ozone**	**Daily mean PM_{10}**
NMB	0.05	–0.34
r (correlation coefficient)	0.69	0.47
FAC2 (%)	76.7	26.8
Number of sites	~40	~40

The CMAQ ozone *NMB* in *Table 4* for the UK in 2003 is within ±0.1. The ozone predictions in 2003 from the OSRM Lagrangian trajectory model for the UK in 2003, described by *Hayman et al.* (2010), also satisfy this performance measure. The AQMEII project provided CMAQ performance statistics for ozone for the many hundreds of ozone monitoring sites in Europe and North America. The *NMB* for daytime ozone over a year is generally within ±0.1, apart from during the summer season in Europe. Given the similar level of model performance from diverse models, a *NMB* of within ±0.1 appears to be achievable for the daytime ozone concentrations over a year with the current generation of photochemical models.

[5] The paper by Chemel *et al*. (2010) contains many more performance statistics than those summarised in *Table 4*. The daily PM_{10} *NMB* is equivalent to the annual average PM_{10} *NMB* in *Table 1* apart from rounding errors.

Using the seer OSRM and constructor CMAQ models, the impact of an oil refinery in southern England, with VOC emissions ~0.2 kg/s and NOx emissions ~0.2 kg/s, on annual average ozone concentrations in 2003 was assessed in CREMO. The predicted change in the annual average ozone concentration along a horizontal transect through the refinery (*Hayman et al.*, 2013d) shows that the emissions from the refinery led to a decrease in ozone on average in both models (of magnitude <0.2 μg m^{-3}) out to distances of a few hundred kilometres. The decrease is thought to be caused by the reaction of ozone with NO releases. The similarity in the operational performance of the OSRM and CMAQ models gives confidence in the simpler, seer OSRM model, which was designed to develop national policy for regulating ozone. However a full understanding of the performance of regional models over the complete range of conditions of interest cannot be based purely on operational performance metrics. Diagnostic evaluation is also required.

7. Examples of diagnostic evaluation

The modelling of the $PM_{2.5}$ and PM_{10} footprint from a power station ought be better than the regional predictions discussed in the previous section, because (1) there is no locally derived coarse particulate, (2) the primary source strength, which consists largely of inorganic compounds, is better known and (3) the plume chemistry without secondary organics is simpler. Examples from CREMO of the footprint of particulate matter from a power station source, are given in *Figs. 2* and *3* using the distance weighted footprint introduced schematically in Figure 1. The average concentration along a radial trajectory is normalised by the concentration in the near field, 30 km from the source.

Footprints used in diagnostic evaluations cannot be compared directly with observations. However they are useful for diagnosing how models are treating processes within the model. In *Fig. 2* the distance-dependent weightings of the annual average PM_{10} concentration in a power station footprint are compared for three types of models. Broadly the models show little dependence on distance suggesting that the removal by wet and dry deposition is largely in balance with production by chemical transformation. The detailed behaviour is different, presumably because of differences in details of the models' structure, parameters and the input data sets, even for the two models whose structure is similar (CMAQ v4.6 and CMAQ v4.7). *Sutton et al.* (2013) showed that apparently small differences in the temporal profile of ammonia emissions over the year in 2003 (although the spatial distribution of the annual total ammonia emission was identical) can make differences to the prediction of the annual spatial distribution of components of $PM_{2.5}$ and of acid deposition. This suggests that differences in the set-up of model runs can make differences to the spatial distribution, even if the models themselves (CMAQ v4.6 and CMAQ v4.7) are formulated in a similar way and demonstrate similar behaviour.

The distance-dependent weighting of the sulphur deposition from a power station shows a general decrease out to distance of 500km, while the distance-dependent weighting of the nitrogen deposition shows an increase out to the same distance[6].

In *Fig. 3*, the distance-dependent weightings of $PM_{2.5}$ and NO_2 for different kinds of sources investigated in CREMO are shown. Different distance dependencies are demonstrated. Small-scale variations in the plots may be the result of edge effects in the domains from which the footprints are plotted.

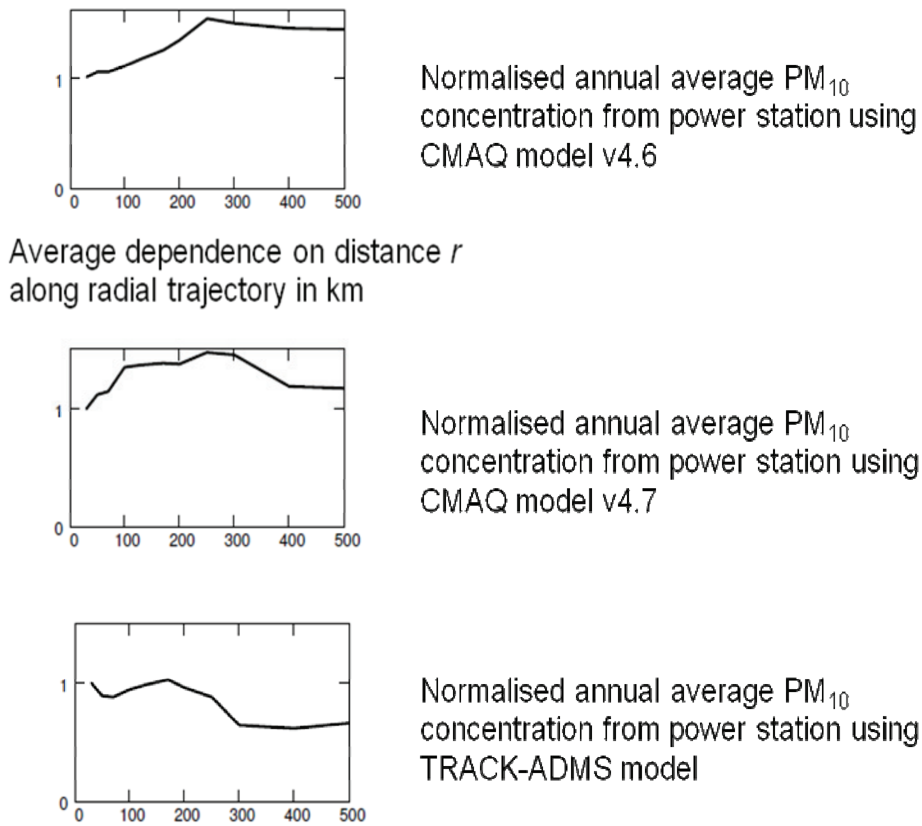

Fig. 2. Dependence of the PM_{10} concentration on distance in kilometres along a radial trajectory, derived from the footprint of a power station source in central southern England, using Eq. (2), normalised by the value of the concentration near the source at 30km, for two versions of the constructor model CMAQ (v4.6 and v4.7), *top two figures*, and for the seer model TRACK-ADMS, *bottom figure*.

[6] The deposition footprints are not shown in this paper.

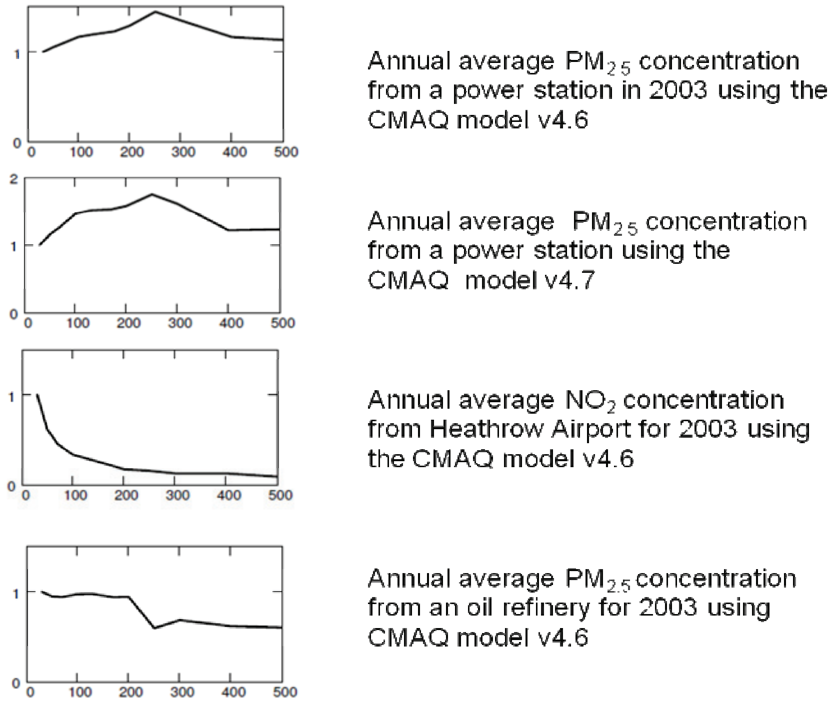

Fig. 3. Top two figures: Normalised dependence of the $PM_{2.5}$ concentration on distance in kilometres along a radial trajectory, derived from the footprint of a power station source, using Eq. (2), for two versions of the constructor model CMAQ (v4.6 and v4.7). *Bottom two figures*: Normalised dependence of the NO_2 concentration on distance in kilometres for Heathrow Airport and normalised dependence of the $PM_{2.5}$ concentration on distance in kilometres for an oil refinery in southern England from results of the constructor model CMAQ v4.6. Different distance dependencies are demonstrated in the footprints.

The EMEP model (NMI, 2010) is a constructor model and has been used to derive the contribution from individual countries to the regional concentration of $PM_{2.5}$ over Europe using grid cells of dimension 50 km×50 km. A diagnostic metric is the set of individual country footprints for a 15 per cent change in emissions, equivalent to the annual country-to-grid source–receptor matrices, see http://www.emep.int/SR_data/index_sr.html [accessed 21 May 2013]. *Fig. 4* shows the circumferentially averaged footprint of the UK centred on the middle of a grid cell in central England. The footprint is plotted at distances greater than 150 km from the centre of the country, because the national footprint of $PM_{2.5}$ concentration is distorted by the irregular distribution of sources over the country. The EMEP footprint for the UK is compared with the footprint from the typical power station analysed in CREMO. It is seen that, out to distances of 500 km, the balance between production and loss is approximately maintained in both the footprint calculations. As expected, atmospheric loss mechanisms from wet and dry deposition begin to dominate beyond this distance. The footprints are expressed in concentration units. The maximum $PM_{2.5}$ concentrations in the two cases are about 0.8 and 0.2 $\mu g m^{-3}$, respectively.

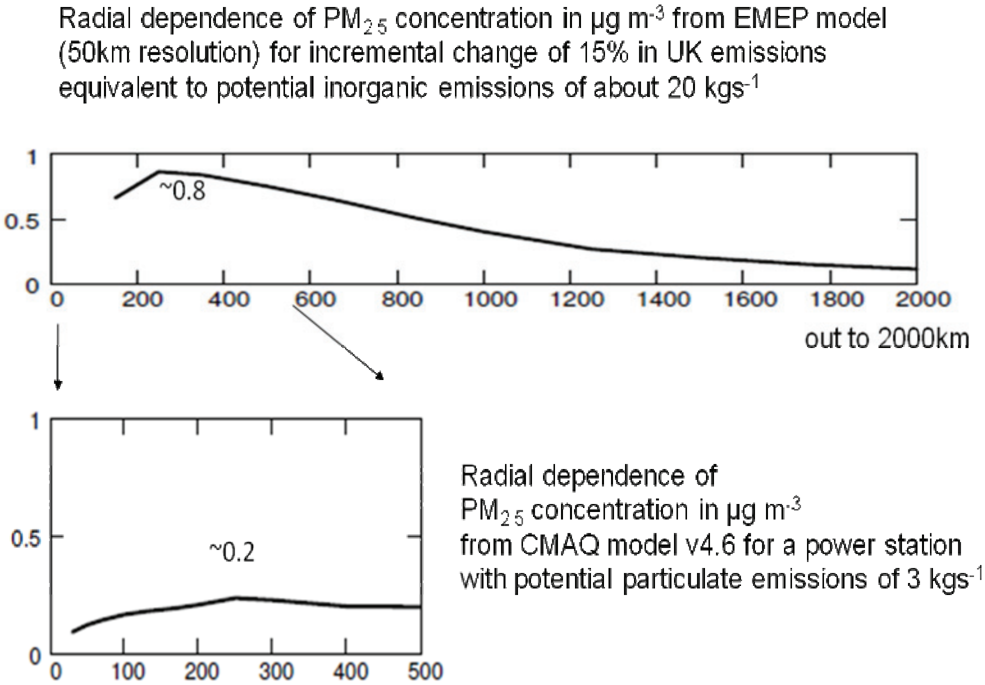

Fig. 4. Top: Dependence of the $PM_{2.5}$ concentration in μgm^{-3} on distance in kilometres, in 2003, along a radial trajectory starting in central England, based on a 15% reduction in UK emissions, based on the EMEP model (Klein *et al*. 2011). *Bottom:* Dependence of $PM_{2.5}$ concentration in μgm^{-3} on distance in 2003, along a radial trajectory, from the footprint of a power station source. Notional potential source strengths for the two types of sources (in kgs^{-1}) and the maximum values of the $PM_{2.5}$ concentrations in the two calculations (0.8 and 0.2 μgm^{-3}) are shown.

Although the species profiles (SO_2 : NOx: NH_3) of the UK emissions and a single coal-fired power station source are different, an approximate estimate can be made of the dominant precursor emissions, using the sum of primary, precursor species (SO_2 + NOx + NH3 + primary $PM_{2.5}$). Although only rough estimates, there is consistency in magnitude in the two cases. The 15% incremental change in UK emissions is six times the power station source strength and gives roughly four times the maximum $PM_{2.5}$ concentration of the power station. The distance-dependent weighting is different in detail. This may be a result of the different spatial resolution of the two models or the result of detailed differences in the treatment of particle formation in the two models.

The footprint is seen to be a valuable metric for diagnosing the behaviour of both constructor and seer air quality models. There is no objective way of deriving diagnostic metrics, though the structure of seer models are likely to indicate aspects of both types of models from which diagnostic metrics can be formulated.

8. Conclusions

The evaluation of models is thus not a wholly objective procedure i.e. it cannot be reduced to a set of rules. However it can be made more systematic by the use of checklists (*Risbey et al.*, 2001) and by systematically evaluating the performance of air quality models, as described by *Thunis et al*. (2011a,b) and *FAIRMODE* (2011) and illustrated by *Chemel et al*. (2010) and *Pederzoli et al*. (2011). Moreover if consistency is shown between the predictions of models which have been developed for performance (the model constructors' approach) and models used to develop understanding (the model seers' approach), then this gives greater confidence in the decisions made.

The operational and diagnostic evaluation of regional air quality models of the seer and constructor type for regulating secondary atmospheric pollutants, such ozone, PM_{10} and $PM_{2.5}$ have been illustrated. There was no clear separation between the operational performance of a seer and a constructor regional air quality model, using observational data sets and the commonly applied operational metrics *FA*C2 and *NMB*. A margin of error shown by operational evaluations of a model should be added to the predictions when the model is used to show compliance with a limit value. However it is not possible from the results of operational evaluations shown in this paper to set a minimum margin of error though this could be at least 20% for annual average concentrations of PM_{10} and $PM_{2.5}$. For long-term average ozone the margin of error appears to be smaller, but interest is really in episodic ozone for which no generally accepted metric has been agreed.

Diagnostic evaluation is needed to justify the use of regional air quality models under conditions where no observational data is available. Although widely applied metrics based on statistical variables are used in operational evaluation, no commonly agreed, standard set of diagnostics exists, which can be used to understand the performance of constructionist regional models. However it is hoped that the continued use of seer models will generate metrics for diagnostic evaluation on a case by case basis.

The footprint metric is shown to be a useful diagnostic for both user and constructor models. The footprint shows similar behaviour in the regional seer and constructor air quality models analysed. It illustrates clearly the balance between the production and loss of PM_{10} and $PM_{2.5}$ from a specified point source, such as a power station, out to distances of 500km beyond which removal processes dominate. For ozone the system behaviour is fundamentally different in episodic and non-episodic conditions. Unlike operational metrics there is no obvious way of deriving diagnostic metrics. However seer models are likely to reveal directly the structure of a model's underlying mathematical equations from which diagnostic metrics can be formed.

Acknowledgements: The principal author acknowledges valuable discussions held with colleagues within the context of the CREMO project when he was working at the Environment Agency and under the letter of agreement between Defra, the Environment Agency and the US Environmental Protection Agency establishing cooperation on exposure science. The CREMO project was supported by the Environment Agency of England under contract (20073/R&D SC060037). The partners included the University of Hertfordshire, AEA Technology, the Joint Environment Programme of the power industry, CEH Edinburgh, RD Scientific and Hayman Atmospheric Consulting. Some of these organisations made significant contributions in kind to the work programme. Results of this work may not necessarily reflect the views of the Environment Agency or any individual CREMO participant and no official endorsement should be inferred.

References

Abbott, J., Hayman, G., Vincent, K., Metcalfe, S., Dore, T., Skeffington, R., Whitehead, P., Whyatt, D., Passant, N. and *Woodfield, M.*, 2003: Uncertainty in acid deposition modelling and critical loads. *Environment Agency Technical Report P4-083(5)/1*. Bristol: Environment Agency.

Abbott, J.A., Vincent, K.J. and *Stedman, J.R.*, 2007: Methods for auditing the contribution of Environment Agency regulated processes to pollution. *Environment Agency Science Report SC030172/SR2*. Bristol: Environment Agency. https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=b839fd39-b1db-4fdd-9d8a-40e3ee476c19 [accessed May 21, 2013].

Abbott, J. and *Vincent, K.*, 2007: Auditing the contribution of Environment Agency regulated processes to pollution. *Environment Agency Science Report SC030172/SR3*. Bristol: Environment Agency.https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=65c9d2a8-87f9-4a2e-b67e-a69c5b244905 [accessed May 21, 2013].

Appel, W., Chemel, C., Roselle, S.J., Francis, X.V., Hu, R.-M., Sokhi, R.S., Rao, S.T. and *Galmarini, S.*, 2012: Examination of the Community Multiscale Air Quality (CMAQ) model performance over the North American and European domains. *Atmos. Environ. 53*,142–155.

Carslaw, D., 2011a: Defra urban model evaluation analysis – Phase 1. London: Defra. Available from: http://uk-air.defra.gov.uk/library/reports?report_id=654 [accessed May 21, 2013].

Carslaw, D., 2011b: Defra regional and transboundary model evaluation analysis – Phase 1. London: Defra. Available from: http://uk-air.defra.gov.uk/library/reports?report_id=653 [accessed May 21, 2013].

Carslaw, D., 2011c: Defra deposition model evaluation analysis – Phase 1. London: Defra. Available from: http://uk-air.defra.gov.uk/library/reports?report_id=652 [accessed May 21, 2013].

Chemel, C., Sokhi, R.S., Yu, Y., Hayman, G.D., Vincent, K.J., Dore, A.J., Tang, Y.S., Prain, H.D. and *Fisher, B.E.A.*, 2010: Evaluation of a CMAQ simulation at high resolution over the UK for the calendar year 2003. *Atmos. Environ. 44*, 2927–2939.

Chemel, C., Sokhi, R.S., Dore, A.J., Sutton, P., Vincent, K.J., Griffiths, S.J., Hayman, G.D., Wright, R., Baggaley, M., Hallsworth, S., Prain, H.D. and *Fisher, B.E.A.*, 2011: Predictions of UK regulated power station contributions to regional air pollution and deposition: a model comparison exercise. *J. Air Waste Manage. Ass. 61*, 1236–1245.

Chemel, C., Fisher B.E.A., Kong X., Francis X.V., Sokhi R.S., Good N., Collins W.J. and *Folberth G.A.*, 2014: Application of chemical transport model CMAQ to policy decisions regarding $PM_{2.5}$ in the UK. *Atmos. Environ. 82*, 410-417.

Denby, B., 2011: Guide on modelling nitrogen dioxide (NO_2) for air quality assessment and planning relevant to the European Air Quality Directive**.** Result of activities in the Forum for Air Quality Modelling in Europe. *FAIRMODE Working Group 1, Version 4.6. ETC/ACM Technical Paper 2011/15.* Bilthoven, The Netherlands, European Topic Centre on Air Pollution and Climate Change Mitigation, RIVM.

Dennis, R., Fox, T., Fuentes, M., Gilliland, A., Hanna, S., Hogrefe, C., Irwin, J., Rao, S.T., Scheffe, R., Schere, K., Steyn, D. and *Venkatram, A.*, 2010: A framework for evaluating regional-scale numerical photochemical modeling systems. *Environ. Fluid Mech. 10*, 471–489.

Derwent, R. and *Nelson, N.*, 2003: Development of a reactivity index for the control of the emissions of organic compounds. *R&D Technical Report P4-105 RC8309*. Bristol: Environment Agency.

Derwent, R., Witham, C., Redington, A., Jenkin, M., Stedman, J., Yardley, R. and *Hayman, G.*, 2009: Particulate matter at a rural location in southern England during 2006: Model sensitivities to precursor emissions. *Atmos. Environ. 43*, 689–696.

Derwent, D., Fraser, A., Abbott, J., Jenkin, M., Willis, P. and *Murrells, T.*, 2010: Evaluating the performance of air quality models. *Report prepared for the UK Department for Environment, Food and Rural Affairs*, Issue 3/June 2010. London: Defra. Available from:http://uk-air.defra.gov.uk/reports/cat05/1006241607_100608_MIP_Final_Version.pdf [accessed May 21, 2013].

Derwent, R.D., 2013: Chemical mechanism choice: impacts of ozone precursor emissions reductions in the United Kingdom on episodic peak ozone in the United Kingdom. *Environment Agency Report SC060037e/R*. Bristol: Environment Agency. https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed May 20, 2013].

Dore A.J., Vieno, M., Tang, Y.S., Dragosits, U., Dosio, A., Weston, K.J. and *Sutton, M.A.*, 2007: Modelling the atmospheric transport and deposition of sulphur and nitrogen over the United Kingdom and assessment of the influence of SO_2 emissions from international shipping. *Atmos. Environ. 41*, 2355–2367.

Dore, A.J., Carslaw, D., Chemel, C., Derwent, R.G., Fisher, B.E.A., Griffiths, S.J., Lawrence, S., Metcalfe, S.E., Redington, A., Simpson, D., Sokhi, R., Sutton, P., Vieno, M. and *Whyatt, J.D.*, 2013: Evaluation and inter-comparison of acid deposition models for the UK. *Air Pollution Modeling and its Application XXII* (ed. D G. Steyn, P.J.H. Builtjes and R.M.A. Timmermans), pp 505-10. Dordrecht: Springer.

EC, 2008: Directive 2008/50/EC of the European Parliament and of the Council of 21 May 2008 on Ambient air quality and cleaner air for Europe. *Official Journal of the European Union, L152*, 11 June 2008, 1–44.

Edwards, P.N., 2010: A vast machine Computer models, climate data and the politics of global warming, The MIT Press, Cambridge Massachusetts, London, England.

FAIRMODE (Forum for Air Quality Modelling in Europe), 2011: The application of models under the European Union's Air Quality Directive: A technical reference guide. *Technical Report No 10/2011*. Copenhagen: European Environment Agency. Available from: http://www.eea.europa.eu/publications/fairmode [accessed May 21, 2013].

Fisher, B.E.A., 2013: Comparison of simple and advanced regional models (CREMO): Outcomes for the Environment Agency**.** *Environment Agency Report SC060037/R*. Bristol: Environment Agency.https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed May 20, 2013].

Fisher, B., Chemel, C., Hu, R.-M. and *Sokhi, R.*, 2010: Emulating complex calculations for regulatory decisions. *Air Pollution Modeling and its Application XX*, (ed. D.G. Steyn and S.T. Rao), pp. 195–199. Dordrecht: Springer.

Fisher, B. E. A., Chemel, C., Francis, X.V., Hu, R.-M., Sokhi, R.S., Hayman, G.D., Vincent, K.J., Dore, A.J., Griffiths, S., Sutton, P. and *Wright, R.D.*, 2011: Diagnostic evaluation of complex and simple atmospheric chemical transport models by considering single source impacts in the UK. *Air Pollution Modeling and its Application XXI*, (ed.D.G. Steyn and S. Trini Castelli), pp 427–431. Dordrecht: Springer.

Francis, X., Chemel, C., Sokhi ,R.S., Norton, E.G., Ricketts, H. and *Fisher, B.*, 2011: Mechanisms responsible for the build-up of ozone over South East England during the August 2003 heat wave. *Atmos. Environ. 45*, 6880–6890.

Galmarini S. and *Rao S.T.*, 2011: The AQMEII two-continent Regional Air Quality Model evaluation study: Fueling ideas with unprecedented data. *Atmos. Environ. 45*, 2464.

Hall, D.J., Spanton, A.M., Dunkerley, F., Bennett, M. and *Griffiths, R.F.*, 2000a: A review of dispersion model inter-comparison studies using ISC, R91, AERMOD and ADMS. *Environment Agency Technical Report P353*, Environment Agency, Bristol.

Hall, D.J., Spanton, A.M., Dunkerley, F., Bennett, M. and *Griffiths, R.F.*, 2000b: An inter-comparison of the AERMOD and ADMS and ISC dispersion models for regulatory applications. *Environment Agency Technical Report P362*, Environment Agency, Bristol.

Hayman, G.D., Abbott, J., Davies, T., Thomson, C.L., Jenkin, M.E., Thetford, R. and *Fitzgerald, P.*, 2010: The Ozone Source–Receptor Model: a UK ozone policy tool. *Atmos. Environ. 44*, 4283–4297.

Hayman, G., Sokhi, R., Chemel, C., Griffiths, S., Vincent, K., Dore, A.J., Sutton, P. and *Wright, R.*, 2013a: Comparison of simple and advanced regional models (CREMO): Model evaluation protocol. *Environment Agency Report SC060037a/R.* Bristol: Environment Agency. https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed May 20, 2013].

Hayman, G., Sokhi, R., Chemel, C., Griffiths, S., Vincent, K., Dore, A.J., Sutton, P. and *Wright, D.R.*, 2013b: Comparison of simple and advanced regional models (CREMO): Model evaluation report. *Environment Agency Report SC060037b/R*. Bristol: Environment Agency.https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed May 20, 2013].

Hayman, G., Sokhi, R., Chemel, C., Griffiths, S., Vincent, K., Dore, A.J., Sutton, P. and *Wright, R.*, 2013c: Comparison of simple and advanced regional models (CREMO): Ozone diagnostics. *Environment Agency Report SC060037c/R*. Bristol: Environment Agency.https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed May 20, 2013].

Hayman, G., Sokhi, R., Chemel, C., Griffiths, S., Vincent, K., Dore, A.J., Sutton, P. and *Wright, R.*, 2013d. Comparison of simple and advanced regional models (CREMO): Model evaluation: Ground-level ozone. *Environment Agency Report SC060037d/R*. Bristol: Environment Agency. https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=cf825d8d-f061-44cb-9293-51b96939d244 [Accessed 20 May 2013].

Im, U., Bianconi, R., Solazzo, E., Kioutsioukis, I., Badia, A., Balzarini, A., Baró, R., Bellasio, R., Brunner, D., Chemel, C., Curci, G., van der Gon, H.D., Flemming, J., Forkel, R., Giordano, L., Jiménez-Guerrero, P., Hirtl, M., Hodzic, A., Honzak, L., Jorba, O., Knote, C., Makar, P.A., Manders-Groot, A., Neal, L., Pérez, J.L., Pirovano, G., Pouliot, G., San Jose, R., Savage, N., Schroder, W., Sokhi, R.S., Syrakov, D., Torian, A., Tuccella, P., Wang, K., Werhahn, J., Wolke, R., Zabkar, R., Zhang, Y., Zhang, J., Hogrefe, C. and *Galmarini, S.*, 2014. Evaluation of operational online-coupled regional air quality models over Europe and North America in the context of AQMEII phase 2. Part II: Particulate matter., *Atmos. Environ.* http://dx.doi.org/10.1016/j.atmosenv.2014.08.072

Klein, H., Gauss M., Nyiri A. and *Steensen B.M.*, 2011: Transboundary air pollution by main pollutants (S, N, O_3) and PM: The United Kingdom. *MSC-W Data Note 1/2011*. Oslo: Norwegian Meteorological Institute.

Middleton, D.R., Luhana L. and *Sokhi R.S.*, 2007: Overview of indicators for estimating the contribution of Environment Agency regulated sources to regional ozone. *Environment Agency Report SC030171/SR3*. Bristol: Environment Agency.

NMI, 2010: Transboundary acidification, eutrophication and ground-level ozone in Europe for 2008. *EMEP Status Report 2010*. Oslo: Norwegian Meteorological Institute. Available from: http://emep.int/publ/reports/2010/status_report_1_2010.pdf [accessed May 21, 2013].

Pederzoli, A., Thunis, P., Georgieva, E., Borge, R. and *Carruthers, D.*, 2011: Performance criteria for the benchmarking of air quality model regulatory applications: the 'target' approach.*14th Conference on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes*, 2–6 October 2011, Kos, Greece.

Pilkey-Jarvis, L. and *Pilkey, O.H.*, 2008: Useless arithmetic: ten points to ponder when using mathematical models in environmental decision making. *Public Administration Review 68*, 470–479.

Rao, S.T., Porter, P.S., Mobley, J.D. and *Hurley, F.*, 2011: Understanding the spatio-temporal variability in air pollution concentrations, *EM*, the Air and Waste Management Association Magazine for Environmental Managers, November 2011, 42–48.

Risbey, J., Van Der Sluijs, J.P., Ravetz, J. and *Jannsen, P.*, 2001: A check list for quality assurance in environmental modelling. *Report Number NW&S-E-2001-11.* Utrecht: Department of Science, Technology and Society, Utrecht University. Available from: http://www.nusap.net/downloads/reports/E2001-11.pdf.

Schere, K., Vautard, R., Solazzo, E., Hogrefe, C. and *Galmarini, S.*, 2012: Results and lessons learned from Phase 1 of the AQMEII. *EM*, the Air and Waste Management Association Magazine for Environmental Managers, July 2012, 30–37.

Solazzo, E., Bianconi, R., Pirovano, G., Matthias, V., Vautard, R., Moran, M.D., Appel, K.W., Bessagnet, B., Brandt, J., Christensen, J.H., Chemel, C., Coll, I., Ferreira, J., Forkel, R., Francis, X.V., Grell, G., Grossi, P., Hansen, A.B., Miranda, A.I., Nopmongcol, U., Prank, M., Sartelet, K.N., Schaap, M., Silver, J.D., Sokhi, R.S., Vira, J., Werhahn, J., Wolkem, R., Yarwood, G., Zhang, J., Rao, S.T. and *Galmarini, S.*, 2012: Operational model evaluation for particulate matter in Europe and North America in the context of AQMEII. *Atmos. Environ. 53*, 75–92.

Sutton, P., Chemel, C., Griffiths, S. and *Sokhi, R.S.*, 2013: Investigation, using CMAQ, of sensitivity of air quality modelling to input ammonia emissions. *Air Pollution Modeling and its Application XXII* (ed. D G. Steyn, P.J.H. Builtjes and R.M.A. Timmermans), pp 571-6. Dordrecht: Springer.*Thunis, P., Georgieva, E.* and *Galmarini, S.*, 2011a: A procedure for air quality models benchmarking. Version 2. Ispra, Italy: European Commission, Joint Research Centre. Available from: http://fairmode.ew.eea.europa.eu/models-benchmarking-sg4. [accessed July 2, 2013].

Thunis, P., Georgieva, E. and *Pederzoli, A.*, 2011b: The DELTA tool and Benchmarking report template. Concepts and user guide. Version 2. Ispra, Italy: European Commission Joint Research Centre. Available from: http://fairmode.ew.eea.europa.eu/models-benchmarking-sg4 [accessed July 2, 2013].

Vincent, K. and *Abbott, J.*, 2008: Air quality and deposition benefits from Environment Agency regulation. *Environment Agency Science Report SC060108*. Bristol: Environment Agency. https://publications.environment-agency.gov.uk/skeleton/publications/ViewPublication.aspx?id=a98e1cd3-7d34-40ef-8f68-2e401db1bfa4 [accessed May 20, 2013].

Wild, O., Fiore, A.M., Shindell, D.T., Doherty, R.M., Collins, W.J., Dentener, F.J., Schultz, M.G., Gong, S., MacKenzie, I.A., Zeng, G., Hess, P., Duncan, B.N., Bergmann, D.J., Szopa, S., Jonson, J.E., T. J. Keating, T.J. and *Zuber, A.*, 2012: Modelling future changes in surface ozone: a parameterized approach. *Atmos. Chem. Phys. 12,* 2037–2054.

Yu, Y., Sokhi, R.S., Kitwiroon, N., Middleton, D.R. and *Fisher, B.*, 2008: Performance characteristics of MM5–SMOKE–CMAQ for a summer photochemical episode in southeast England, United Kingdom. *Atmos. Environ. 42*, 4870–4883.