Treating Predictions From More Than One Model

Bernard Fisher

National Centre for Risk Assessment and Options Appraisal
Environment Agency
Kings Meadow House
Kings Meadow Road
READING RG1 8DG
Tel 01189 535244 Fax 01189 535265
Bernard.Fisher@environment-agency.gov.uk

1 INTRODUCTION

The usual approach to the comparison of models is to compare the models' performance against some data sets and then to choose the model which performs best, according to some criteria. A series of Harmonisation Conferences have been held at which datasets have been introduced, protocols for undertaking intercomparisons developed and recommendations made. A summary is available at:

http:www.dmu.dk/atmosphericenvironment/harmoni/conferences/belgirate/Outcome.asp

In practice insufficient data usually exists to make a clear preferred selection, but for operational reasons one model is adopted. The selected model is then used to make decisions about the appropriate regulation of new and existing sources. In this paper an alternative approach is proposed for situations when results from several models are available assuming that each model has an acceptable level of performance. The basic assumption is that accepting the inevitable inaccuracies in each model it is better to try to obtain a combined estimate based on all the information available.

One should make a clear distinction between the situation when one is comparing a number of models against measurements and the situation when one is dealing with just model predictions. In the former case is illustrated in the paper by Vardoulakis et al (2001), in which one finds for street canyons that the measurements are broadly encompassed within a range of models and input assumptions. This means that the assumptions of one of the available models at a given time represent approximately the real behaviour. Unfortunately no one model is able to represent accurately conditions all of the time. The latter case, when no measurements are available for testing, is the one dealt with in this paper. An example of an approach in this situation is that of Hall et al (2000b) who used a protocol of tests cases to examine the behaviour of three well tried dispersion models. The test cases cannot encompass all possible uses of the models but attempt to cover as wide a range of practical cases as possible.

2 THE DECISION MAKER

Generally the decision maker faces a situation in which results from several models are available, but there are no measurements to assess which model performs best. Deriving a combined estimate based on all the information available may seem sensible qualitatively, but a decision maker has to make use of the extra information in a quantitative way. This can be illustrated by an example. Supposing that one has the results of two models. If the results are presented as just two numbers, one does not have much extra information on which to base a decision. The two predictions could be regarded as either end of a range, so that the predictions provide an indication, perhaps, of the likely range of predictions. The air quality objective might lie above, below or within the range. Only in the case of the objective lying above the range would one be able to feel that there was no need for further investigation. Also if the model results are 'close' one has increased confidence in the models predictive capability. A simple case is one in which one has predictions from one or more dispersion models.

It is much more useful if predictions were always associated with some kind of uncertainty. There are various ways of expressing the uncertainty but it would need to be derived from verification exercises e.g. comparisons of dispersion model predictions with measurements reviewed in Hall et al (2000a). These authors conclude that not enough has been done to verify current atmospheric dispersion models.

One should not describe the prediction of a model as a single number. The number may be the best estimate that a model can produce but there is a range of possible numbers. One can describe this uncertainty in a range of ways. One way is shown in Fig 1 in which the predictions of two models are shown as fuzzy numbers (Ferson, Root and Kuhn, 1999). Fuzzy numbers are just one of a number of ways of expressing an uncertain number. It involves a single best estimate and range described by the upper and lower bound. The idea is that the uncertain or fuzzy number can be manipulated like ordinary numbers. The approach is somewhat similar to associating a probability distribution to the prediction of a model, but not as rigorous. The definition implies that there is a wide range of just **possible** values (with a low **possibility**). It is the most conservative estimate of the uncertainty about a quantity. At the other end of possibilities there is a best estimate, equal to the prediction, representing the most optimistic view of the degree of uncertainty. The wide range of uncertainty could be selected on the basis of the factor of two often assumed in dispersion models. There are other ways of representing the uncertainty, all of which depend on some subjective judgement, but the application of fuzzy numbers will suffice for an example.

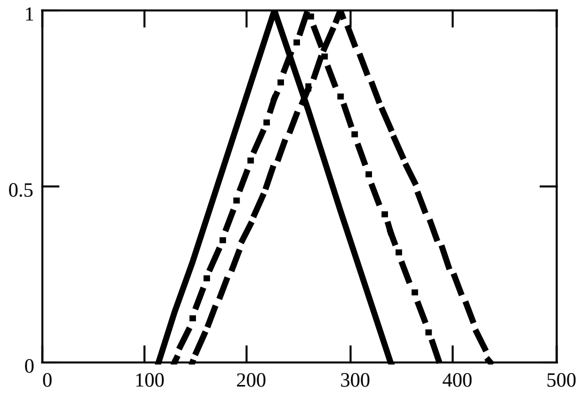

Figure 1. **Comparison of two models predictions using fuzzy numbers. Model A (——) and model B (--------) are the models, the x-axis is the range of concentrations and the y-axis is the possibility or weighting on the predictions. The combined model (·-·-·-·-·) is an estimate taking the average of the two model predictions. These numbers are based on a real case of predictions from the AERMOD and ADMS dispersion models for a tall stack with buoyant plume in unstable conditions (Hall et al, 2000b) assuming a factor of two uncertainty to define upper and lower bounds.**

In this example taken from Hall et al's report (2000b) for a straightforward case in flat terrain the well known factor of two has been adopted to represent the uncertainty. One cannot justify rigorously from a pure model intercomparison that a factor of two is justified. Certainly there are many straightforward cases where model predictions differ by more than a factor of two. R Timmis (private communication) has proposed a hierarchy for treating model uncertainty in dispersion calculations. In straightforward cases of a flat terrain and no buildings, or chemistry or peak concentrations, one should assume that models are accurate within a factor of 2 (or some other value accepted by the modelling community). For the more complex cases additional factors of uncertainty arise (Timmis again suggest further factors of 2) to account for the additional uncertainty in the way models account for buildings or complex terrain etc. He assumes for simplicity that this introduces further multiplicative factors. One may be concerned that this could soon lead to very large factors of uncertainty as one multiplies the factors together. In practice except for a few cases, the additional complexity just involves a single extra process. Sometimes complexity is not necessarily multiplicative and bounds on possible ranges arise. This appears to be the case for estimating the effective height of chimneys for emissions of sulphur dioxide from small boilers (Vawda, Moorcroft, Khandelwal and Whall, 2001). Though these authors recommend a simple approach for stack height allowing for a building correction, the method leads to stack heights, which are not very different from earlier methods. It appears that the combination of adequate stack height without a building, and the building height, leads to a limited range of stack heights which take into account both buildings and dispersion.

This still leaves open the question of how to combine uncertainties. For the decision maker the question is how to make use of uncertain predictions from two models. This is an example of “addition” and taking the mean of fuzzy numbers. For the case of additional complexity one would be taking the “products” of fuzzy numbers.

3 WAYS OF COMBINING FUZZY NUMBERS

For the example considered above one sees that there is fair degree of overlap between the estimates and therefore a reasonable way to proceed is to combine the two fuzzy numbers. If one sets 0.05 as the possibility level (the qualitative description of the scale on the y-axis) one considers acceptable i.e. one is rather conservative about the range of possible values that could arise, one would estimate that the prediction lies between 135 and 381.

One may then compare the two fuzzy number representing predictions. The shape of the fuzzy number is a matter of opinion. It is argued by those who favour the use of fuzzy numbers that the method is not sensitive to the detailed shape of the possibility function e.g. whether it were Gaussian or a top hat distribution. However it depends on some estimate of the uncertainty in prediction so that some kind of bound, together with a best estimate, is needed from every model prediction. This is not routinely available from dispersion models, but it could be made available or could be requested from model users. This is the situation assuming that the model is being applied in the usual way, in order to obtain the best estimate of the maximum ground-level concentration.

In contrast a screening model would be used in another way. A screening model would have a much wider range of possible predictions, without necessarily weighting the maximum possibility (best estimate) towards the upper bound of the range. The screening model is used to incorporate all results, or as many results as possible. It is anticipated that normally the predictions of a screening model would have to be refined using a more accurate model, in order to reduce the range of possible predictions.

Fuzzy numbers can be manipulated in a consistent manner. They arise as a refinement of interval bounds, which are rather crude encapsulation of uncertainty (Ferson, Root and Kuhn, 1999). For instance, intervals imply that although we cannot give a quantity’s value exactly, one can give exact bounds on the quantity. Fuzzy numbers generalise intervals. A fuzzy number can be thought of as a nested stack of intervals containing many levels of confidence about uncertainty. These levels of confidence range between 0 (corresponding to the most conservative, widest interval) and 1 (the narrowest interval, which assumes one is really good at making predictions). By definition, all fuzzy numbers must reach one and must be convex (single humped). The scale between 0 and 1 is said to measure the possibility that a number is within the interval at a particular level. This introduces the notion that possibilities can be graded. Fuzzy numbers and their arithmetic provide a simple and workable methodology that is valid for handling non-statistical uncertainty in calculations.

In the application under consideration one might wish to take the average of the two estimates and this leads to the combined estimate shown in the figure. As long as two fuzzy predictions contain a significant amount of overlap this seems a sensible way to proceed. However in situations where the fuzzy numbers are in ranges which do not overlap, one would conclude that the predictions are significantly different. Model A is predicting values definitely lower than model B and taking some average would not be appropriate.

The reasons for a **significant difference** between model predictions are varied. They could arise because of an input error in one of the models. The models may be using different algorithms. For example the models may differ in the way they treat a plume near the top of the boundary layer or what happens to a plume near the top of the hill. One model may predict zero concentration at the ground, while the other model may predict substantial ground-level concentrations. The only way to deal with this situation is to consider the full specification of the model. This is the reason for transparency in the descriptions of models. Of particular interest are the algorithms for which a small variation in parameter values leads to a large change in concentration.

An alternative approach to combining two estimates is random sampling of two independent probability distributions. If one considered that the predictions were uniformly distributed between the upper and lower bounds with equal probabilities, one can then obtain an average estimate by sampling each distribution at random. This assumes that one chooses values from the two distributions in an independent way to obtain a combined distribution. If one chooses the 5 and 95 percentile points in the combined distribution as bounds on the estimate, one obtains the range 167 to 344.

The conventional, simple approach is just to quote a range with the upper and lower bounds chosen from the results from the best estimates of the two models e.g. between 226 and 290. This is a start to defining uncertainty, but the measure of uncertainty is arbitrary depending on predictions of models in specific situations without regard to model performance. The range can appear unrealistically small. An alternative, simple approach is to take the range from the higher of the two upper bounds of the predictions and the lesser of the two lower bounds of the predictions, namely a range 113 to 425 in this case. The fuzzy number approach gives a more bounded range, 135 to 381, while the assumption that there are random independent errors associated with each model gives the smaller range, 167 to 344.

Thus taking ranges directly from models' best predictions does not appear to give a realistic estimate of uncertainty. Taking the maximum range has the advantage of simplicity but may be regarded as encompassing too broad a range of possibilities. The other two approaches are to some extent qualitative. Taking the mean of fuzzy numbers is related to taking the mean end points of intervals.

4 SUBJECTIVE ELEMENT TO THE USE OF FUZZY NUMBERS AND FUZZY SETS

It is apparent that there is a subjective element to the use of fuzzy numbers, and as will be seen later, in the use of fuzzy sets. Basically one is trying to combine, or aggregate, two quantities in a way that is not normally done in conventional arithmetic. Ways of doing this have been proposed in the literature (Klir and Folger, 1988) but are not universally accepted, nor commonly applied in the environmental field.

However it should be recognised that the application of IPPC (Integrated Pollution Prevention Control) faces a similar dilemma. It requires judgements about the relative weighting of concentrations of different chemical species, which may be in different media (air, water or soil). These may be scaled individually against the appropriate environmental standards, but one is still making an optimisation of some kind of penalty function, which depends on the combined concentrations C_A and C_B of species A and species B whose effects cannot necessarily be compared on a common basis. Practical methods have to be applied to make decisions about the best practicable environmental option. However combining concentrations of different chemical species in the optimisation is more subjective than combining different estimates of the concentration of the same chemical species, which is the situation considered in this paper.

5 COMBINING TWO DIFFERENT MODELS

The operation of fuzzy aggregation, a generalisation of the addition of fuzzy numbers, can be adopted when one has separate sub-models describing different parts of system. This might be an approach to quantifying the extra uncertainties introduced by complexity in the Timmis hierarchy. It would apply to situations when one is not just adding two concentrations together, but where the relationship between parts of the model is more complex.

Fig 2 shows the annual average NO_2 from all road transport sources using a multi-source dispersion model at a scale of 1km by 1km (Fisher, private communication). This does not show the impact from the roads. One needs different models to treat pollution dispersion on the local (metres) and regional scale (kilometres). If one treats the urban area on a regional scale, the local impact of emissions from major roads is not apparent. In an earlier paper, methods for estimating roadside and background concentrations were described (Fisher and Sokhi, 2000). These have been extended to treat an urban area allowing for the variation in emissions with distance from the centre of the urban area (in a model called urbanGRAM). The contributions of roadside and background NOx are added and the annual average NO2 is estimated using a non-linear relationship between NOx and NO2.

To emphasise the effect of length scale, urbanGRAM has been applied to an idealised version of London in which three major circular roads are imagined to surround areas of the city: an inner ring road, an outer ring road and an orbital road on the outskirts of the city. A cross-section of the annual average NO_2 concentration along a profile across the city is shown in Figure 3. The NO_2 air quality standard of 21ppb is exceeded within a central zone by the background concentration. However the exceedence around the outer ring road arises from the combined contributions of the background and the roadside concentrations.

Colvile et al (2001) have discussed ways of estimating the uncertainties of modelled urban concentrations using a probabilistic mapping technique, which is very similar in principle to the technique used in this paper. The principle of probabilistic mapping is to replace concentration values on a map with qualitative descriptions of environmental quality relative to the relevant standards. From comparisons between predictions of an urban model and measurements, Colvile et al (2001) have assessed the standard deviation in the precision of the NO2 annual average predictions to be of order 10% of the mean.

They did not distinguish between roadside and background monitoring sites because of the limited number of measurement sites available. One precautionary way of applying probabilistic mapping when estimating areas of air quality exceedence, taking account of uncertainty, would be assume that air concentrations one standard deviation lower than the air quality standard of 21ppb (i.e. 19ppb) are not acceptable. This would have the effect of greatly enlarging the size of the central air quality management area. The gradient in the background concentration as a function of distance is small, so that small changes in the air quality standard produce large changes in the size of the air quality management area. In contrast the gradient in concentration close to the road is large, so that small changes in the air quality standard applied have a smaller effect on the size of the air quality management area.

Applying a discrete criterion to define an air quality management area might be regarded as a far from ideal way of taking account of uncertainty. The estimates of the air quality management area are not robust and are very sensitive to qualitative estimates of the uncertainty. There are significant differences in the size of the air quality management area, if uncertainty is included, depending on whether most of the pollution arises from background sources or from nearby road transport sources.

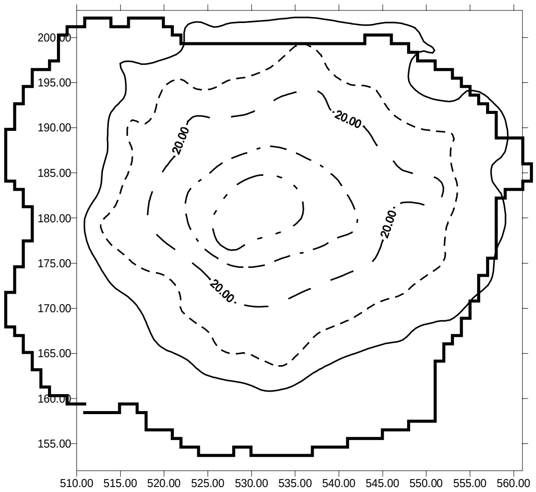

Figure 2. Background annual average NO_2 concentration in ppb over London from road transport emissions

An alternative approach is to assume that all estimates have some fuzziness, or uncertainty, associated with them, which will tend to smooth out the boundaries of any area to which a discrete, crisp criterion is applied. This approach can be applied to the example of a road within a city using the measure of fuzziness applied in an earlier paper (Fisher and Newlands, 2000). In this paper a simple form for the multiplicative error associated with estimates of concentration C(x) at a point x of the form $1/(1+C_0/C(x))^2$ was adopted where C_0 is the air quality standard. This had the advantage that functions could be easily manipulated. The membership function of a fuzzy set representing an air quality management area takes the simple form $1/(1+C_0/C(x))$, where higher values of the membership function denote greater certainty that the point x lies within the air quality management area and vice versa. (This choice is preferable to a step function for which $C>C_0$ implies exceedence and $C>C_0$ implies no exceedence.)

It was possible to obtain an analytical formula for the membership of an air quality management area based on the aggregation of the membership functions of background and roadside air quality management areas considered separately. The same approach may be possible for combining parts of a dispersion calculation such as the dispersion and building effect, or the dispersion and complex terrain. This would enable the errors or overall fuzziness in a complex situation to be assessed. As an alternative sensitivity analysis and associated techniques generate error assessments, but these can become very complex and computationally intensive. The operations on fuzzy sets, such as aggregation, in this case lead to a simpler error assessment. This is preferable to just multiplying together factors of 2.

If this approach is applied to the air quality management area around the outer ring road about 10km from the city centre, one obtains the membership function shown in Fig 4. The membership function, or weighting, is shown on the y-axis and the distance from the city centre is shown on the x-axis in kilometres. Hence only distances of a few hundred metres from the road are being considered. High values represent more certainty that the point lies within an air quality management area and low values, or weightings, represent less certainty that the point lies within an air quality management area.

The result is dependent on choosing a form for the membership functions and that the errors in the roadside concentration are independent of the errors in the background concentration. The aggregation operation involves a convolution integral over the weighting of errors in the roadside and background concentrations. The basic conclusion is that the choice of an air quality management area is not a discrete step function, but should be a smooth weighting function as shown in Fig 4. This is much more intuitively appealing since the profile of the annual average NO_2 concentration is itself a smooth function, albeit with a steeper gradient near a road. In conclusion, estimates of harm or damage based on concentration values are to be preferred to estimates that artificially distinguish areas of exceedence (harm) from areas of no exceedence (no harm). The latter approach does have the advantage of simplicity, highlighting possible areas of concern in scoping or screening studies.

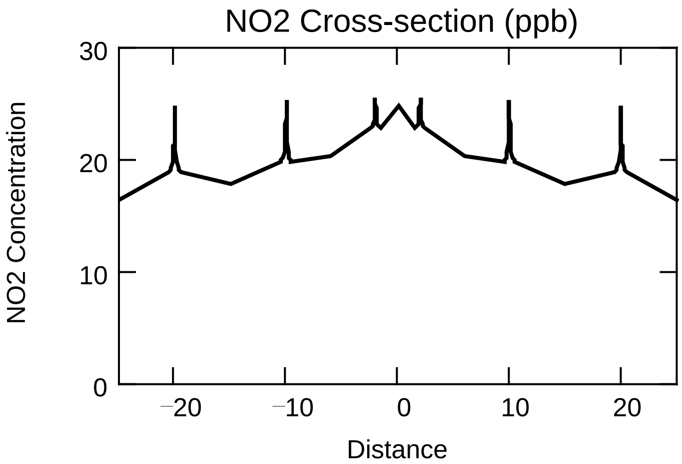

Figure 3. Idealised profile of the annual average NO_2 concentration in ppb in London along a cross-section through the centre of the city. The spikes denote high concentrations in the vicinity of major roads.

The approach illustrated here for combining the uncertainty results from two models is analogous to the problem of combining two concentration frequency distributions. This question arises when, for example, one has two different contributions, such as the urban background concentration and the concentration from a major stationary source. Similar aggregation methods based on the assumption of independent fluctuations may be applied.

6 METHODS ADOPTED IN ASSESSING HEALTH EFFECTS

The health assessments of the Committee on the Medical Effects of Air Pollutants (COMEAP, 2001) are based on annual concentrations of PM_{10} and have led to proposals to revise the national air quality objectives for particles (DEFRA, 2001). The health effect is taken to be proportional to the PM_{10} concentration. In this case the health effect community does not recognise any sharp boundaries between acceptable and unacceptable concentrations (the air quality objective). There appears to be an inconsistency between the methods of analysis. One way of looking at the COMEAP method is that one should not assign a sharp boundary to an air quality management area. Instead the PM_{10} concentration is proportional to the health outcome. The PM_{10} concentration is proportional to the percentage increase in death rates. This approach may be compared with the definition of air quality management areas. The difference depends on the stage at which other arguments are introduced into the decision making. The 'air quality management area' approach sets an objective, in terms of a fixed concentration, although in reality there must be some considerable uncertainty to defining where the concentration lies. The 'health outcomes' approach allows a sliding scale related to concentration. Inevitably at some stage an acceptable level of outcome is considered tolerable in the sliding scale, to allow for wider societal considerations.

Another way of illustrating how sharp boundaries or demarcations result in conclusions which are very sensitive to the assumptions would be to consider the number of people, who are exposed to exceedences of the annual average NO_2 standard. For this case it should be noted that the area, where there is a high concentration, is very near to the road. It is also an area where people may not be exposed. Houses are usually set back some way from the road carriageway so that the area of very high concentration does not necessarily correspond to a high number of people exposed. It is also not clear whether a street canyon increases or decreases the population exposed.

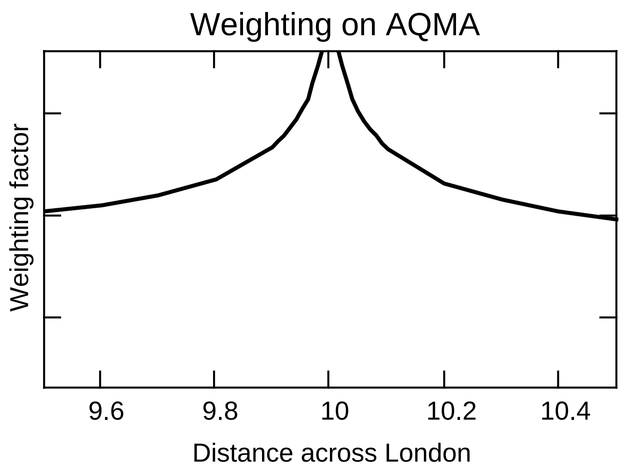

Figure 4. Weighting function representing the likelihood of exceeding the air quality standard for the annual average NO_2 concentration in the vicinity of an idealised major road 10km from the centre of London. High values of the weighting represent a high likelihood of an exceedence and vice versa.

For the idealised example of a major road 10km from the centre of London, the running of a dispersion model (urbanGRAM but with no allowance for uncertainty) suggests that all persons within 240m of the road would be within an area where the annual average NO_2 standard was exceeded. If instead the air quality standard was assumed to be 20.5 ppb, then it is estimated that all persons within 540m of the road would live in an area of exceedence. The number of people exposed to an exceedence is very sensitive to uncertainty. These results apply to that part of the curve in Fig 4 where the gradient is shallow. The overall result is expressed better by the curve in Fig 4, which shows that the area exceeding the air quality objective is very sensitive to distance far the major road, but less sensitive to distance near to the road. There is a danger that a single estimate of the number of persons exposed would not show this sensitivity.

7 COMPARISON OF TWO MODELS' PREDICTIONS AGAINST AN AIR QUALITY OBJECTIVE

In a third example of the application of fuzzy sets, one may consider the situation when one has two models, leading to predictions C_A and C_B of the annual average NO2 concentration. One needs to know how likely is the air quality objective to be exceeded, and what is the effect of having two independent estimates.

For illustrative purposes the weighting on the uncertainty in each of two estimates of concentration C is assumed to be $1/(1+C_0/C)^2$, where C_0 is the air quality objective (=21ppb). One may then compare the uncertainty in the joint weighting of $(C_A + C_B)/2$ compared with C_0. A similar analysis to that in Section

4, based on a simple analytical formula, leads to the aggregate weighting shown in Fig 5, which is a function of the two concentrations. The result is expressed as the weighting relative to the weighting, if both concentrations were equal to the air quality objective. Thus one sees that if both predictions give concentrations higher than the objective, then the relative weighting on an exceedence would be greater than one, and the objective is likely to be exceeded. If both predictions give concentrations lower than the objective, the relative weighting on an exceedence would be less than one, and the objective is unlikely to be exceeded. In this case the question of deciding whether an air quality objective is exceeded, is complex. It requires knowledge of a two-dimensional function describing the joint uncertainty in the two estimates. Similar multi-dimensional surfaces arise if one considered the dependence of model predictions on the uncertainty in input values and the values of parameters assumed within the model. The benefits of being able to generate these surfaces in simple ways are obvious.

8 CONCLUSIONS

All assessments should be associated with some uncertainty. Deterministic modelling neglects uncertainty. The treatment of uncertainty is inevitably associated with some subjective judgements. Using results from more than one model has distinct advantages in decision making, if used with care.

Decisions should reflect the underlying uncertainty. This paper serves to illustrate that there are different approaches, using fuzzy sets and fuzzy numbers, which would deserve investigation. These potentially could lead to better descriptions of imperfect knowledge. Air quality management areas are an ideal example of the application of the concepts. The disadvantage is that the decision may become more complex.

The views expressed in this paper are personal and do not necessarily represent those of the Environment Agency, or the National Centre for Risk Assessment and Options Appraisal. A shorter version of this paper was previously published as:

Fisher B and Ireland M, 2001, Using fuzzy numbers to treat predictions from more than one model, NATO/CCMS Internal Technical Meeting on Air Pollution Modelling and its Applications, Louvain-la-Neuve, Belgium 15-19 October 2001

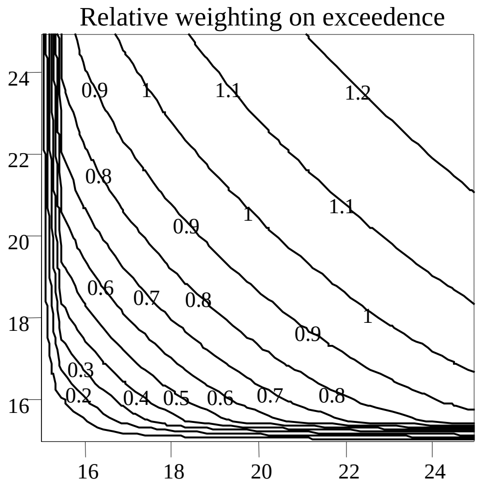

Figure 5. **Relative weighting of an exceedence of the long-term air quality standard for NO_2 as a function of two estimates of the concentration. The x-axis is the best estimate from one model and the y-axis is the best estimate from another model. If both models give a high estimate then the standard is likely to be exceeded, while if both give low estimates then it is unlikely that the standard will be exceeded.**

9 REFERENCES

Colvile R N, Woodfield N K, Carruthers D J, Fisher B E A, Rickard A, Neville S and Gudgin A, 2001, Uncertainty in urban air quality mapping, submitted for publication

COMEAP, 2001, Report on the long-term effects of particles on mortality, Department of Health

DEFRA, 2001, Consultation document on proposals for air quality objectives for particles, benzene, carbon monoxide and polycyclic aromatic hydrocarbons

Ferson S, Root W and Kuhn R, 1999, RAMAS RiskCalc: Risk assessment with uncertain numbers. Applied Biomathematics, Setauket, New York. http://www.ramas.com

Fisher B E A and Newlands A G, 2000, The designation of fuzzy air quality management areas, in: *Air Pollution Modeling and its Application*, S-E Gryning and F A Schiermeier, ed, Kluwer, New York, pp 97-105.

Fisher B and Sokhi R S, 2000, Investigation of roadside concentrations in busy streets using the model GRAM - Conditions leading to high short-term concentrations, *Int J of Env Technology and Management*, **14**, 488-495.

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

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

Klir G J and Folger T A, 1988, *Fuzzy sets, uncertainty and information*, Prentice Hall, New Jersey

Vardoulakis S, Fisher B, Gonzalez-Flesca N and Pericleous K, 2001, NATO/CCMS Internal Technical Meeting on Air Pollution Modelling and its Applications, Louvain-la-Neuve, Belgium 15-19 October 2001

Vawda Y, Moorcroft J S, Khandelwal P and Whall C, 2001, Sulphur dioxide emissions from small boilers – supplementary assistance on stack height determination, Clean Air, 31, 89-95