COMPARISON OF MODEL EVALUATION METHODOLOGIES WITH APPLICATION TO ADMS 3 AND US MODELS

C.A. McHugh, D.J. Carruthers, H. Higson & S.J. Dyster
Cambridge Environmental Research Consultants
3 Kings Parade, Cambridge, CB2 1SJ, UK

ABSTRACT
At previous Harmonisation Workshops model comparisons have largely used the Model Validation Kit's BOOT statistical package to assess model performance against field data sets based on comparisons of maximum arcwise concentrations. The emerging methodology from the US that uses near-centreline values has been introduced at the Workshops and compared briefly with the Model Validation Kit (MVK) method. This paper compares the results from both methodologies which have been applied to results from three models for three well-known field data sets: Prairie Grass, Kincaid and Indianapolis, which cover a range of release heights and meteorological conditions.

KEYWORDS
Dispersion modelling, model evaluation, maximum arcwise concentration, near-centreline concentration, fractional bias.

INTRODUCTION

At previous Harmonisation Workshops model comparisons have largely used the Model Validation Kit's BOOT statistical package to assess model performance against field data sets based on maximum arcwise concentrations. The emerging methodology from the US, which uses near-centreline values has been introduced at the Workshops and compared briefly with the Model Validation Kit (MVK) method. This paper compares the results from both methodologies which have been applied to results from three models for three well-known field data sets: Prairie Grass, Kincaid and Indianapolis, which cover a range of release heights and meteorological conditions.

The Model Validation Kit method for model evaluation, that has been presented at previous workshops, uses the maximum arcwise concentration observed (MAC). The method has the advantage of being straightforward but, after identification of the arcwise maximum, most of the measured data are discarded. It also ignores the inherent uncertainty in the arcwise maximum due to turbulence. The highest measured concentrations are likely to be higher than the ensemble mean concentration due to turbulent fluctuations. By contrast the near-centreline method uses more of the measured data to fit a Gaussian distribution to the observations and in boot strapping the observed data to produce a large number of near-centreline values. It attempts to account for the uncertainty in the measurements through the use of near-centreline values, but the analysis is more complex. As with the MVK method it assumes that models predict only ensemble averages. ADMS 3 can predict the effect on concentrations of short term turbulent fluctuations and this ability, although not explored here, could provide a direct comparison between modelled uncertainty and a treatment of uncertainty in the observed values.

In this study the models used are ADMS 3, which was released this year and has an improved treatment of high plumes compared with ADMS 2, and the US models, ISCST3 and AERMOD. The study explores the two evaluation methods and shows how the different analytical methods can give a contrasting impression of model performance and, hence, the importance of understanding evaluation measures.

METHODS OF ANALYSIS

Model Validation Kit (MVK): Maximum Arcwise Concentration

The scheme referred to in this paper as the Model Validation Kit (MVK) methodology uses the BOOT statistical package developed by Hanna et al (Hanna & Chang, 1991) that has been widely distributed with the Model Validation Kit by Olesen (Olesen, 1998). Results using the MVK method have been reported at several previous Harmonisation Workshops.

The method compares the maximum observed concentration for each arc, i.e. the maximum arcwise concentration (MAC), with model predictions of centreline concentrations. The most commonly presented outputs of the package are statistics of the observed and modelled maximum arcwise concentration (MAC): mean, standard deviation, bias, normalised mean square error, correlation, proportion of values within a factor 2, fractional bias and its standard deviation. The fractional bias, F_b, is defined as:

$$F_b = \frac{\overline{C}_0 - \overline{C}_p}{0.5(\overline{C}_0 + \overline{C}_p)}$$

$\overline{C}_0$ is the mean of observed MAC and $\overline{C}_p$ is the mean of model predicted MAC.

A value of 1.99 in absolute fractional bias indicates no agreement between model and observations, 1.00 corresponds to model predictions within a factor of 3 of the observations and for a value of 0.67 they are within a factor 2. Negative values of F_b indicate a model over-prediction and positive values a model under-prediction. Fractional bias is non-linear and bounded by ± 2. For example, an over-prediction of 2 times has a fractional bias of –0.67 and an over-prediction of 20 times has a fractional bias of –1.81 but, an over-prediction of 200 times has a fractional bias of –1.98. The consequence of this is that if a model tends to over or under-predict considerably, though by quite different factors, the variance of the fractional bias will be low. A low variance in fractional bias can be taken as indicating confidence in the model predictions but taken on its own this would be misleading.

The MVK analysis presents an average fractional bias for all the arc-hours. Taking a simple average means that over-predictions (negative) will cancel out under-predictions (positive), as with the mean concentrations, which may give a false impression of model performance. However, the other statistical parameters help to give the complete picture.

Analysis of model performance based on the MAC is based solely on the maximum observed concentrations for each arc. Concentrations measured at other locations along each arc that are less than the maximum value are used only to establish the maximum value.

The maximum observed value is subject to significant inherent uncertainty due to small scale turbulence. If an experiment were repeated under identical conditions, the maximum concentration measured would be different each time. The ensemble mean predicted by the models, which represents the average over a large number of repetitions, is the value compared here with measurements. Therefore, even a 'perfect' ensemble would be expected to under-predict compared with the highest observed values.

ASTM Methodology: Near-centreline values

The American Society for Testing and Materials (ASTM) is developing a Standard Practice (Irwin, 1997), for comparing results from air quality models with field data. This method compares predicted centreline concentrations against near-centreline concentrations (NCC) based on the observations within a regime, not simply for one arc-hour of data. A Gaussian distribution, centred on the centre of mass of the observations, is fitted to the observed values for each arc-hou of data. For each arc-hour values within a distance of the centreline, typically $0.67 s_y$, are bootstrapped to produce, for each regime, a large data set from which an average value is calculated. The method seeks to take into account the variability in maximum values due to turbulence through the use of these near-centreline values.

The average of the near-centreline concentrations will be lower than the maximum observed concentration and, therefore, the ASTM method compares a model prediction of ensemble centreline concentrations with a lower value than the MVK method. A trend to over-prediction identified by the MVK method would appear as greater over-predictions according to the ASTM method.

The method relies on defining several regimes, for instance ranges of atmospheric stability, within which dispersion mechanisms are similar. For each regime the fractional bias is calculated. (Note that the fractional bias is defined similarly to the MVK method but with the opposite sign.) A polled value of absolute fractional bias is calculated from weighting the modulus of fractional bias for each regime. The weighting uses the inverse of the variance of the fractional bias for each regime. As described above the variance of the fractional bias is low for models which consistently over or under-predict by a significant factor, even if the amount by which they over or under-predict varies considerably. Therefore, this weighting method can give undue prominence to regimes where the predictions are poor.

Both methods have a limitation due to receptor spacing in that the observed values are unlikely to record the arcwise maximum value that would be captured by a continuous array of monitors and so the maximum recorded values under-estimate the true maximum.

Some of the consequences of the ASTM method are discussed in a paper by Irwin and Rosu (1998) and both methods are discussed by Olesen (1997). His main conclusion is that a model which is 'perfect' at predicting ensemble concentrations according to the near-centreline (ASTM) method, under-predicts in far more cases than a model which is 'perfect' according to the maximum arcwise (MVK) method. He deduces that the near-centreline method is therefore biased towards models which under-estimate the ensemble mean.

VALIDATION DATA SETS

In each of the three field experiments the measured data have been collected on arcs downwind of a source. In the Kincaid and Indianapolis experiments a quality indicator, from 0 to 3, has been assigned to each hour's measurements according to how well the maximum concentration is defined. It has been recommended that data with a quality indicator of 2 or 3 should be used when analysing model behaviour and this has been followed for the Kincaid and Indianapolis data sets. For the Prairie Grass experiment all the hours of data have been used.

Kincaid Power Plant

The Kincaid data set contains 171 hours of tracer experiments, which were performed at the Kincaid power plant in Illinois, USA, which has a 187m stack (TRC, 1983, Hanna & Paine, 1989). Meteorological conditions ranged from neutral to convective. In the ADMS and AERMOD model runs shown here, the boundary layer height was calculated using each model's own meteorological pre-processor.

Indianapolis experiment – Perry K Power Plant

The Indianapolis data set contains the results of 170 hours of SF_6 tracer experiments carried out for EPRI (Electric Power Research Institute) in 1985 at the Perry K power plant on the south-west edge of Indianapolis (TRC,1986). The power plant has a stack 84m tall and is located in a mixed industrial/commercial/urban area. Tracer concentrations were recorded hourly from arrays of up to 160 receptors on arcs at distances ranging from 0.25km to 12 km from the stack, which were moved according to the prevailing wind direction. Meteorological conditions covered a range of stability classes and wind speeds throughout the experimental period, which covered daytime and night-time.

Project Prairie Grass

Project Prairie Grass was carried out in north central Nebraska in the summer of 1956, (Barad, 1958). The site was located on virtually flat land covered with natural prairie grasses for which a roughness length of 0.006m was used. Small amounts of SO_2 tracer were released over 10-minute periods from approximately 0.5m above ground level. Concentration measurements were made at a height of 1.5m along arcs at five downwind distances: 50, 100, 200, 400 and 800m.

About half of the 70 trials were conducted during unstable (convective) daytime conditions and the rest were held at night with temperature inversions present (stable conditions). Extensive meteorological measurements were taken on-site during the trials.

THE DISPERSION MODELS

ISCST is a U.S. Environmental Protection Agency (EPA) model. It is a so-called 'old generation' model, based on the 1970's description of boundary layer physics (US-EPA, 1995). It uses Pasquill stability categories to describe the stability of the atmospheric boundary layer and assumes a Gaussian distribution of the concentration in the crosswind vertical and horizontal directions under all stabilities.

ADMS and AERMOD are both described as 'new generation' models (CERC, 1999, US-EPA 1998). These models describe the state of the atmospheric boundary layer using two parameters: boundary layer depth and Monin-Obukhov length. The vertical concentration distribution is Gaussian in neutral and stable atmospheres but is a skewed-Gaussian in convective conditions. A Gaussian distribution is assumed in the crosswind horizontal direction for all stabilities. The models differ in several respects, but the differences most relevant to the three data sets presented here are, firstly the treatment of plume rise and, secondly, the use of mean plume height in ADMS. ADMS uses a Runge-Kutta scheme to solve the conservation equations and calculates pluem variabels at the mean plume height, whereas AERMOD uses a modified Briggs formula and calculates plume variables at the plume centreline.

ADMS model runs have been carried out by CERC using ADMS version 3.0. US model runs (AERMOD, ISCST3) have been provided by the US EPA (Peters, 1998). The version of ADMS used was a pre-release version of ADMS 3 and the version of AERMOD used was AERMOD 98022 with meteorological pre-processor AERMET 98022.

ADMS has been developed in and is widely used throughout the U.K. by industry, consultants and regulatory bodies. AERMOD is under development in the US and will be submitted to the US EPA for consideration as a US regulatory model.

MODEL EVALUATION

Kincaid

The MVK results in Table 1 show ADMS predictions agreeing well with the observations in terms of mean and variance in particular. AERMOD shows significant under-estimates with the predicted mean 50% of the observed and standard deviation 61% of the observed. ISCST3 also shows significant under-predictions of the mean with over-estimates of the variance.

Figure 1 shows the fractional bias of each of the 29 regimes defined for the ASTM methodology for Kincaid. The regimes are distance-stability regimes, so the first regime refers to the nearest arc to the source and most unstable conditions and the second refers to the nearest arc and the next stability case. (Up to 12 arcs of data were used in 4 stability regimes.) ADMS has a fractional bias of modulus less than 1 i.e. predictions within a factor of 3, for all but one regime. AERMOD has predictions within a factor of three for all but three regimes. The single indicator of model performance, the pooled absolute fractional bias (PAFB) is lower for AERMOD (0.39) than for ADMS (0.53).

The pooled absolute fractional bias (PAFB) for ISCST3 shows that predictions are typically not within a factor of 3 of observation. An un-weighted average of the fractional bias in each regime would produce an overall value of approximately 1.0, but the weighting according to the inverse of the variance of fractional bias has emphasised those regimes where there are consistent, large under-predictions, giving a PAFB nearer to the limit of 2.0. As noted previously, these consistent, large over and under-predictions are accompanied by a low value of variance.

Considering quality 3 observation data only it would be expected that, as the observed value of MAC is likely to be closer to the maximum possible observed concentration, that there will be an increasing tendency to under-estimate and this is the case with the AERMOD results in Table 3. The under-estimate of 50% of the mean for quality 2 and 3 data in Table 1 compares with an under-estimate of 40% of the mean in Table 3, with a slightly reduced variance. The small over-prediction of the mean by ADMS 3 for quality 2 and 3 data has been replaced by a small under-prediction as anticipated. ISCST3 predicts about 56% of the mean in both cases, although the variance is increased for quality 3 data alone.

Model	**Mean** $ng/m^3/g/s$	**Sigma** $ng/m^3/g/s$	**Bias** $ng/m^3/g/s$	**NMSE**	**Cor**	**Fa2**	**Fb**	**Fs**
Observations	41.0	39.3	0.0	0.00	1.000	1.000	0.000	0.000
ADMS	43.2	33.5	-2.2	0.77	0.494	0.581	-0.053	0.159
AERMOD	20.3	24.1	20.7	2.28	0.348	0.329	0.675	0.480
ISCST3	23.1	53.3	17.9	3.83	0.259	0.256	0.558	-0.303

NMSE: Normal mean square error, **Cor:** Correlation, **Fa2:** Fraction within a factor of 2, **Fs:** Fractional variance

Table 1. MVK statistics: Kincaid, Data quality:2&3

Model	**Pooled Absolute Fractional Bias**	
	Average	**Standard deviation**
ADMS	0.5336	0.0179
AERMOD	0.3856	0.0206
ISCST3	1.9947	0.0003

Table 2. ASTM statistics: Kincaid, Data quality: 2&3

In Table 4 ADMS and ISCST3 have a better (lower) PAFB with just quality 3 data than quality 2 and 3 data, so they perform better when compared with concentrations including the arcwise maximum closer to the true arcwise maximum. AERMOD gives a higher PAFB for quality 3 data alone, which is consistent with the MVK results, that showed a worse mean under-prediction, but it is unexpected that the measure of overall performance should be worse. The change in PAFB for ISCST3 is small, suggesting that it is under-estimating by such significant factors such that it is insensitive to how well defined the observed MAC value is. This conclusion is supported by the MVK results support. Figure 2 shows the fractional bias in each regime.

Model	**Mean** $ng/m^3/g/s$	**Sigma** $ng/m^3/g/s$	**Bias** $ng/m^3/g/s$	**NMSE**	**Cor**	**Fa2**	**Fb**	**Fs**
Observations	54.3	40.3	0.00	0.00	1.000	1.000	0.000	0.000
ADMS	51.7	34.7	2.7	0.57	0.446	0.669	0.051	0.147
AERMOD	21.8	21.8	32.6	2.07	0.403	0.293	0.855	0.593
ISCST3	30.0	60.0	24.3	2.80	0.258	0.275	0.577	-0.394

Table 3. BOOT statistics: Kincaid, Data quality: 3

Model	**Pooled Absolute Fractional Bias**

	Average	Standard deviation
ADMS	0.4582	0.0165
AERMOD	0.4894	0.0198
ISCST3	1.9942	0.0003

Table 4. ASTM statistics: Kincaid, Data quality: 3

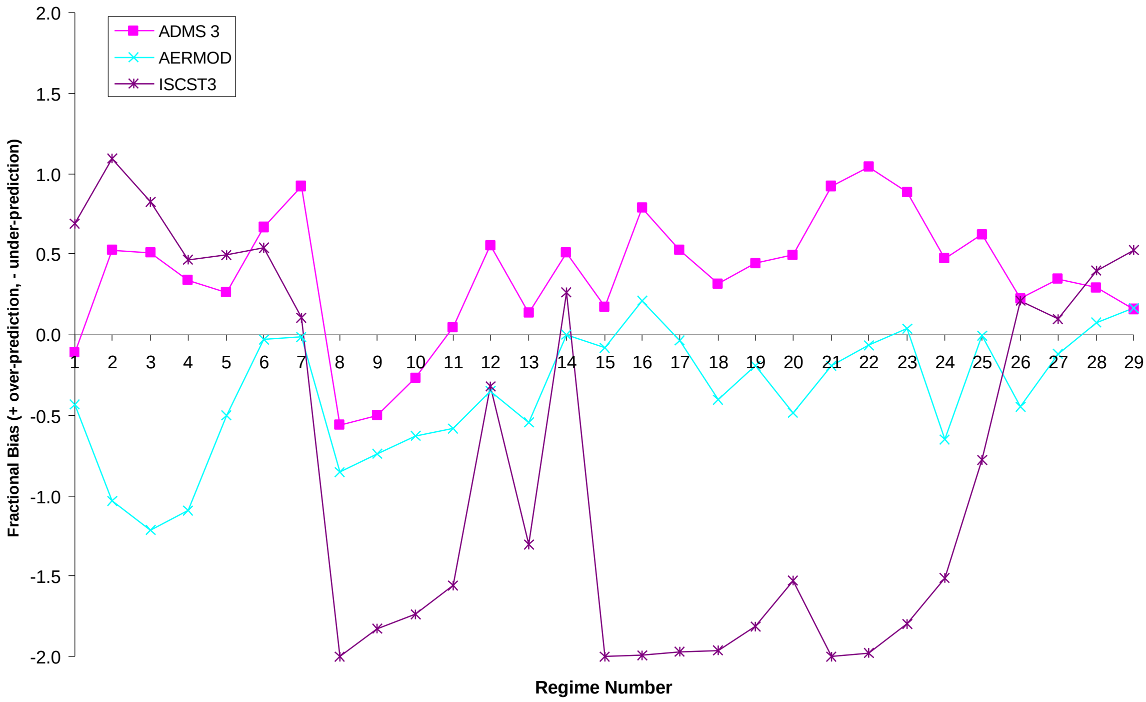

Fig. 1 Kincaid, Quality 2&3 data, Fractional Bias for each regime

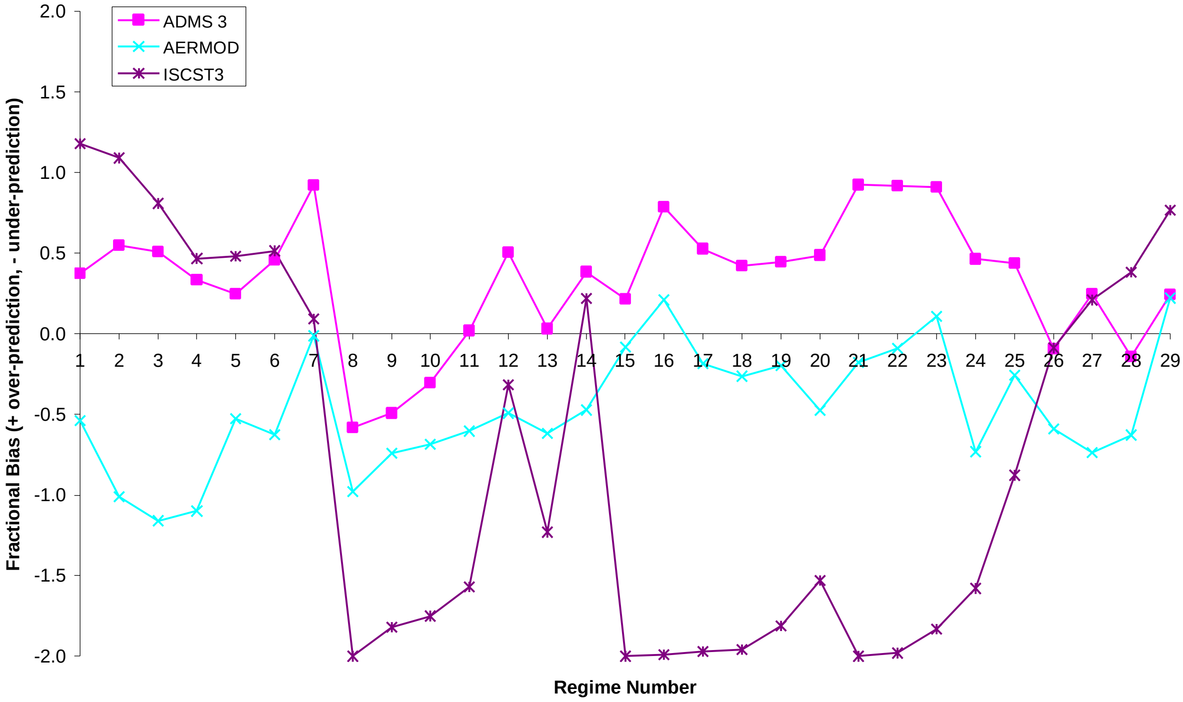

Fig. 2 Kincaid, Quality 3 data, Fractional Bias for each regime

Indianapolis

In the Indianapolis experiment high concentrations were reported at night that were unexpected given the height of the stack, the plume rise parameters and the stable meteorological conditions reported. However all the quality 2 and 3 data, daytime and night-time, have been used here.

Results from MVK, Table 5, show ADMS slightly over-estimating the mean (+3%) and AERMOD slightly under-estimating (-13%) with ISCST3 predicting a mean approximately one and a half times the observed. All the models have a similar fraction of predictions within a factor of 2 of observed. The other MVK statistics show slightly better performance from ADMS than from AERMOD. The ASTM results, Table 6, show AERMOD performing better than ADMS in terms of PAFB. Although AERMOD under-predicted by more than ADMS over-predicted when compared with the MAC using the MVK method, in the ASTM method the comparison is between modelled centreline concentrations and the average of NCC values, which is lower than the MAC. The results are therefore consistent with the difference between the methods. Figure 3 shows the smaller fractional bias in each regime for AERMOD than for ADMS when compared with the mean NCC.

Model	**Mean ng/m^3/g/s**	**Sigma ng/m^3/g/s**	**Bias ng/m^3/g/s**	**NMSE**	**Cor**	**Fa2**	**Fb**	**Fs**
Observations	257.8	221.6	0.00	0.00	1.000	1.000	0.000	0.000
ADMS	265.3	255.1	-7.53	1.25	0.256	0.423	-0.029	-0.141
AERMOD	224.6	195.5	33.22	1.28	0.168	0.408	0.138	0.125
ISCST3	404.2	321.0	-146.40	1.44	0.162	0.445	-0.442	-0.366

Table 5. BOOT statistics: Indianapolis, Qualities 2&3

Model	**Pooled Absolute Fractional Bias**	
	Average	**Standard deviation**
ADMS	0.5161	0.0225
AERMOD	0.3393	0.0276
ISCST3	0.8665	0.0200

Table 6. ASTM statistics: Indianapolis, Qualities 2&3

As with the Kincaid quality 2&3 results, ADMS has a negative (over-prediction) in fractional bias that is smaller in magnitude than AERMOD's positive value (under-prediction) according to the MVK method, but a higher PAFB according to the ASTM method. This may be partly due to the comparison with higher lower values in the ASTM method or due to over-estimates in some regimes being cancelled by under-estimates in others in the MVK method, which does not occur with the use of PAFB.

Prairie Grass

The MVK statistical analysis, Table 7, shows AERMOD predicting a mean concentration identical to the observed value with zero fractional bias. It should be noted that the Prairie Grass results were used directly in the formulation of plume spread parameters for surface sources for AERMOD, (Cimorelli et al, 96-TP24B.04). As would be expected, the number of observations within a factor 2 and the correlation are both high (0.76 and 0.75, respectively). It is interesting to see how this good performance when comparing maximum arcwise concentrations using MVK is interpreted by ASTM.

Figure 4 shows the fractional bias by regime, and AERMOD over-predicts by more than a factor of 3 in 6 out of the 35 regimes. It has a worse (higher) PAFB (0.40) than for the Kincaid experiment (quality 2 and 3 data) (PAFB=0.39) and the Indianapolis experiment (0.34), in both of which MVK indicated under-predictions by AERMOD.

Conversely, ADMS, which performed quite well for Kincaid according to MVK with a small over-prediction, (PAFB of 0.53), but for the Prairie Grass experiment where MVK suggests consistent under-predictions the values of PAFB is only slightly worse (0.56). ISCST3 performs relatively well according to the MVK method with a small mean under-prediction, but it is the worst model judged by the ASTM measure of PAFB. This may once again be due to significant over-predictions in several regimes (7, 8, 15, 22, 29, 30) being given undue prominence when pooled.

Model	**Mean** $mg/m^3/g/s$	**Sigma** $mg/m^3/g/s$	**Bias** $mg/m^3/g/s$	**NMSE**	**Cor**	**Fa2**	**Fb**	**Fs**
Observations*	2.14	3.75	0.00	0.00	1.000	1.000	0.000	0.000
ADMS	1.20	2.68	0.94	3.62	0.641	0.456	0.566	0.333
AERMOD	2.14	4.35	0.00	1.87	0.749	0.759	0.000	-0.148
ISCST3	2.01	3.50	0.13	1.76	0.716	0.621	0.064	0.069

* observed concentrations corrected for evaporative loss

Table 7. BOOT statistics: Prairie Grass, All qualities

Model	**Pooled Absolute Fractional Bias**	
	Average	**Standard deviation**
ADMS	0.5584	0.0100
AERMOD	0.3979	0.0104
ISCST3	0.9414	0.0088

Table 8. ASTM statistics: Prairie Grass, All qualities

DISCUSSION

The three data sets used in this validation study cover a range of source elevation and meteorological conditions. The model results have been compared with observations and analysed using two methodologies, the Model Validation Kit (MVK) method which compares modelled centreline concentrations against maximum arcwise concentrations (MAC), and the emerging ASTM method which compares the model prediction against near-centreline (NCC) values. The comparison demonstrated several points :

(i) NCC values will be lower than MAC values and therefore a model which agrees perfectly with the observed values when comparing with the MAC, will over-estimate when compared with NCCs.

(ii) a low value of variance of the fractional bias should not on its own be taken as an indication of confidence in model results as large over and under-predictions will be accompanied by a low variance even if the factor varies considerably, due to the non-linear nature of fractional bias.

(iii) use of the inverse of the variance of fractional bias as a weighting for combining results from different regimes in the ASTM method gives prominence to regimes with large over or underpredictions.

(iv) Alternative weighting measures should be considered e.g. number of arc-hours within each regime

The study demonstrates the necessity of understanding the behaviour of evaluation parameters and the importance of looking at a variety of parameters in assessing model performance.

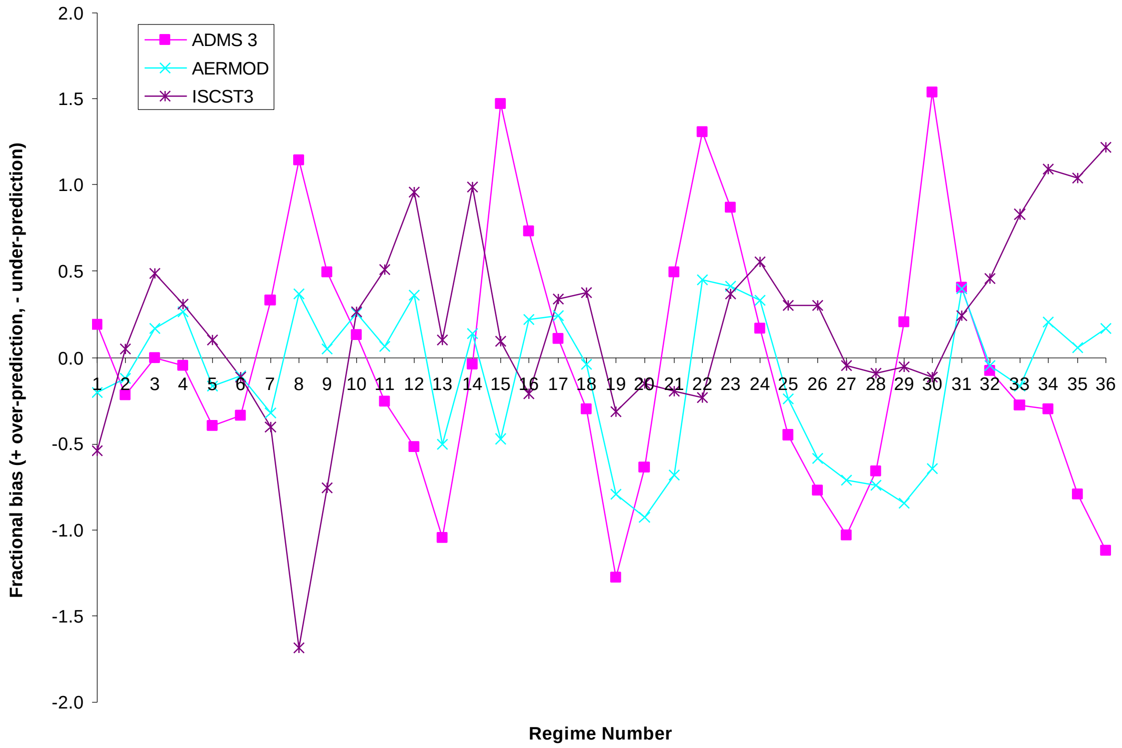

Fig 3. Indianapolis, Quality 2&3 data, Fractional Bias for each regime

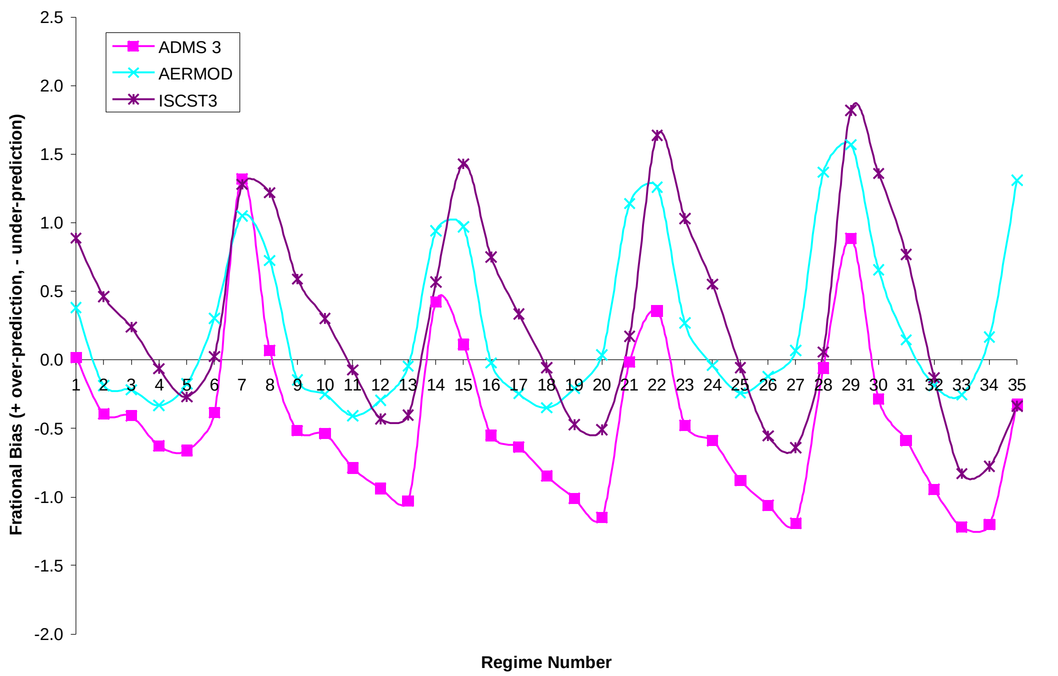

Fig 4. Prairie Grass, All data, Fractional Bias for each regime

REFERENCES

Barad, M.L., 1958. *Project Prairie Grass, a field program in diffusion*. Geophysical Research Papers no. 59, vols I-III.

Carruthers, D.J., S. Dyster and C.A. McHugh, 1998*. Contrasting methods for validating ADMS using the Indianapolis dataset*. Proc. 5th Int. Conf. On Harmonisation within Dispersion Modelling for Regulatory Purposes pp 104-110. To be published in J. Envt. & Pollution.

CERC, 1999. *ADMS 3 User Guide,* CERC, Cambridge, UK

Cimorelli, A.J.; Perry, S.G.; Lee, R.F.; Paine, R.J.; Venkatram, A.; Weil, J.C. and R.B. Wilson., 1996 *Current progress in the AERMIC model development program*. 96-TP24B.04.

Hanna S. and Paine R., 1989, *Hybrid Plume Dispersion Model (HPDM) development and evaluation*, J. Applied Met. **28**, pp206-224

Hanna S., Strimaitis D. and Chang C, 1991*, Hazard Response Modelling Uncertainty (a Quantitative Method). Vol 1 : User's Guide for Software for Evaluating Hazardous Gas Dispersion Models*. Sigma Research Corp., Westford, MA

Irwin, J.S., 1997*. Draft standard practice for statistical evaluation of atmospheric dispersion models*. American Society for Testing and Materials, Z6849Z.

Irwin, J.S. and M-H Rosu, 1998. *Comments on a draft practice for statistical evaluation of atmospheric dispersion models*. US-EPA.

Olesen, H.R., 1997. *Tools for model evaluation.* Proc.22nd NATO/CCMS International Technical Meeting on Air Pollution Modelling and its Applications, Clermont-Ferrand France

Olesen, H.R., 1998. Model Validation Kit – status and outlook. Proc. 5th International Conference on Harmonistaion within atmospheric dispersion modelling for regulatory purposes, Rhodes, Greece. To be published in J. Envt. & Pollution.

Peters, W., 1998. *The development of centreline concentrations from AERMOD, ISCST3 and HPDM for the Prairie Grass, Indianapolis and Kincaid field experiments*. US EPA Technical Memorandum for contract no. 68D70069. February 4, 1998 I:\504.

TRC, 1983, *Overview, Results and Conclusions for the EPRI Plume Model Validation and Development Project : Plains Site*, EPRI, EA-3074, EPRI, Palo Alto, CA

TRC, 1986, *Urban power plant plume studies*, EPRI report EA-5468, EPRI, Palo Alto, CA

US EPA, 1998, *Revised Draft Users' Guide for the AMS/EPA Regulatory Model – AERMOD*

US EPA, 1995, *User's Guide for the Industrial Source Complex (ISC3) Dispersion Models*, EPA-454/B-95-003a