USER EXPERIENCE WITH MODEL VALIDATION EXERCISES

Kathrin Baumann-Stanzer, Martin Piringer, Erwin Polreich, Marcus Hirtl, Erwin Petz, Marianne Bügelmayer

Central Institute for Meteorology and Geodynamics, Vienna, Austria

Abstract: Gaussian and Lagrangian model runs are evaluated in comparison to field data from the Odour Release and Odour Dispersion project and to wind tunnel data from the Mock Urban Setting Test (MUST). Different statistical metrics are discussed. To conclude which model performs best in the two cases, a weighted multiplier proposed by Sornette et al. (2007) is calculated based on each metric and finally multiplied to one score per model and experiment. The results illustrate once again that a good model performance is strongly dependent on the model input (e.g. terrain data, roughness length). Promising results are received from a combination of the Lagrangian dispersion model LASAT with wind field simulations calculated with the CFD model MISKAM.

Key words: *model validation, statistical measures, Gaussian and Lagrangian models, COST 732.*

1. INTRODUCTION

The validation of air quality models is increasingly requested from model users to prove the legitimacy and reliability of the model to decision makers (e.g. industry and authorities).

Oreskes et al. (1994) aver that validation of numerical models of natural systems is limited to the inherently partial confirmation of models by the demonstration of agreement between observation and prediction within acceptability criteria. Validation is the process of determining the degree to which a model is an accurate representation of the real world from the perspective of its intended uses. Model validation exercises thus can enhance our confidence in the model if it is not rejected according to predefined criteria in a number of tests. Sornette et al. (2007) propose a mathematical approach to quantify the relative state of validation of a model by a multiplier F based on quantitative measures and expert judgement:

$$F\left[p(Model \mid Observations), q, c_{novel}\right] = \left[\frac{\tanh\left(\frac{p}{q} + \frac{1}{c_{novel}}\right)}{\tanh\left(1 + \frac{1}{c_{novel}}\right)}\right]^4 \qquad (1)$$

p stands for the probability of the model given the observational data and q is a (predefined) reference likelihood. Thus, p/q is a metric for the agreement between model and observations: p/q=0.1 (poor fit), p/q=1 (marginally good fit), p/q=10 (good fit). The parameter c_{novel} is a measure of the impact of the experiment on the validation process: c_{novel}=1 (marginally useful new test), c_{novel}=10 (substantially new test), c_{novel}=100 (important new test).

In this study, this approach is applied for a simple Gaussian model ONG (described in Pechinger and Petz, 1997), the advanced Gaussian model ADMS (CERC, 2001), the Lagrangian model LASAT (Janicke, 2005) and other Gaussian and Lagrangian models applied by groups participating in COST 732. Results from two experiments described in the following are presented. In prior validation exercises, field concentration data from line sources in street canyons (Lohmeyer et al., 2002; Hirtl and Baumann-Stanzer, 2007), from a high stack in flat terrain (Pechinger and Petz, 1997) and in complex terrain (Hirtl et al., 2007) were used.

Field experiment data from tracer experiments conducted within the **Odour Release and Odour Dispersion (OROD) project** (Bächlin et al., 2002) are available from the German environmental programme BWPLUS. The source is a pig fattening unit in fairly flat terrain. Concentration data sets from 13 (SF6) tracer experiments of 10 minutes duration, meteorological data, odour intensity data estimated by a panel and the exact source values (source flow rate, volume flux and emission velocity at source) are available. All experiments took place under neutral conditions, wind speeds ranging from 2.5 to 7.9 ms^{-1}.

The **Mock Urban Setting Test (MUST) experiment** (run at the Dugway Proving grounds in western Utah in September 2001) was designed to study airflow and plume transport in urban areas and to provide a test case for model validation (Biltoft, 2001; Yee and Biltoft, 2004). 120 standard size shipping containers (12.2 m length, 2.42 m width and 2.54 m height) were set up in a nearly regular array of 10 by 12 obstacles, covering an area of around 200 m by 200 m. The terrain of the field site is flat with bushes and grass land with a height of approximately 0.5 to 1 m.
In this study, only wind tunnel data of the MUST experiment carried out at the Environmental Wind Tunnel Laboratory at Hamburg University are used (Bezpalcova et al., 2005; Harms, 2005).

Reference wind speeds in the wind tunnel test runs ranged between 7.5 and 8.5 ms^{-1}, source flow rates between 10 and 14 $lhour^{-1}$. Thus, the model user first has to decide the proper values for wind speed, emission rate, emission velocity and roughness length for the model runs. Validation results for a large number of CFD models for this exercise collected within COST Action 732 are presented by Olesen et al. at this Conference. This presentation

therefore mainly concentrates on the results for non-CFD models. Besides of the non CFD model applications, wind fields are simulated also with the CFD model MSIKAM (Eichhorn, 1989) and used as input to the LASAT model (referred to as LASAT c in the following section).

Excel workbooks including graphics and statistical metrics to support an exploratory analyses of model performance have been developed in COST Action 732 (see also the presentation by Olesen et al.). Other statistical methods for model evaluation are additionally applied in this paper.

2. RESULTS

The choice of the metrics and the interpretation of the results are essential for the outcome of the model validation. In this context, it furthermore has to be kept in mind that the model is always evaluated for a certain purpose (e.g. the prediction of the concentration maxima).

Kobayashi and Salam (2000) present a method for the comparison of model to observational data mainly using deviation-based statistics but including the correlation coefficient (r) as a constituent: The difference between the model result m and the observation o is calculated with the mean squared deviation MSD, which becomes small if the simulation is close to the measurements.

$$MSD = \frac{1}{n}\sum_{i=1}^{n}\left(m_i - o_i\right)^2 \tag{2}$$

MSD can be separated into three parts, the bias *SB*, the difference between the standard deviations of the simulation and measurements *SDSD* and the lack of positive correlation weighted by the standard deviations *LCS*:

$$MSD = SB + SDSD + LCS \tag{3}$$

$$SB = \left(\overline{m} - \overline{o}\right)^2 \tag{4}$$

$$SDSD = \left(SD_{model} - SD_{obs}\right)^2 = \left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(m_i - \overline{m}\right)^2} - \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(o_i - \overline{o}\right)^2}\right)^2 \tag{5}$$

$$LCS = 2SD_{model}SD_{obs}\left(1 - r\right) \tag{6}$$

The squared bias *SB* describes the difference of the simulated and observed mean values. A larger *SDSD* indicates that the model fails to simulate the magnitude of fluctuation of the observational data. Large values of *LCS* reveal that the model fails to simulate the distribution pattern of the observational data.

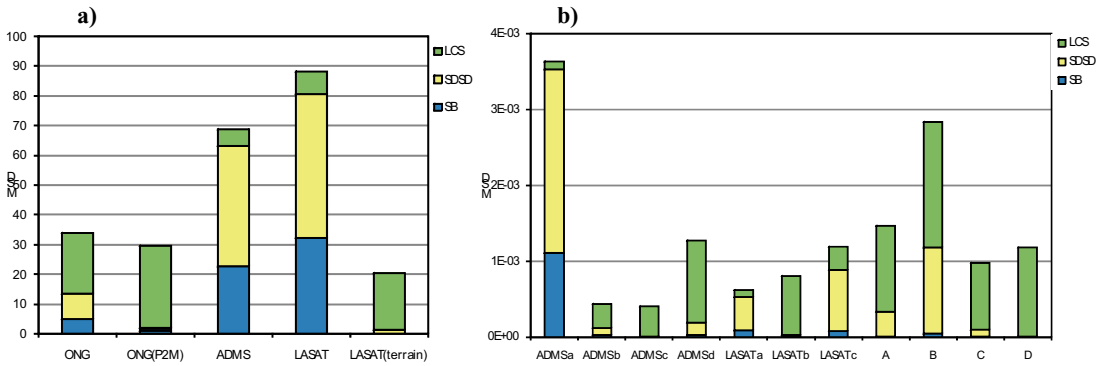

Figure 1. Mean squared deviation (MSD) and its components (see text for explanation) a) OROD (n=152) b) MUST experiment (n=257).

As can be seen from Figure 1a and b, the MSD values of different experiments are not comparable as this statistical measure is dependent on the range of the compared data-sets. For the OROD experiment, model and observational data are given in gm^{-3}, in the MUST experiments, the results are given as dimensionless concentrations.

For the OROD experiment (Fig. 1a), the relatively small MSD for the ONG model is still improved when a peak-to-mean (P2M) approach (Piringer et al., 2007) is applied to the model results. With this P2M-correction, the magnitude of fluctuation of the model data is in agreement with the observations (SDSD is near zero) and the mean values are very close (SB near zero). The remaining MSD is due to failures to reproduce the distribution pattern of the measured plume (according to LCS). The MSD for ADMS is much higher in this case. LASAT applied without terrain data (just considering the relative height difference between the measuring points and the source) renders an even worse MSD, while LASAT based on a regular grid of terrain data (with 3 m distance between the grid points) achieves the best result (the smallest MSD).

In Fig. 1b, MSD for four different ADMS runs of the MUST case (conducted by different groups) are depicted. It is obvious that the performance of the model is strongly dependent on the chosen input configuration and the selection of input parameters as e.g. roughness length: ADMS a (without parameterization of the flow around buildings, roughness length = 0.1 m) gives a comparatively high MSD, ADMS d (no buildings, roughness length =0.381) a better result. ADMS b (with buildings, roughness length =0.1m) and ADMS c (no buildings, roughness length =0.268 m) the results are obviously much improved (MSD=0.0004).

LASAT is applied to this case with the grid oriented according to the north and east direction (LASAT a), in a second run with rotated grid parallel to the side wall of the buildings (LASAT b) and based on three-dimensional wind fields calculated with the CFD model MISKAM (LASAT c). The difference in standard deviations (SDSD) and the difference between the mean values (SB) are decreased significantly from LASAT a to LASAT b while the differences in the distribution pattern of the modelled and observed plume are increased (LCS enlarged). LASAT b therefore renders a higher MSD although the correlation with the measurements is much improved (from 0.5 to 0.8). LASAT c achieves a better agreement concerning the shape of the plume, but the difference in standard deviations is enlarged.

Model A, B, C are results for the MUST experiment achieved with other Gaussian models applied by different groups contributing to COST Action 732, D with a Lagrangian puff model based on diagnostic wind fields. Model run B and C are conducted with the same model with a very small roughness length of 0.037m (B) and a roughness length of 0.3 m (C) which is obviously the better choice in this case. MSD from different CFD model runs conducted for the MUST exercise in COST Action 732 range between 0.001 and 0.003.

In the next step, residual plots as suggested by Hanna et al. (2003) are used to compare the data-sets. The lower and upper ends of the vertical bars represent the 2nd and 98th percentile of the residuals (model value to observation). The lower and upper limits of the boxes indicate the 16th and 84th percentiles of the residuals, the horizontal line in between the median. At least 50% of the model data should be within a factor of 2 to the observations (within the dotted lines) according to Chang and Hanna (2004).

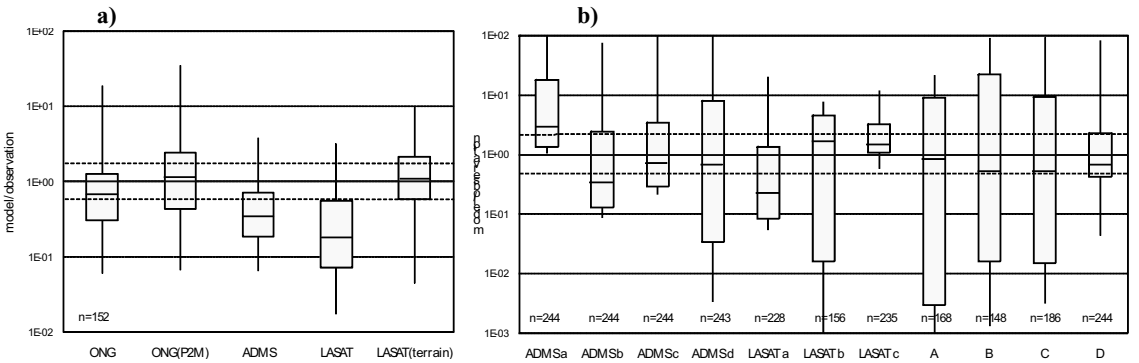

Figure 2. 2nd, 16th, 50th, 84th and 98th percentiles of residuals (predicted to observed concentrations) a) OROD b) MUST experiment.

For the OROD experiment (Fig. 2a), ONG(P2M) achieves on average a closer agreement between model and observations (median close to 1) but a larger range of residuals than ONG. ADMS and LASAT significantly underestimate the observations (more than 98 percent of the model values smaller than the observations (residuals less than 1). LASAT with terrain data gives the best result with the median near 1 and about 95 percent of the model values within a factor of two to the observations (most of the box within the dotted lines).

ADMS a (without buildings) in the MUST experiment (Fig. 2b) significantly overestimates, LASAT a (not rotated grid) significantly underestimates the concentrations (most of the residuals above respectively below 1). The residuals for ADMS d, LASAT b, A, B and C cover a comparatively wide range indicating that at a large portion of points the observations are over- and underestimated by a factor of 10 or more (note: logarithmic scaling on vertical axis). The best agreement between model and observations is found in Figure 2b for LASAT c (based on MISKAM wind fields) and for Model D (Lagrangian puff model).

Finally, the usual performance measures, geometric mean (*MG*), geometric variance (*VG*), fractional bias (*FB*) and normalized mean square error (*NMSE*) are discussed, where

$$MG = \exp\left(\overline{ln\left(\frac{o}{m}\right)}\right),\ FB = \frac{\overline{o} - \overline{m}}{0.5\left(\overline{o} + \overline{m}\right)},\quad VG = \exp\left[\overline{\ln\left(\frac{o}{m}\right)^2}\right] \text{ and } \quad NMSE = \frac{\overline{(o-m)^2}}{\overline{o}\,\overline{m}} \qquad (7)$$

Chang and Hanna (2004) suggest the following values for performance measures as "acceptable" for CFD models based on experience from several model comparison and validation exercises: the fractional bias -0,3 < FB < 0,3 (or the geometric mean 0,7 < MG < 1,3) and the normalized mean square error NMSE < 4 (or the geometric variance: VG < 1,6).

Table 1. Geometric mean (MG), geometric variance (VG), fractional bias (FB) and normalized mean square error (NMSE) and #acceptable' limits as proposed by Chang and Hanna (2004).

a) OROD experiment

	ONG	ONG(P2M)	ADMS	LASAT	LASAT(terrain)	'acceptable'
MG	1.2	1.0	1.6	2.1	1.0	0,7 < MG < 1,3
VG	1.4	1.4	1.8	2.4	1.4	VG < 1,6
FB	0.3	-0.2	0.9	1.1	0.04	-0,3 < FB < 0,3
NMSE	1.3	0.7	5.7	10.5	0.5	NMSE < 4

b) MUST experiment

	ADMS a	ADMS b	ADMS c	ADMS d	LASAT a	LASAT b	LASAT c	A	B	C	D
MG	1.3	0.3	1.1	0.9	2.0	1.4	0.7	1.4	1.0	1.1	1.2
VG	5.2	12.0	2.7	6.1	3.0	4.8	1.6	6.6	16.1	6.4	2.9
FB	-1.1	-1.2	0.03	-0.3	1.2	0.2	-0.7	0.2	-0.5	-0.03	0.1
NMSE	100.8	14.0	2.6	5.7	14.3	6.0	17.5	10.7	10.6	5.6	8.0

For the OROD experiment, ONG, ONG(P2M) and LASAT(terrain) render an 'acceptable' geometric mean MG (Table 1a). The fractional bias FB is also within the 'acceptable' limits for ONG(P2M) and LASAT(terrain). MG as well as FB measure the relative mean biases comparable to SB (Fig. 1a), whereas the normal mean square error NMSE and the geometric variance VG are metrics for the scatter comparable to SDSD (Fig 1a). VG and NMSE are 'acceptable' for ONG, ONG(P2M) and LASAT. For ADMS, these statistical measures are somewhat higher than the suggested limits in this case. The LASAT results without considering the terrain data definitely are rejected. ONG(P2M) and LASAT(terrain) are the only two model runs which are classified 'acceptable' in the OROD experiment according to all four measures.

An 'acceptable' geometric mean in the MUST experiment is calculated for ADMS b, c and d, LASAT b and c and for the model runs B, C and D (Tab. 1b). If FB is compared to the proposed limits, only ADMS c, LASAT b, model run A, C and D are accepted. The geometric variance is larger than the proposed limit of 1.6 for all model runs of the MUST experiment depicted in Table 1b. Only LASAT c (LASAT based on MISKAM wind fields) achieves a VG of 1.6. Nevertheless, the NMSE for this model run is relatively large (17.5). ADMS c is in this case the only model run with an 'acceptable' NMSE.

3. DISCUSSION

Different measures for the comparison of simulated concentrations and observations for two experiments are presented in the previous section. To conclude which model performs best in the two cases, F (Eq. 1) is calculated for each statistical evaluation and finally multiplied to one score per model and experiment: p/q is judged according to the different metrics. In order not to prefer any method, all weights c_{novel} are set to 1.

For the OROD experiment, the Lagrangian model LASAT with terrain data achieves the highest score (F=3), followed by the Gaussian model runs ONG(P2M) and ONG (F=2), while the Gaussian model ADMS and LASAT without terrain data are evaluated as poor.

For the MUST experiment, ADMS c is rated highest (F=2.4). The two Lagrangian model applications, LASAT c (based on MISKAM wind fields) and model run D (based on diagnostic wind fields) are scored marginally good (F=0.5). The evaluation results for the other (non CFD) model runs discussed here are comparatively poor ($F \leq 0.2$).

Applying different statistical measures (supported by a previous graphical investigation of the model results which is not described here) and combining the results with equation 1 as proposed by Sornette et al. (2007) is found a helpful tool for summarizing the numerous results of any model evaluation experiment. The presented results furthermore reveal the importance of an optimum model input (e.g. terrain data, roughness length) as model performance may vary widely for the same model applied with different input parameters. The application of Lagrangian dispersion modelling (with LASAT) based on wind fields simulated with the CFD model MISKAM is found to offer an interesting alternative to CFD simulations for applications with complex building structure as represented by the MUST experiment.

Acknowledgements: *We thank the Environmental Wind Tunnel Laboratory at Hamburg University, especially B. Leitl for the MUST wind channel data. A. Lohmeyer and the BWPLUS program are acknowledged for the data from the Odour Release and Odour Dispersion project. The colleagues of COST Action 732 are thanked for providing their non CFD model results.*

REFERENCES

Bächlin, W., A. Rühling, A. Lohmeyer, 2002: Bereitstellung von Validierungsdaten für Geruchsausbreitungsmodelle – Naturmessungen. *Forschungsbericht FZKA-BWPLUS*, Förderkennzeichen BWE 20003, 183 pp.

Bezpalcová, K., Z. Jaňour, B. Leitl, M. Schatzman, 2005: Concentration Cross-Correlations within Passive Tracer Plumes in Regular Arrays of Obstacles. *Proceedings International Workshop on Physical Modelling of Flow and Dispersion Phenomena*, London, Ontario, Aug. 24-26.

Biltoft, C.A., 2001: Customer report for Mock Urban Setting Test (MUST). DPG Doc. No. WDTC-FR-01-121, West Desert Test Center, U.S. Army Dugway Proving Ground, Dugway, UT 84022-5000.

CERC (Cambridge Environmental Research Consultants), 2001: ADMS 3, Version 3.1, User Guide.

Chang , J.C. and S.R. Hanna, 2004: Air quality model performance evaluation. *Meteorol. Atmos. Phys.*, **87**, 1-3.

Eichhorn, J., 1989: Entwicklung und Anwendung eines dreidimensionalen mikroskaligen Stadtklima - Modells. PhD Thesis, Univ. Mainz.

Hanna, S.R., J. Chang, R. Britter, M. Neophytou, 2003: Overview of Model Evaluation History and Procedures in the Atmospheric Air Quality Area. *QNET-CFD Network Newsletter*, **2**, 5, 1-4.

Harms, F., B. Leitl, M. Schatzmann, 2005: Comparison of tracer dispersion through a model of an idealized urban area from field (MUST) and wind tunnel measurements. *Proceedings International Workshop on Physical Modelling of Flow and Dispersion Phenomena*, London, Ontario Aug. 24-26.

Hirtl, M., and K. Baumann-Stanzer, 2007: Evaluation of two dispersion models (ADMS-Roads and LASAT) applied to street canyons in Stockholm, London and Berlin. *Atmos. Environ.* **41**, 5959-5971.

Hirtl, M., K. Baumann-Stanzer, A. Kaiser, E. Petz, G. Rau, 2007: Evaluation of three dispersion models for the Trbovlje power plant, Slovenia. *Proceedings of the 11th Int. Conf. on Harmonisation*, 2-5.7.2007, Cambridge, UK, 21-25.

Janicke, U., 2005: Ausbreitungsmodell LASAT. Referenzbuch zu Version 2.14.

Kobayashi, K. and M. U. Salam, 2000: Comparing simulated and measured values using mean squared deviation and its components. *Agron. J.*, **92**, 345-352.

Lohmeyer, A., W.J. Mueller, W. Baechlin, 2002: A comparison of street canyon concentration predictions by different modellers: final results now available from the Podbi-exercise. *Atm. Environ.*, **36**, 1, 157-158.

Oreskes, N., K. Shrader-Frechette and K. Berlitz, 1994: Verification, validation and confirmation of numerical models in the earth sciences. *Science*, **263**, 641-646.

Pechinger, U. and E. Petz, 1997: Model evaluation of the Austrian Gaussian model ON M9440: comparison with the Copenhagen and the Lillestrom datasets. *Int.J.of Environment and Pollution*, **8**, No.3-6, 287-294.

Piringer, M., E.Petz, I. Groehn and G. Schauberger, 2007: A sensitivity study of separation distances calculated with the Austrian Odour Dispersion Model (AODM). *Atmos. Environ.*, **41**, 1725-1735.

Sornette, D., A.B. Davis, K. Ide, K.R. Vixie, V. Pisarenko and J.R. Kamm, 2007: Algorithm for model validation: Theory and applications. *PNAS*, **104**, 16, 6562-6567.

Yee, E., and C.A. Biltoft, 2004: Concentration fluctuation measurements in a plume dispersing through a regular array of obstacles. *Bound. Layer Meteorol.*, **111**, 363-415.