THE MUST MODEL EVALUATION EXERCISE: STATISTICAL ANALYSIS OF MODELLING RESULTS

J. Franke[1], J. Bartzis[2], F. Barmpas[3], R. Berkowicz[4], K. Brzozowski[5], R. Buccolieri[6], B. Carissimo[7], A. Costa[8], S. Di Sabatino[6], G. Efthimiou[2], I. Goricsan[9], A. Hellsten[10], M. Ketzel[4], B. Leitl[11], R. Nuterman[12],H. Olesen[4], E. Polreich[13], J. Santiago[14], R. Tavares[8]

[1]Department of Fluid- and Thermodynamics, University of Siegen, Germany
[2]Univ. of West Macedonia, Greece; [3]Aristotle Univ., Greece; [4]NERI, Univ. of Aarhus, Denmark; [5]Univ. of Bielsko-Biala, Poland; [6]Univ. of Salento, Italy; [7]EDF, France; [8]Univ. of Aveiro, Portugal; [9]Budapest Univ. of Technology and Economics, Hungary; [10]Helsinki Univ. of Technology, Finland; [11]Univ. of Hamburg, Germany; [12]Tomsk State Univ., Russia; [13]ZAMG, Austria; [14]CIEMAT, Spain

Abstract: The first validation exercise of the COST action 732 lead to a substantial number of simulation results for comparison with the MUST wind tunnel experiments. Validation metrics for selected simulation results of the flow field and the concentrations are presented and compared to the state of the art. In addition mean metrics and corresponding scatter limits are computed from the individual results.

Key words: *COST 732, MUST, CFD, atmospheric dispersion, model evaluation, model validation, metrics, N-Version testing.*

1. INTRODUCTION

The increasing use of computational simulation for dispersion predictions at micro-scale aggravates the need for a clear and commonly accepted procedure for the quality assurance of the codes and the simulation results. To that end the European COST action 732 *Quality Assurance and Improvement of Micro-Scale Meteorological Models* has proposed a protocol for the evaluation of codes (Britter and Schatzmann, 2007). This protocol has been applied by the action's participants to the mocked urban setting test (MUST) case. For this case both field (Yee and Biltoft, 2004) and wind tunnel measurements (Leitl et al., 2007) are available. In the first step the validation of the codes was done against wind tunnel measurements of velocities and concentrations for several wind directions.

The validation part of the protocol recommends exploratory data analysis and statistical performance measures to assess the predictive capability of the codes. First results of the exploratory data analysis are presented by Olesen et al. (2008) in a companion paper. Here selected results for the statistical performance measures are presented. These measures are defined as validation metrics, allowing a quick assessment of the simulation quality by delivering one value for a large number of measurement points.

Furthermore the metrics are analysed statistically. This is possible because of the large number of simulation results that were produced by the 13 different research groups participating in the exercise. To that end the scatter of the different results is viewed as reproducibility of the computational process. This is known as N-Version testing in computation and equals the N-th order replication in experiments (Coleman and Steele, 1999). The outcome is a collective metric of all simulations, together with scatter limits, which can also be used for the certification of CFD codes (Stern et al., 2006).

2. EXPERIMENTAL SETUP AND DEFINITION OF STATISTICS

Measurement positions

Extensive wind tunnel measurements of the flow field and the concentrations were conducted for several wind directions (Bezplacova, 2007; Harms et al., 2005). In this work only velocity measurements at 18 towers and dispersion measurements for the –45 degree approach flow case are used for the calculation of metrics, as these data are also used in the exploratory data analysis of Olesen et al. (2008). In Figure 1 the tower positions and the concentration measurement positions are shown. Altogether there are 498 Velocity measurement positions and 256 concentration measurement positions.

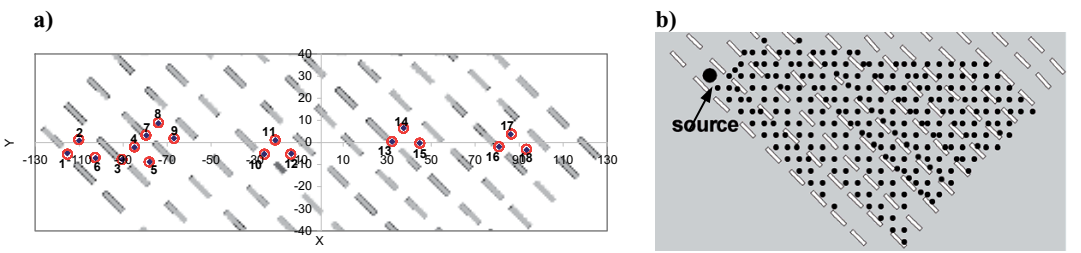

Figure 1. Measurement positions for the –45 degree case. a) towers for velocities, b) concentrations. Wind blows from the left.

Validation metrics

For the velocity data and the concentrations the hit rate q is proposed as one metric (VDI, 2005).

$$q = \frac{1}{I}\sum_{n=1}^{I} i_n \qquad \text{with} \qquad i_n = \begin{cases} 1 & \text{if } \left|(O_n - P_n)/O_n\right| \le \Delta_r \text{ or } \left|O_n - P_n\right| \le \Delta_a \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

Here N = number of measurement positions; O = observed value; P = predicted value; Δ_r = 0.25, the allowed relative difference; Δ_a = allowed absolute difference. For the allowed absolute difference the measurement uncertainties for the corresponding variables are used, see Table 1.

Table 1. Allowed absolute differences Δ_a used in the calculation of the hit rate, Equation (1).

U/U_{ref}	W/U_{ref}	k/U_{ref}^2	C^*
0.008	0.007	0.005	0.003

Here C^* is the scaled concentration

$$C^* = C \cdot U_{ref} \cdot H^2 / Q_{source} \tag{2}$$

where U_{ref} = reference velocity in x-direction at (x,y,z) = (-144,-2.25,7.29) m and Q_{source} = volumetric flow rate of the source, see Figure 1b.

For the concentrations in addition the metrics also used in the BOOT software (Chang and Hanna, 2004) are calculated from the observed (index o) and predicted (index p) concentrations C^*. These are the factor of two $FAC2$, the fractional bias FB, the normalised mean square error $NMSE$, the geometric mean MG and the geometric variance VG, which are defined as

$$\begin{aligned} &FAC2 = \text{fraction of data with } 0.5 \le C_p^*/C_o^* \le 2, \\ &FB = 2\left(\langle C_o^* \rangle - \langle C_p^* \rangle\right)/\left(\langle C_o^* \rangle + \langle C_p^* \rangle\right), \qquad NMSE = \left\langle \left(C_o^* - C_p^*\right)^2 \right\rangle / \langle C_o^* \rangle \cdot \langle C_p^* \rangle, \\ &MG = exp\left(\langle ln\, C_o^* \rangle - \langle ln\, C_p^* \rangle \right), \qquad VG = exp\left[\left\langle \left(ln\, C_o^* - ln\, C_p^* \right)^2 \right\rangle \right], \end{aligned} \tag{3}$$

where angular brackets denote an average over all measurement points. For MG and VG the experimental uncertainty from Table 1 is used as threshold, i.e. for the observed and predicted values $max(0.003, C^*)$ is used. In the calculation of $FAC2$ it is checked for low concentrations if both, observed and predicted are below the threshold. If so, then the position is within a factor of two. All metric calculations are performed within the Excel workbooks described by Olesen et al. (2008).

Statistics of metrics

Assuming that all simulation results belong to a Gaussian or normal parent population the mean Y and the standard deviation S of the sample population are defined as

$$Y = 1/M \sum_{j=1}^{M} X_i, \qquad S = \left[1/(M-1)\sum_{j=1}^{M} \left(X_j - Y\right)^2\right]^{1/2} \tag{4}$$

where X_j = one of the metrics presented above for simulation j and M = 20 is the number of simulation results. As this definition of the mean is known to be easily contaminated by outliers (Müller, 2000), additionally the median Z of the sample population is computed and the standard deviation T based on the sample median absolute deviation (MAD), see e.g. Hemsch (2000).

$$T = 1/0.6745 \sqrt{M/(M-1)}\, median\left(\left|X_i - Z\right|\right) \tag{5}$$

From the standard deviations scatter limits are computed as P_S = 2.093 S and P_T = 2.093 T, corresponding to 95% confidence intervals around the mean, $Y \pm P_S$ and $Z \pm P_T$ (Coleman and Steele, 1999).

3. RESULTS

Up to 30 simulation results from 13 different research groups are at present available. The statistical results are shown for 20 of these simulation results, which are all obtained with prognostic models, comprising general purpose commercial CFD codes, commercial micro-scale CFD codes and in house CFD or micro-scale meteorological models. As a general overview of the predictive capabilities shall be given the results are presented anonymously and no ranking of the codes is performed.

Metrics for the flow field at the towers of the –45 degree case

In Figure 2 the hit rates for the two measured velocity components at the 18 towers are shown. Also included is the lower limit of q = 0.66 as broken line, which is used by the VDI (2005) for its test cases to define successful validation.

While the velocity component in flow direction, U/U_{ref}, is predicted very well by nearly all simulations with hit rates larger than 0.66, the vertical velocity component, W/U_{ref}, is not well predicted by all simulations. Olesen et al. (2008)

showed that at the towers W/U_{ref} is mostly underpredicted. This shortcoming is however well reflected in the low hit rates. The quickly available metrics are therefore useful for identifying possible problems that can then be further analysed by an detailed inspection of the simulation results at the measurement positions. This does however not mean that a detailed analysis of the data is obsolete in the case of good metrics. E.g. the high rates of U/U_{ref} are mostly due to the expected good agreement between simulation and experiment well above the containers, while below container height the agreement deteriorates.

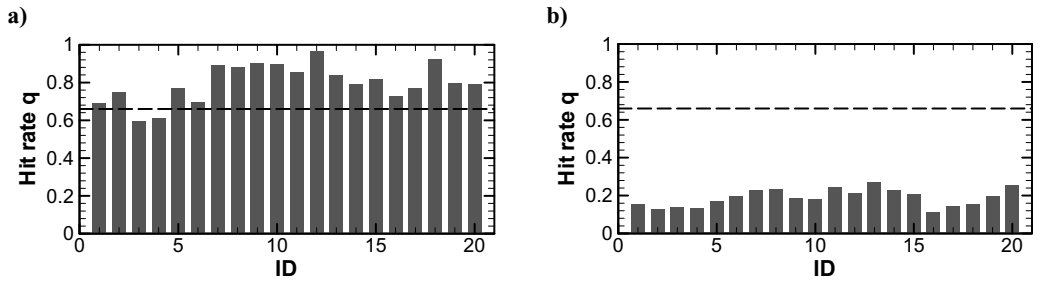

Figure 2. Hit rate for mean velocity components at the towers for the –45 degree case. a) U/U_{ref}, b) W/U_{ref}.

The hit rate for the turbulent kinetic energy k/U_{ref}^2 at all towers is shown in Figure 3a. Here large differences between the different simulations are visible. As k/U_{ref}^2 is a positive quantity also other metrics which are normally only used for concentrations are meaningful and can be analysed. E.g. the large positive *FB*s in Figure 3b clearly indicate that those simulations with low hit rates underpredict the turbulent kinetic energy.

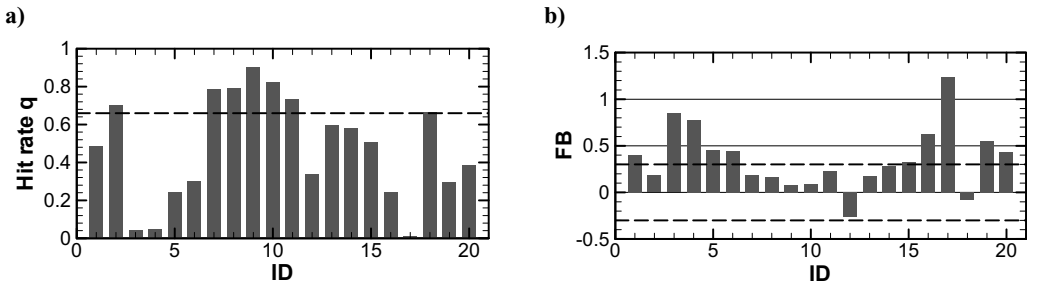

Figure 3. Metrics for the turbulent kinetic energy k/U_{ref}^2 at the towers for the –45 degree case. a) Hit rate *q*, b) *FB*.

Metrics for the concentrations of the –45 degree case

The metrics of dispersion results are shown in Table 2. Numbers in italics are results which are not within the range defined by Chang, J.C. and S.R. Hanna (2004) as state of the art, $FAC2 > 0.5$, $|FB| < 0.3$, $NMSE < 4$, $0.7 < MG < 1.3$ and $VG < 1.6$.

Table 2. Dispersion metrics for the –45 degree case. Italics indicate results being not state of the art.

Simulation ID	Hit rate q	FAC2	FB	NMSE	MG	VG
1	*0.46*	0.71	*-0.54*	*7.52*	0.82	*1.65*
2	*0.65*	0.87	*-0.36*	*7.07*	0.93	1.23
3	*0.39*	0.65	*-0.91*	*20.39*	*0.69*	*1.79*
4	*0.37*	0.67	*-0.82*	*13.08*	0.72	*1.66*
5	*0.43*	0.64	*-0.50*	*8.80*	0.94	*1.67*
6	0.80	0.88	*-0.33*	*10.34*	1.00	1.26
7	0.74	0.86	-0.30	*4.52*	0.86	1.23
8	*0.57*	0.89	*-0.38*	2.72	0.84	1.20
9	0.68	0.97	-0.27	1.65	0.86	1.10
10	*0.42*	0.58	*0.40*	*4.19*	*1.36*	*1.70*
11	*0.54*	0.83	*0.41*	3.54	*1.42*	1.36
12	*0.63*	0.79	0.27	1.76	*1.34*	1.43
13	*0.48*	0.62	*-0.57*	*8.49*	0.89	*1.72*
14	*0.38*	0.52	*-0.62*	*12.51*	0.86	*2.27*
15	0.79	0.86	*-0.41*	*9.83*	0.81	1.41
16	*0.58*	0.79	0.15	2.89	1.15	1.37
17	*0.64*	0.82	*-0.36*	*4.56*	0.89	1.39
18	*0.64*	0.95	*-0.40*	2.21	0.84	1.13
19	*0.54*	0.73	*-0.67*	*8.65*	*0.68*	*2.05*
20	*0.52*	0.79	*-0.43*	*4.94*	0.87	1.35

While there are only four simulations that have larger hit rates for the concentrations than 0.66, the limit defined by the VDI (2005), all simulations have $FAC2 > 0.5$. Looking at all simulation results there is no obvious correlation between the flow field results presented in the previous section and the concentration metrics. While the low hit rates of simulations 3 and 4 correspond to low hit rates for U/U_{ref} (Fig. 2a) and k/U_{ref}^2 (Fig. 3a), simulation 17 has a rather high hit rate for the concentrations despite the extremely low hit rate for k/U_{ref}^2. The most probable reason for this behaviour is that errors in the computation of the flow field and errors in the prediction of the concentrations cancel each other, leading to rather good results for the concentration metrics. Taking all metrics into account is therefore necessary to identify situations where good results are predicted due to the wrong reasons.

From Table 2 it can be also seen that most of the simulations overpredict the concentrations as a negative *FB* corresponds to an overprediction. In general the concentrations close to the source (see Fig. 1b) are overpredicted. The large differences at these few positions lead to *FB*s with too high magnitudes. Therefore only four simulations have a *FB* within the range defined as state of the art. From these four results one has *NMSE* larger than 4, indicating that this simulation substantially overpredicts as well as underpredicts the observations. While the over- and underpredictions cancel in *FB*, they lead to a too high *NMSE*. As can be seen very clearly from Figure 4a) only three simulations therefore have both, *FB* and *NMSE* in the range defining the state of the art.

The geometric metrics *MG* and *VG* are not dominated by high concentration values. Therefore they do not suffer from the overpredictions close to the source. Only five simulations have a geometric mean *MG* outside the range defining state of the art. However, if also the results for the geometric variance are taken into account, there are only 10 simulations within the admissible range spanned by both metrics, see Figure 4b).

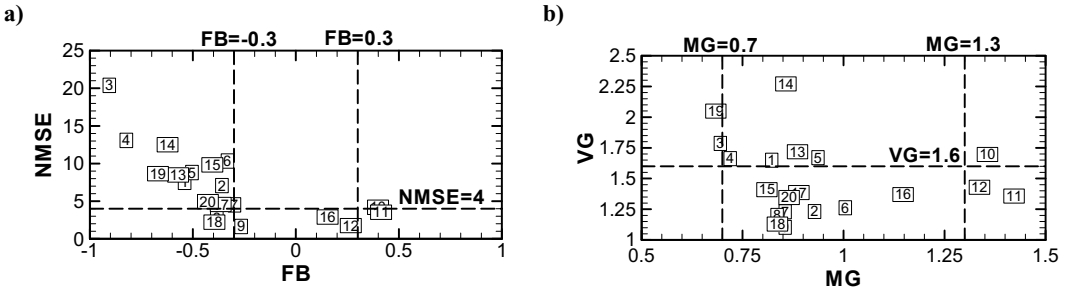

Figure 4. Metrics for the concentrations C^* of the –45 degree case. a) *NMSE* vs. *FB*, b) *VG* vs. *MG*.

Statistics of dispersion metrics for the –45 degree case

In Figure 5 the running records of the hit rate q and the *FB* are shown together with the corresponding mean, median and associated scatter limits defining the 95% confidence interval. On the basis of statistics differences of the individual values within the confidence limits are considered as noise. Only those simulation results which lie outside the confidence intervals are then significant for further analysis.

For the hit rate the mean and the median are equal up to two digits, $Y = Z = 0.56$. The 95% confidence interval around the mean value is $P_S = 0.28 = 0.5\,Y$. For the median it is $P_T = 0.31 = 0.55Z$. Both intervals are large with regards to the mean value. All simulation results are within these intervals, there are no outliers. For *FB* the situation is different. Based on the median, $Z = -0.39$, and its confidence interval, $P_T = 0.44 = 1.13|Z|$, there are 5 outliers. The mean value, $Y = -0.33$, and its even larger confidence interval with $P_S = 0.76 = 2.30|Y|$ do not have any outliers. This demonstrates the better suitability of the median to identify outliers which in the case of the mean value are only reflected by a larger confidence interval.

Except for *FAC2*, *MG* and *VG* the collective metrics mean and median are all outside the limits defined as state of the art. This displays a systematic shortcoming of the models and indicates the need for at least an improved model set up, if not for model improvement.

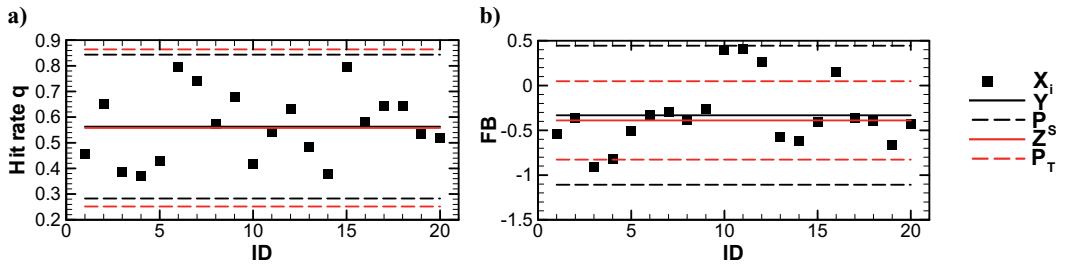

Figure 5. Running records of the metrics for the concentrations C^* of the –45 degree case. a) Hit rate q, b) FB.

4. CONCLUSIONS

Within the first validation exercise of the COST action 732 for the MUST wind tunnel experiment a large number of simulation results was produced by different research groups, using different models. The validation metrics for 20 of these results have been presented for one wind direction, comprising metrics for the flow field at selected locations and metrics for the concentrations at all concentration measurement positions. While all concentration results have a $FAC2 > 0.5$, other metrics are often outside the range defined as state of the art. In fact there is only one simulation result with all concentration metrics being state of the art.

This shortcoming of the individual results is also reflected in the collective results for the metrics, computed by statistical methods with the assumption that the 20 individual results are samples from a Gaussian parent population. From the collective metrics only $FAC2$, MG and VG are within the range being state of the art. Another result from this statistical analysis is that the median as more robust estimate for the sample mean value, together with its standard deviation based on the median of the absolute deviations, is better suited for identifying outliers, i.e. individual simulations that differ substantially from the core region of the statistical distribution.

These results for the validation metrics indicate at least a need for an improved model set up for the MUST test case. After further detailed analysis of the individual simulation results recommendations for the conduction of simulations for the MUST case will be included in the official documents of the action, which can be found through URL 1.

REFERENCES

Bezpalcova, K., 2007: Physical Modelling of Flow and Dispersion in an Urban Canopy, PhD thesis, Faculty of Mathematics and Physics, Charles University, Prague, 193 pp.

Britter, R. and M. Schatzmann, 2007: Model evaluation guidance and protocol document, COST office, Belgium, 28 pp.

Chang, J. C. and S.R. Hanna, 2004: Air quality model performance evaluation. *Meteo. Atmos. Phys.,* **87**, 167-196.

Coleman, H.W. and W.G. Steele, 1999: Experimentation and Uncertainty Analysis for Engineers, 2nd Edition, John Wiley & Sons, USA, 275 pp.

Harms, F., B. Leitl and 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, August 24-26.

Leitl, B., K. Bezpalcova and F. Harms, 2007: Wind Tunnel Modelling Of The MUST Experiment. *11th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes,* Cambridge, July 2-5.

Hemsch, M., 2002: Statistical analysis of CFD solutions from the drag prediction workshop. *AIAA paper,* 2002-0842.

Müller, J.W., 2000: Possible advantages of a robust evaluation of comparisons. *J. Res. Natl. Inst. Stand. Technol.*, **105**, 551-555.

Olesen, H.R., A. Baklanov, J. Bartzis, F. Barmpas, R. Berkowicz, K. Brzozowski, R. Buccolieri, B. Carissimo, A. Costa, S. Di Sabatino, G. Efthimiou, J. Franke, I. Goricsan, A. Hellsten, M. Ketzel, B. Leitl, R. Nuterman, E. Polreich, J. Santiago and R. Tavares, 2008: The MUST model evaluation exercise: Patterns in model performance. *12th Int. Conf. on Harmonisation within Atmospheric Dispersion Modelling for Regulatory Purposes,* Cavtat, October 6-9.

Stern, F., R. Wilson and J. Shao, 2006: Quantitative V&V of CFD simulations and certifications of CFD code. *Int. J. Numer. Meth. Fluids*, **50**, 1335-1355.

VDI, 2005: Environmental meteorology – Prognostic microscale windfield models – Evaluation for flow around buildings and obstacles. VDI guideline 3783, Part 9, Beuth, Berlin.

Yee, E. and C.A. Biltoft, 2004: Concentration Fluctuation Measurements in a Plume Dispersing Through a Regular Array of Obstacles. *Boundary-Layer Meteorology,* **111**, 363-415.

URL 1: Official web site of COST 732: http://www.mi.uni-hamburg.de/index.php?id=464