H14-132
HOW TO CHOOSE THE BEST SIMULATION FOR A SPECIFIC PURPOSE?

A. Martilli[1], J. L. Santiago[1], T. G. Reisin[2], A. Baklanov[3], J. Bartzis[4], R. Buccolieri[5], A. M. Costa[6], S. Di Sabatino[5], G. Efthimiou[4], J. Franke[7], A. Hellsten[8], R. Nuterman[3,,9], R. Tavares[10]

[1]Unidad de Modelización y Ecotoxicidad de la Contaminación Atmosférica, CIEMAT, Madrid, Spain
[2]SOREQ NRC, Yavne, 81800, Israel
[3]Danish Meteorological Institute, Copenhagen, Denmark
[4]University of West Macedonia, Dept of Mechanical Engineering, Environmental Technology Lab, Kozani, Greece
[5]Dipartimento di Scienza dei Materiali, University of Salento, Lecce, Italy
[6]IDAD - Instituto do Ambiente e Desenvolvimento, Aveiro, Portugal
[7] Department of Fluid and Thermodynamics, University of Siegen, Germany
[8] Finnish Meteorological Institute, Finland
[9]Mechanics and Mathematics Faculty, TomskStateUniversity, Tomsk, Russia
[10]CESAM & Department of Environment and Planning, University of Aveiro, Portugal

Abstract: The decision on the fitness for purpose of a simulation should be based on the quantity of interest. However, in general, models are used because there is no complete experimental information available on the quantity of interest, so a direct judgement is not possible. The aim of this article is to put in light this dichotomy, and propose a methodology to decide if a simulation is fit for purpose, based on the experimental data available and an ensemble of simulations. The methodology is illustrated with one example of microscale simulations.

Key words: *Validation, fit-for-purpose,air quality, modelling* **.**

INTRODUCTION

Numerical models are used to complete the information obtained from measurements. Whether they are used to assess air quality, to evaluate air pollution reduction strategies, to forecast pollution levels, or for research purposes, the main information expected from a simulation is an estimate of one, or more, quantities of interest (QI), that may vary with the purpose of the study, but for which measurements are not available. There can be several reasons why experimental values of QI are not available. Among the most important:

- They can be technically difficult to obtain. For example, if QI is the maximum of concentration in a certain region, or the area with a concentration above a threshold, it is often difficult or even impossible, to have a measurement network dense enough to recover this information.
- They can refer to a "hypothetical" situation. This is the case, for example, when QI is the amount of pollution reduction due to a certain abatement strategy.
- They can refer to the future, as in the case of a forecast.

In the modelling activity there may be several sources of uncertainties. These can be due to model formulation (e. g. parameterizations), numerics, or to the lack of detailed knowledge of the initial or boundary conditions. For the same case study, it is often possible to produce an ensemble of simulations by varying model's options (or by using different models), or by perturbing initial and boundary conditions. The problem is, then, to choose among the simulations of the ensemble, the simulations that fit a specific purpose (e. g. results are good enough that they can be used for the intended application).

We are stating here that this choice should be based on the distance between the value of QI of the real world, and the simulated value of the same quantity (SQI). With the word "distance" we indicate a statistical operator or a metric (function of QI and SQI, $d_{purpose}(QI,SQI)$) that can measure, in a quantitative manner, how "close" QI and SQI are. It is important to stress the quantitative aspect because to separate fit from non-fit for purpose simulations, it will be necessary to define an acceptance criteria *H*, such that all the simulations with $d_{purpose}(QI,SQI) < H$ can be accepted, while those with $d_{purpose}(QI,SQI) > H$ are not accepted. The choice of $d_{purpose}$, and H must depend on the purpose of the simulation.

The problem here is that, as explained above, the real value of QI is not known. This is the motivation of model use (obviously, if QI was known no modelling is needed). The judgement on simulation fitness must be based on some experimental quantities (EQ) and on the simulated values of these quantities (SEQ). We need to find another distance (or metric) based on EQ and SEQ, $d_{Xbest}(EQ,SEQ)$, that can surrogate $d_{purpose}(QI,SQI)$. Similarly we will need to find another separator value K that can be used to decide if a simulation is fit or not fit for the purpose. In summary, if simulation *j*computes SQI_j, SEQ_j, we are looking for a d_{Xbest} and K such that

$$d_{Xbest}(EQ,SEQ_j) < K \Leftrightarrow d_{purpose}(QI,SQI_j) < H .$$

In the next section we propose a methodology, based on ensemble of simulation, to define d_{Xbest} and K.

METHODOLOGY

To search for the d_{Xbest} and K for a specific purpose, it is necessary to consider a situation where $d_{purpose}$ and d_{Xbest} can be computed. This situation can be given by the ensemble of simulations itself. The proposed steps are as follows:

- For each couple of simulations*i,j* a distance is calculated based on $d_{purpose}\left(SQI_j, SQI_i\right)$. Note that this is possible to do because SQI can be derived from simulations, since simulations give an approximate, but complete representation of the real world.
- The same is done for several possiblemetrics d_X , involving simulated experimental quantities, that are candidate to be the surrogate of $d_{purpose}$. The following distances are computed $d_X\left(SEQ_i, SEQ_j\right)$.
- The best surrogate metric d_{Xbest} is the one that gives the most similar ranking of simulation couples to the ranking of $d_{purpose}$, and those that gives the best values of the separator *K*.

Before explaining how the metrics can be compared, few comments.The basic assumption behind this methodology is that if $d_{purpose}\left(SQI_i, SQI_j\right)$ and $d_{Xbest}\left(SEQ_i, SEQ_j\right)$ behaves in a similar way (e. g. metrics applied to simulation-to-simulation comparison), also $d_{purpose}\left(QI_i, SQI_j\right)$ and $d_{Xbest}\left(EQ_i, SEQ_j\right)$ will do (metrics applied to real world-to-simulation comparison). Clearly this is a leap of logic, and there is no warranty that this will happen. The validity of the assumption depends on how realistic are the results produced by the different simulations (with the term realistic here we indicate that they represent an equally likely physical state of the atmosphere, not that they represent exactly the state under study). This is the reason why it is important that the models used to produce the simulations passed a Scientific Evaluation and Verification steps. Moreover, we think that this logical leap is different in nature, but not bigger, than the logical leap implicitly done by assuming that the error performed by the model at the measurement points is the same error performed in estimating the quantity of interest, which is the assumption usually done.

It is also important that the members of the ensemble are well chosen. It goes beyond the scope of this paper to provide a rigorous criterion to define a proper ensemble design, and in this work it is assumed that the results of the methodology do not depend on the specific ensemble design.

Kendall's Tau

One technique to compare rankings is based on the Kendall's Tau test (Kendall M., 1938). This is computed as $\tau_{kendall} = \frac{n_c - n_d}{N^2}$, where *N* is the total number of simulation couples, n_c is the total number of duplets of simulation couples such that one of the two relationships below holds

$$\begin{cases} d_{purpose}\left(SQI_i, SQI_j\right) > d_{purpose}\left(SQI_k, SQI_m\right) \\ d_X\left(SEQ_i, SEQ_j\right) > d_X\left(SEQ_k, SEQ_m\right) \end{cases} \text{or} \begin{cases} d_{purpose}\left(SQI_i, SQI_j\right) < d_{purpose}\left(SQI_k, SQI_m\right) \\ d_X\left(SEQ_i, SEQ_j\right) < d_X\left(SEQ_k, SEQ_m\right) \end{cases}$$

while n_d is the number of duplets of simulation couples such that none of the two relationships above holds. The range of values for $\tau_{kendall}$ is [-1,1]. The highest the value of $\tau_{kendall}$ the closest is d_X to $d_{purpose}$. In a certain sense $\tau_{kendall}$ is a "measure" of the probability that if simulations *i,j* are closer between them than *k,m* for d_X , they will also be closer between them for $d_{purpose}$. Based on this, d_{Xbest} is the one that has the highest $\tau_{kendall}$.

Separation value

Here, we are looking for a *K* such that if $d_{Xbest}\left(SEQ, EQ\right) \leq K\left(> K\right)$, then also $d_{purpose}\left(SQI, QI\right) \leq H\left(> H\right)$, where *H* is the quantitative acceptance criterion for a simulation to fit the purpose. Again, the idea is to infer the value of *K* based on the intercomparison between simulations. This is done by looking for the *K* that maximizes the number of simulation couples such that $d_{Xbest}\left(SEQ_i, SEQ_j\right) \leq K$ and $d_{purpose}\left(SQI_i, SQI_j\right) \leq H$, or $d_{Xbest}\left(SEQ_i, SEQ_j\right) > K$ and $d_{purpose}\left(SQI_i, SQI_j\right) > H$. More formally, this can be computed defining an index m_{ij} for each couple *(i,j)* as below, and then sum the index over all the couples and divide by the number of couples.

$$m_{ij} = \begin{cases} 1 \Leftrightarrow \left[\left(d_{purpose}\left(SQI_i, SQI_j\right) - H\right) \cdot \left(d_{Xbest}\left(SEQ_i, SEQ_j\right) - K\right)\right] > 0 \\ 0 \quad else \end{cases}$$

$$s(K) = \frac{\sum_{ij} m_{ij}}{N(N-1)}.$$

Here, *s(K)* represents the fraction of simulations couples for which *K* is a good separator. The best separator K_{best} is the onethat gives the highest *s*. In other words, *s(K)* is the probability (based on our set of simulations) that the judgment based on d_{Xbest} , is the same judgment that could be obtained using $d_{purpose}$.

EXAMPLES

In the frame of the European COST 732 Action on Quality Assurance and Improvement of Microscale Meteorological Models, a model intercomparison exercise has been carried out with the aid of the wind tunnel reproduction of the "Mock Urban Setting Test" (MUST, Fig. 1) experiment (Bezpalcova K. and F. Harms, 2005). The wind tunnel data set provided measured information on the global flow field within and above the array of obstacles for wind directions 45° respect to the regular arrangement of obstacles at three different heights above ground. The point release was located at ground level, and concentrations were measured at 0.5 H_{obst}, where H_{obst} is the obstacle height. Flow measurements are available at 0.5 H_{obst}, H_{obst}, 2 H_{obst} and at several selected vertical profiles For this case, several Computational Fluid Dynamics (CFD) models were employed to simulate flow and dispersion (Schatzman M. et al., 2009). In particular, for our analysis 17 simulations were selected, obtained by seven CFD models, used by 10 modellers (see Table 1). A description of the differences between the 17 simulations can be found in Schatzmann M. et al. (2009). Each model passed a Scientific Evaluation.

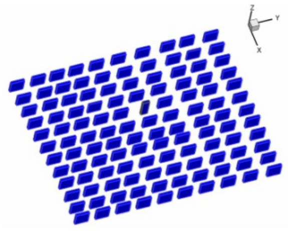

Figure 1. The MUST array.

From the 17 simulations, it is possible to form 272 couples of simulations[1]. In order to test the methodology, we make the hypothesis that the quantity of interest *QI* can be derived from the tracer concentration *at the measurement points*. Then we will apply the methodology to the flow quantities, as we would do if the concentration measurements were not available. This will allow us to select d_{Xbest}. For this special case, where $d_{purpose}(SQI,QI)$ and $d_{Xbest}(SEQ,EQ)$ can both be measured, it is possible to asses if the choice is good or not.

Table 1. Models, developers and users that simulated the MUST case in the framework of the COST 732 action.

Model	*Developer*	*Users*
FINFLO	HelsinkiUniversity of Technology, Finland	Hellstein
FLUENT	ANSYS (commercial code)	Franke, Goricsan, Santiago, Buccolieri
M2UE	Tomsk State University, Russia, and Danish Meteorological Institute	Nuterman, Starchenko and Baklanov
MISKAM	University of Mainz, Germany	Ketzel, Goricsan
STAR CD	CD-ADAPCO (commercial code)	Brzozwski
VADIS	University of Aveiro, Portugal	Costa and Tavares
ADREA	Environmental Research Laboratory of NCSR “Demokritos”, Greece	Efthimiou and Bartzis

We assume, then, that *QI* is the maximum of concentration *at the measurement points* that we want to know with an accuracy of 50%. The appropriate $d_{purpose}(SQI_i,QI)$ is:

$$d_{purpose}(SQI_i,QI)=d_{\max\_rel}(SQI_i,QI)=2\frac{\left|\max(C_i(x))-\max(C_e(x))\right|}{\max(C_i(x))+\max(C_e(x))}$$

Where $max(C_i(x))$ is the maximum concentration at the measurement points estimated by simulation*i*, and $max(C_e(x))$ is the maximum concentration in the experimental data. The acceptance criterion *H* is fixed to 0.5.

For this case, two measurement datasets are available: (1) horizontal (*x* and *y)* components of the wind, and turbulent kinetic energy (TKE) from a horizontal array of points, and (2) *x* and *z* components of the wind and TKE from a series of vertical profiles. Six metrics d_X were considered:

- $d_{hrvv}(M_i,M_j)=1.-HitRate(vect_i,vect_j)$, based on the hit rate computed from horizontal wind velocities at the measurement points of the horizontal array. Relative threshold is 0.25 and absolute 0.014 ms^{-1}
- $d_{hrdd}(M_i,M_j)=1.-HitRate(dir_i,dir_j)$, based on the hit rate computed from horizontal wind directions at the measurement points of the horizontal array. Threshold is 10 Degree.
- $d_{hrtke}(M_i,M_j)=1.-HitRate(tke_i,tke_j)$, based on the hit rate computed from TKE at the measurement points of the horizontal array.Relative threshold is 0.25 and absolute 0.01 m^{2}s^{-2}
- $d_{hrvxz}(M_i,M_j)=1.-HitRate(vx_i,vx_j)$, based on the hit rate computed from x-component of the wind velocities at the measurement points of the vertical profiles.Relative threshold is 0.25 and absolute 0.014 ms^{-1}
- $d_{hrvzz}(M_i,M_j)=1.-HitRate(vz_i,vz_j)$, based on the hit rate computed from z-component of the wind velocities at the measurement points of the vertical profiles. Relative threshold is 0.25 and absolute 0.014 ms^{-1}

[1] Some of the metrics are not symmetric.

- $d_{hrtkez}\left(M_i, M_j\right) = 1. - HitRate\left(tkez_i, tkez_j\right)$, based on the hit rate computed from TKE at the measurement points of the vertical profiles. Relative threshold is 0.25 and absolute 0.01 m^{2}s^{-2}

Many other metrics can be implemented (see for example Hanna S. et al. 2004), but these are used because they are among the most commons.

Kendall's Tau

Results of the Kendall's tau are presented in Figure 2 (grey bars) and they clearly show that the best metrics are d_{hrvv} (horizontal wind velocity), and d_{hrvxz} (*x* component of the wind from the vertical profiles). Then there is d_{hrtke} (TKE), followed by d_{hrdd} (wind direction). Last is the vertical velocity d_{hrvzz}.To check if this result is valid, a similar index is computed based on the 17 couples Simulations-Observations. This is possible since, in the way we built the example, there are experimental data for *QI*. For each metric ($d_{purpose}$ and d_X) presented in the previous section a ranking of the 17 couples is formed, then these rankings are compared using theKendall's Tau test. In the large majority of the cases, the $\tau_{kendall}$ computed from the simulation-to-simulation (*sts*) intercomparison is lower than the $\tau_{kendall}$ computed from the simulation-to-observation (*sto*) comparison (black bars), but the tendency is similar. Both analyses agree that d_{hrvv} and d_{hrvxz} are the metrics that best surrogate $d_{purpose}$, corroborating the methodology.

A physical interpretation is that the most important variable that determines the maximum of concentration is the horizontal wind speed. TKE is also important, and finally the vertical velocity is the less important variable. The vertical concentration transport is dominated by turbulence and not by the mean vertical velocity. *It must be stressed, that this conclusion is valid only for this specific case.*

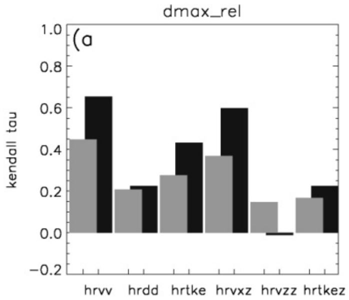

Figura 2. Black bars are for simulations-to-observations, while grey bars are from simulation-to-simulation intercomparisons. *hrvv* hit rate for horizontal velocity, *hrdd* hit rate for direction, *hrtke* hit rate for TKE from horizontal array of measurements, *hrvxz* hit rate for x wind component from the profiles, *hrvzz* hit rate for vertical velocity from the profiles, *hrtkez*, hit rate for TKE from profiles.

Separation value

The previous analysis clearly shows that the best surrogate metrics are those involving the horizontal velocity (d_{hrvv}, d_{hrvxz}) and the one with TKE from the horizontal array d_{hrtke} . The following analysis focuses only on these three metrics. As mentioned above, the acceptance criterion is 50% ($H = 0.5$). The values found for the best separator K_{best} are in Table 2.Such values can be also reported on a graph (Fig. 3). There, the vertical dotted line represents $d_{purpose} = H$, while the horizontal dotted line is representing $d_{Xbest} = K_{best}$ The value of $s\left(K_{best}\right)$ is the fraction of points which are in the lower left and upper right quadrant defined by the two lines. K_{best} is the value that maximizes this fraction. $s\left(K_{best}, Obs\right)$is the fraction of simulation-observation couples that are in the lower left and upper right quadrant. The highest is the value of $s\left(K_{best}, Obs\right)$ the more robust is the methodology. d_{hrvv} is the metrics that gives the highest value $s\left(K_{best}\right)$=0.77 with $K_{best} = 0.34$. This means that in 77% of the cases, one of the two following relationship is true:

$d_{hrvv}\left(M_i, M_j\right) \le 0.34 \Rightarrow d_{\max\_rel}\left(M_i, M_j\right) \le 0.5$ or $d_{hrvv}\left(M_i, M_j\right) > 0.34 \Rightarrow d_{\max\_rel}\left(M_i, M_j\right) > 0.5$

While in the remaining 23% of the cases none is true.

The analysis of the couples simulation-observation (equal to the number of models, 17), shows a lower score for $s\left(K_{best}, Obs\right)$of 0.59, meaning that only in 59% of the cases (10 over 17) one of the following is true:

$d_{hrvv}\left(M_i, O\right) \le 0.34 \Rightarrow d_{\max\_rel}\left(M_i, O\right) \le 0.5$ or $d_{hrvv}\left(M_i, O\right) > 0.34 \Rightarrow d_{\max\_rel}\left(M_i, O\right) > 0.5$

However, there are two couples simulation-observation, which are very close to the lower right and upper left sectors. If we accept these two points as good, the percentage will increase to 70% (12/17), closer to the value obtained from the simulation-to simulation intercomparison.The *x*-component of the velocity derived from the vertical profiles d_{hrvxz} gives similar information, but with a different value of $K_{best} = 0.47$. So, the separation value depends not only on the variable measured, but also on the distribution of measurements points.The TKE derived from the horizontal array (d_{hrtke}) has also a good $s(K_{best})$ of 0.7, and $K_{best} = 0.77$. The comparison with measurements is in agreement with $s(K_{best})$. If a more stringent condition is required for d_{max_rel}, for example *H*=0.35 (relative difference less than 35% for the maximum), different values are found, as presented with the dashed lines in Fig. 3.

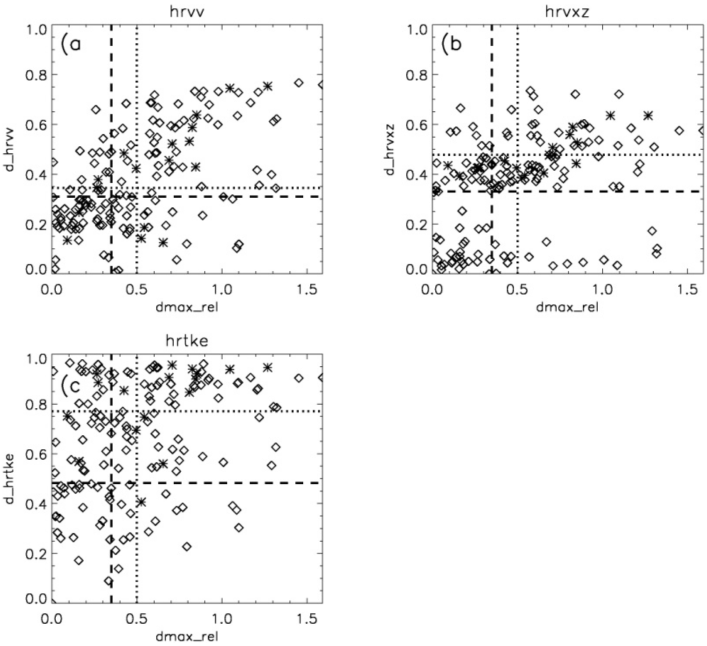

Figure 3. d_{max_rel} vs. d_{hrvv} (a), d_{hrvxz} (b) and d_{hrtke} (c). The diamonds represent the couples of simulations, while the asterisks are the comparison between simulations and observations.

Table 2. Best separators (K_{best} , second column), fraction of couples of simulations for which K_{best} is a good separator ($s(K_{best})$, third column), and fraction of simulations for which K_{best} is a good separator ($s(K_{best}, Obs)$, fourth column) respect to the observations. This is for $d_{purpose} = d_{max\_rel}$, H=0.5, and the metrics indicated in the first column.

d_{max_rel} , H=0.5	K_{best}	$s(K_{best})$	$s(K_{best}, Obs)$
d_{hrvv}	0.34	0.77	0.59
d_{hrtke}	0.77	0.70	0.65
d_{hrvxz}	0.47	0.71	0.70

CONCLUSIONS

The methodology presented in this study has been applied with different $d_{purpose}$ to the same MUST simulations, and also to other case studies with encouraging results. An important conclusion is that while the quantitative acceptance criterion *H*based on $d_{purpose}$ is only a function of the purpose, the criterion K_{best} , based on d_{Xbest} , is also a function of the specific case under study and the distribution of measurements. The important consequence is that is useless to fix universal acceptance criteria, rather, the important target is to define a methodology to deduce such values for every case study and distribution of measurements, which is what we attempted to do in this study.

REFERENCES

Bezpalcova, K., F. Harms, 2005: EWTL Data Report / Part I: Summarized Test Description Mock Urban Setting Test, report, Environmental Wind Tunnel Laboratory, Center for Marine and Atmospheric Research, University of Hamburg, Germany.

Hanna S. R., O. R. Hansen, S. Dharmavaram, 2004: FLACS CFD air quality model performance evaluation with Kit Fox, MUST, Prairie Grass, and EMU observations. *Atmospheric Environment*, **38**, 4675–4687.

Kendall, M., 1938: A New Measure of Rank Correlation, *Biometrika*, **30**, 81-89.

Schatzmann, M., H. Olesen, and J. Franke editors, 2009: COST 732 model evaluation case studies: approach and results, COST Office. available online at:
http://www.mi.uni-hamburg.de/fileadmin/files/forschung/techmet/cost/cost_732/pdf/5th_Docu_May31.pdf