

Validation of numerical cloud models – can we learn from other atmospheric models? Dispersion and microscale meteorological models as an example.

Tamir Reisin

Acknowledgments: Alberto Martilli, David Steinberg,
COST Action 732 members, DTRA/DPG, ZMK/ZMAW

Motivation

- Is the model I'm using "appropriate" to simulate the case in study (a priori)? (ICCP)
- Model evaluation: model-measurements or model-model; is there any "choice"?
- Is a "good" comparison between model and measurements for a certain case a guarantee that the model will perform well in a "similar" case?

* The "words" require more precise definitions

Questions:

- What is the purpose of using (cloud) numerical models?
 - ✦ Basic physics ('theoretical problems')
 - ✦ Phenomenology (explaining what we measure/see)
 - ✦ Prediction (before or without measurements)

Questions:

- What's peculiar in clouds?
 - ✦ Large spatial and temporal variability,
 - ✦ where we are in the cloud,
 - ✦ at what stage of the cloud life we are measuring (when was the cloud “born”?)
 - ✦ Not all the physics well understood (ice phase!)
- What is done in other atmospheric numerical models?

COST Action 732



**BACKGROUND AND JUSTIFICATION
DOCUMENT TO SUPPORT THE MODEL
EVALUATION GUIDANCE AND
PROTOCOL**

Edited by:
Rex Britter and Michael Schatzmann

**COST Action 732
QUALITY ASSURANCE AND IMPROVEMENT OF
MICROSCALE METEOROLOGICAL MODELS**

1 May 2007



**MODEL EVALUATION GUIDANCE AND
PROTOCOL DOCUMENT**

Edited by:
Rex Britter and Michael Schatzmann

**COST Action 732
QUALITY ASSURANCE AND IMPROVEMENT OF
MICRO-SCALE METEOROLOGICAL MODELS**

1 May 2007



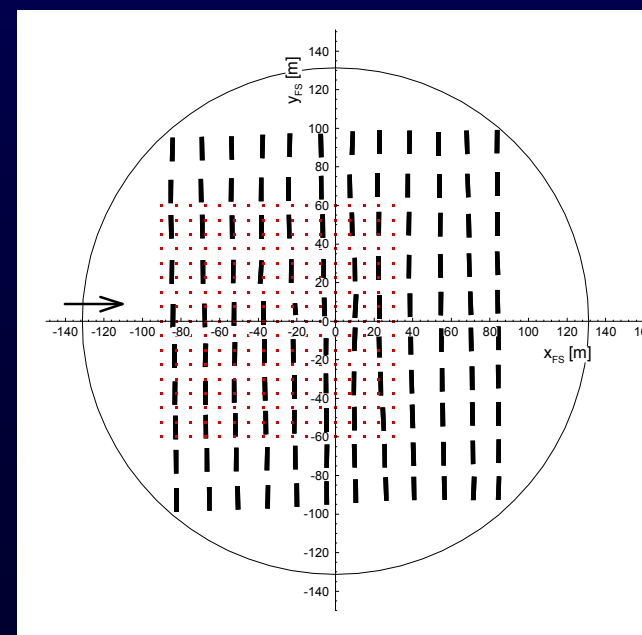
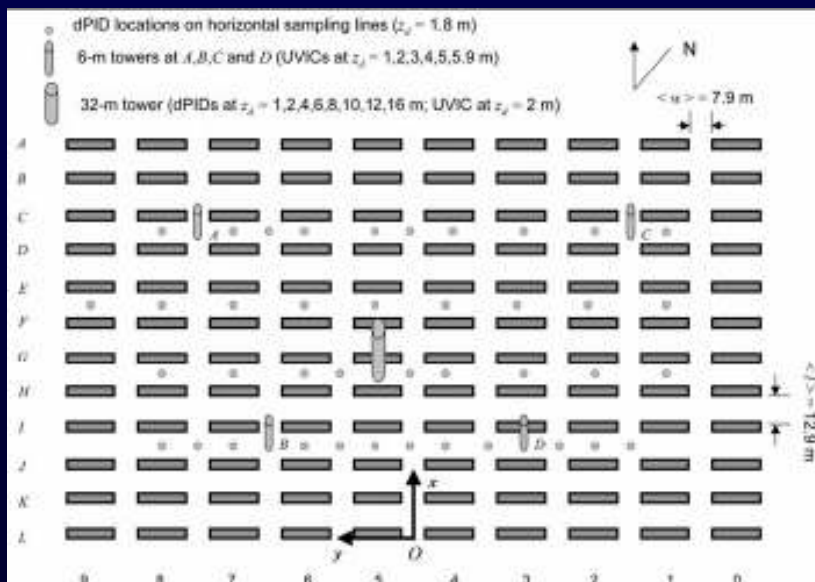
**BEST PRACTICE GUIDELINE
FOR THE CFD SIMULATION OF FLOWS
IN THE URBAN ENVIRONMENT**

Edited by:
Jörg Franke, Anli Heister, Heiko Schatzler, Bernd Carlsino

**COST Action 732
QUALITY ASSURANCE AND IMPROVEMENT OF
MICROSCALE METEOROLOGICAL MODELS**

1 May 2007

The MUST Experiment: Testing microscale meteorological and dispersion models



The MUST Experiment

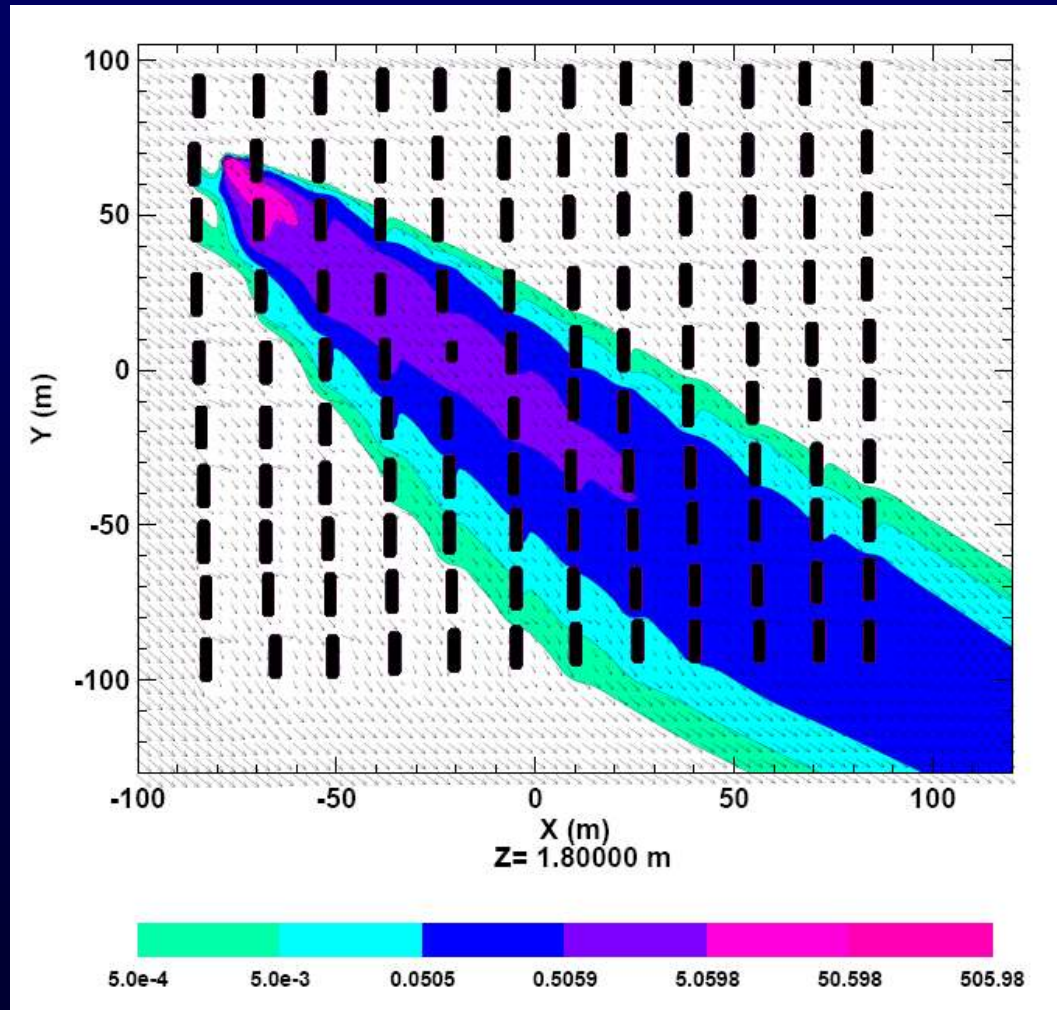
Attention!

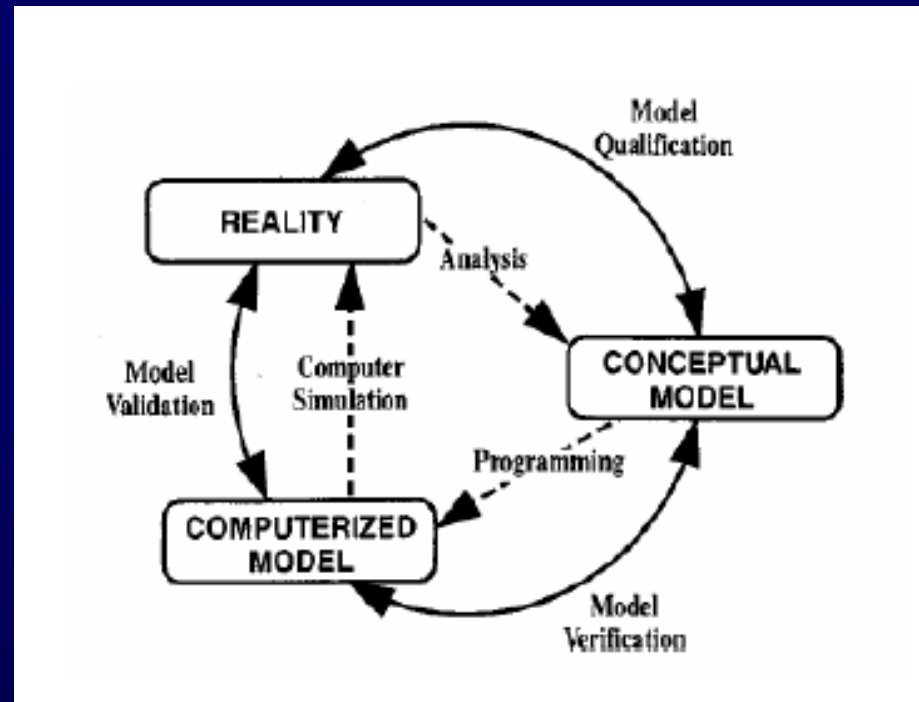
This experiment incorporates three “independent” tools:

- Field experiment
- Wind Tunnel
- Numerical modelling

In the near future a similar approach to Oklahoma City Experiment

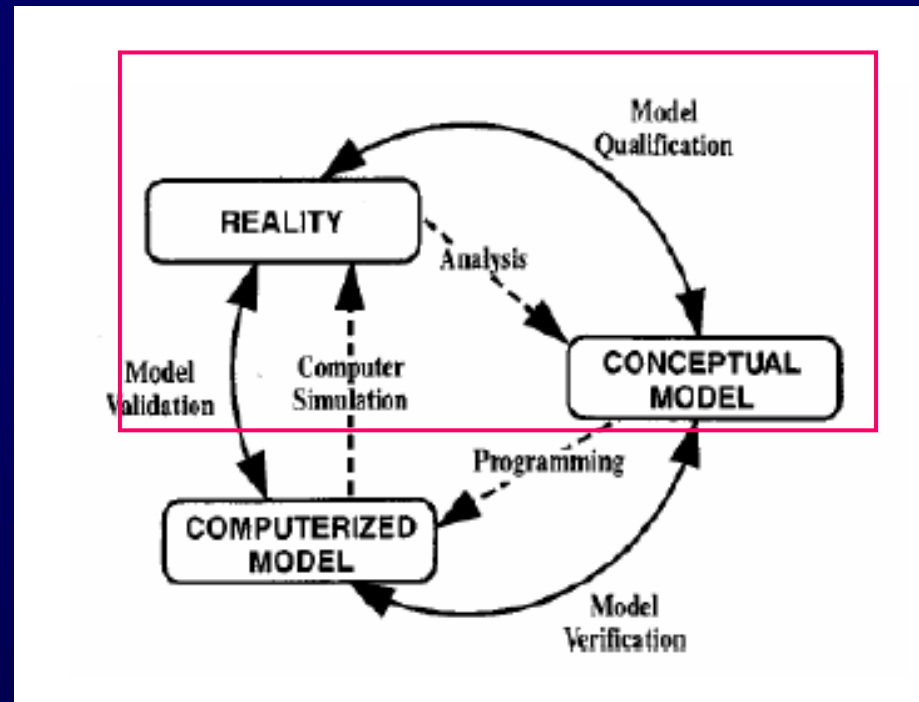
Numerical simulation using the RAMS model





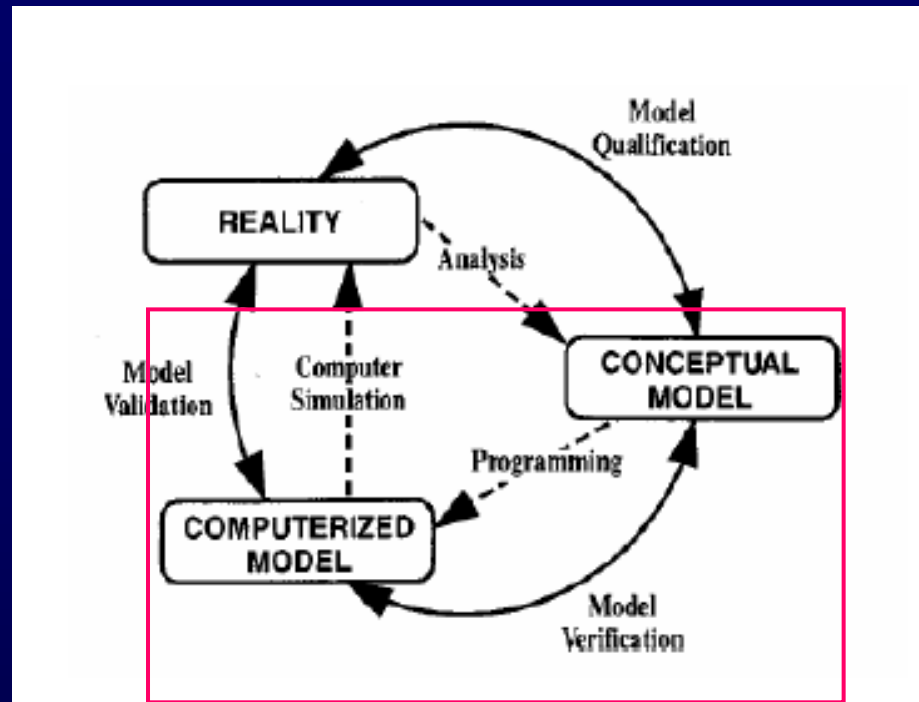
The evaluation protocol has several distinct elements:

- A Scientific Evaluation Process (model qualification)
- A Verification Process that addressed both the code and the solution procedure
- The provision of appropriate and quality assured Validation Data Sets; in particular it was preferred that both field data and associated physical modelling data were utilised
- A Model Validation (Statistical Evaluation may be a better term) - Process in which model results are compared with the experimental data sets.
- An Operational Evaluation Process that reflects the needs and responsibilities of the Model User



A model evaluation requires a Scientific Evaluation (model qualification) whose primary objectives are:

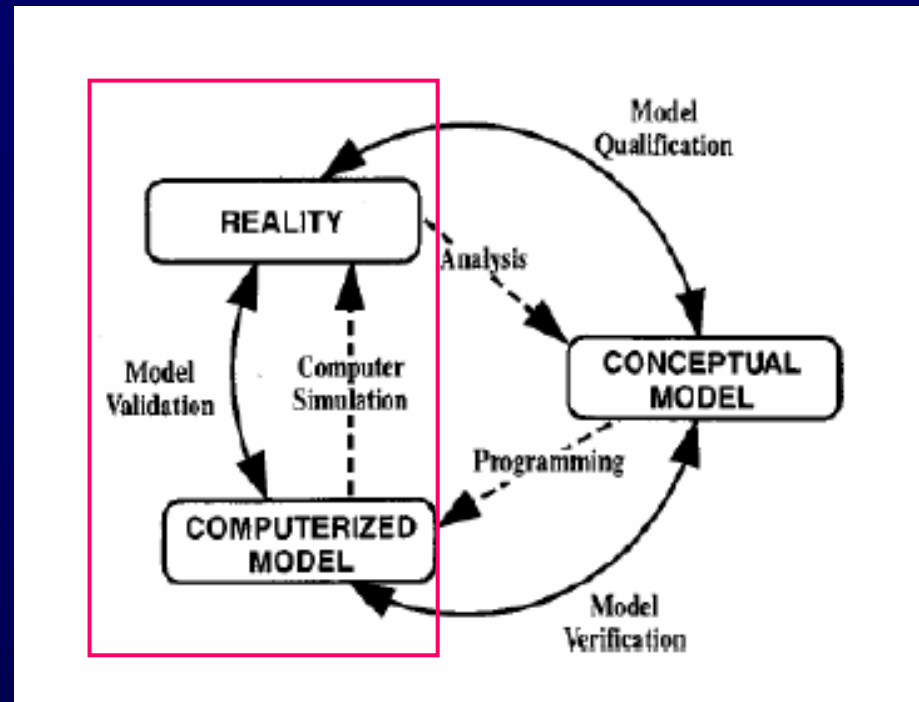
- to ensure that all important phenomena within the models' ranges of application are included
- to ensure that the mathematical modeling of these phenomena and the associated simplifications and parameterizations are well justified in terms of science and model practicality and
- that the limits of model applicability are clear and explicit
- to ensure that the user or prospective user is able to judge the suitability or not of the model for their specific purpose



Verification of models and their results

Verification is an important part of the model evaluation. There are two aspects of verification; the verification of the models also known as code verification and the verification of the solution procedure also known as solution verification.

- Code verification is used to demonstrate that the computational model is consistent with the conceptual model.
- Solution verification is used to estimate the numerical error in a given numerical solution. Both aspects are present in CFD and non-CFD models.



Model validation is the comparison of model predictions with experimental observations (or possibly with the results from more complex models that have themselves undergone formal model evaluation). It is important that there is consensus amongst the scientific and wider community as to the appropriateness of the validation process. In order to conduct a validation, one will have to decide for which purpose (*fitness-for-purpose*) the model results should later be used and thus to decide the variable(s) whose prediction is the most important. In other words the *validation objectives* have to be determined.

Best Practice Guidelines:

Before running a model and before comparing results with other models!

Choice of ...

- Common:
 - * Target variables
 - * Approx. eqs. describing the physics of the flow
 - * Computational domain
 - * BC, IC
 - * Computational grid
 - * Time step
 - * Numerical approximations
 - * Parameterizations
- Clouds: microphysics (bulk, detailed, coefficients, etc.)
- Dispersion: scalar tracer like / lagrangian particles

Measured (and derived) parameters (in situ)

○ Dispersion (gas)

- * u, v, w
- * $P, T, (RH)$
- * Concentration
- * **TKE** (u', v', w')
- * **Boundary layer**
(e.g. u^* , MO
length)

○ Clouds

- * u, v, w
- * P, T, RH
- * **LWC, IC**
- * **Aerosols/Drops/Ice
Spectra**

Remote: Met. Radar, Lidar, Satellites, etc.

Metrics for Model Validation

- **FAC2** = Fraction of predictions within a factor of two of the observations.
- **FB** (Fractional Bias) = $\frac{(\overline{C_o} - \overline{C_p})}{[0.5(\overline{C_o} + \overline{C_p})]}$
- **NMSE** (Normalized Mean Square Error) = $\frac{(\overline{C_o} - \overline{C_p})^2}{\overline{C_o} \overline{C_p}}$
- **MG** (Geometric Mean) = $\exp(\overline{\ln C_o} - \overline{\ln C_p}) = \exp(\overline{\ln(C_o / C_p)})$
- **VG** (Geometric Variance) = $\exp[\overline{(\ln C_o - \ln C_p)^2}] = \exp[\overline{(\ln(C_o / C_p))^2}]$
- **Figure of Merit (FoM)** is defined as the ratio of the area marked by the intersection of the predicted and observed concentrations or dosages within a prescribed contour, divided by the union of the two areas
- **Measure of Effectiveness (MoE)** assigns different weights to false positives and false negatives.
- **Hit Rate**

$$q = \frac{N}{n} = \frac{1}{n} \sum_{i=1}^n N_i \quad \text{with} \quad N_i = \begin{cases} 1 & \text{for } \left| \frac{P_i - O_i}{O_i} \right| \leq D \text{ or } |P_i - O_i| \leq W \\ 0 & \text{else} \end{cases}$$

To evaluate the model performance, normalized values are compared, with the wind speed used for normalisation. From the normalised model results P_i and normalised comparison data O_i a hit rate q is calculated from the equation below, which specifies the fraction of model results that differ within an allowed range D from the comparison data. D accounts for the relative uncertainty of the comparison data. Only those differences are counted that are above a threshold value W , which describes the repeatability of the comparison data.

Additional practices

- Sensitivity analysis
- Uncertainty analysis
- Model intercomparison (comparisons with already evaluated models or more detailed models, comparisons between parameters that were not measured but can be calculated, e.g. Center of Gravity)
 - Not for choosing the “best model” rather for establishing “state-of-the-art”

A methodology to find a “fit for purpose” validation criteria

Question: how to decide if a model/simulation is fit for a specific purpose, in a quantitative way?

Need for an appropriate metric

$$d_{\text{purpose}}(M, R)$$

that “measures” the distance between the simulation results M and the “real world” R .

Define a numerical value, H based on the required accuracy for the purpose, to discriminate between fit and unfit models.

$$d_{\text{purpose}}(M, R) \begin{cases} \leq H & \text{OK!} \\ > H & \text{NO!} \end{cases}$$

$$d_{\text{purpose}}(M, R)$$

Must involve the variables of interest for the purpose, and it must be possible to derive such variables from model’s outputs.

Examples:

I want to know the *maximum of pollutant concentration* (worst case), (or *maximum of LWC*) and I will accept (based on criteria dependent on the purpose) an accuracy of 50%. For this case, then:

$$d_{purpose}(M, R) = 2 \frac{|\max(M) - \max(R)|}{\max(M) + \max(R)}, \quad H = 0.5$$

I want to know the *pollutant concentration* (or the *LWC*) in a certain region with a precision of 25% and I will accept (based on criteria dependent on the purpose) models able to fulfill this condition in at least 66% of the domain. For this case, then:

$$d_{purpose}(M, R) = 1. - HitRate(M, R), \quad H = 0.34$$

Is the variable needed to compute $d_{purpose}$ measured, or not?

YES

Compare model results and measurements using $d_{purpose}$ and decide if the model is fit for purpose or not

NO

For example if there are no concentration measurements or if they are not well distributed in the domain, etc.

Find another metric d_x (involving measured variables, either concentration at certain points or dynamical variables) that can work as surrogate of $d_{purpose}$.

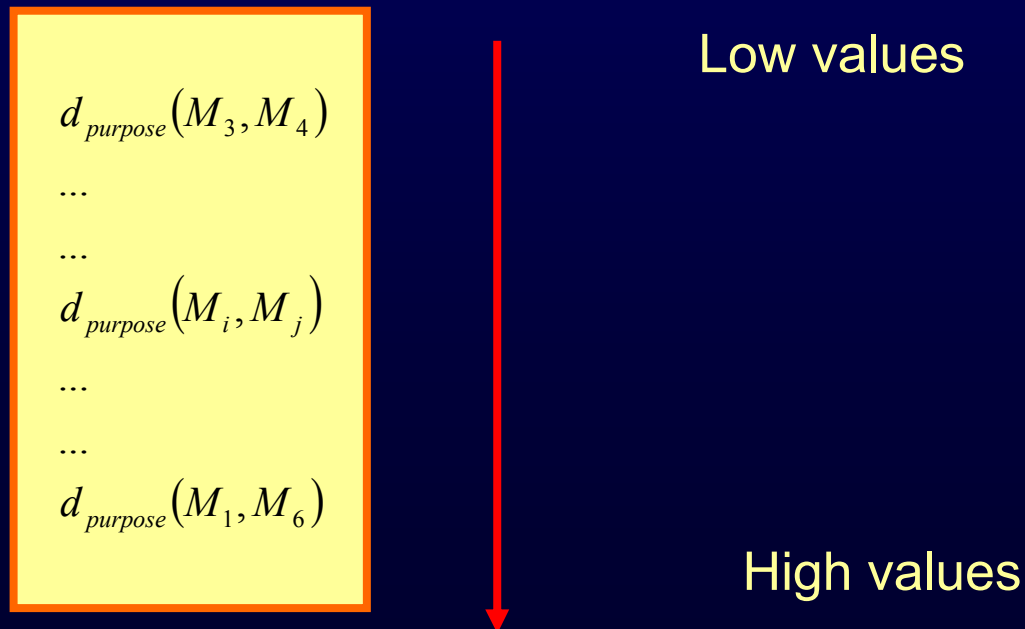
Use d_x to decide if the model is fit for purpose or not.

How to compare two metrics, if one does not involve measured variables?

The idea is to use a model intercomparison technique (if many model results are available as the case of MUST-COST732).

Technique:

- 1) From N models, $N*(N-1)/2$ couples of models (M_i, M_j) can be formed.
- 2) For every couple a distance between models can be computed using $d_{purpose}(M_i, M_j)$
- 3) The couples of models can be ranked based on the $d_{purpose}(M_i, M_j)$



With the same technique rankings for metrics involving measured variables ($d_{X_1}, d_{X_2}, d_{X_3}$ etc.) can be computed

$$d_{X_1}(M_3, M_4)$$

...

...

$$d_{X_1}(M_i, M_j)$$

...

...

$$d_{X_1}(M_3, M_6)$$

$$d_{X_2}(M_1, M_4)$$

...

...

$$d_{X_2}(M_i, M_j)$$

...

...

$$d_{X_2}(M_3, M_1)$$

$$d_{X_3}(M_2, M_4)$$

...

...

$$d_{X_3}(M_i, M_j)$$

...

...

$$d_{X_3}(M_3, M_2)$$

The metric d_{X_i} that gives the most similar ranking to $d_{purpose}$ is the best surrogate.

How to measure similarity between rankings?

- Kendall's tau
- Lift Curve

Using model intercomparison to look at relationships between metrics-
pros and cons:

- Advantage: Model results are complete, the metrics involving the variable of interest can be computed.
- Disadvantage: Model results are not reality (but are the best approximation we have...)

Since a metric is a measure of the capability of a model to reproduce a certain physical aspect of the real world, the comparison between metrics can provide information also on the most important physical mechanisms for a certain purpose.

Example from the MUST-COST 732 results. Models analysed:

1. *'FINFLO_Hellstein_1'*
2. *'FINFLO_Hellstein_2'*
3. *'FLUENT_mskespudf_Franke'*
4. *'FLUENT_RSM_Goricsan'*
5. *'FLUENT_Santiago'*
6. *'FLUENTske_DiSabatino'*
7. *'M2UE_Nuterman_Baklanov'*
8. *'MISKAM_Ketzel'*
9. *'MISKAM05res_Goricsan'*
10. *'STAR_CD_Brzozwski_fine'*
11. *'VADIS_Costa_05m'*
12. *'ADREA_Bartzis'*
13. *'FINFLO_Hellstein'*
14. *'MISKAM_Ketzel_varRoughness'*
15. *'FLUENT_Goricsan_k-e'*
16. *'Miskam_Goricsan_1mes'*
17. *'Miskam_Goricsan_08mes'*

Test 1 (a good one).

Purpose: maximum of concentration.

Accuracy allowed: 50%

$$d_{purpose}(M_i, R) = 2 \frac{|\max(C_i) - \max(R)|}{\max(C_i) + \max(R)}$$

If concentrations were not measured, which is the best surrogate metrics based on dynamical variables?

$d_{hrvv}(M_i, M_j)$ (1-Hit rate) for wind speed in the horizontal grid, $W=0.014$, $D=0.25$

$d_{hrdd}(M_i, M_j)$ (1-Hit rate) for wind direction in the horizontal grid. $W=10^\circ$

$d_{hrke}(M_i, M_j)$ (1-Hit rate) for TKE in the horizontal grid, $W=0.01$, $D=0.25$

$d_{hrvxz}(M_i, M_j)$ (1-Hit rate) for U in the profiles, $W=0.014$, $D=0.25$

$d_{hrvzz}(M_i, M_j)$ (1-Hit rate) for W in the profiles, $W=0.014$, $D=0.25$

$d_{hrtkez}(M_i, M_j)$ (1-Hit rate) for TKE in the profiles, $W=0.01$, $D=0.25$

Metrics with dynamic variables at measurements points.

I computed the rankings for $d_{purpose}$, and the other metrics with the 17 models (136 couples).

Kendall's tau test

The first test to compare the metrics is the Kendall's tau test.

if

$$\begin{aligned}
 & d_{purpose}(M_i, M_j) \geq d_{purpose}(M_k, M_m) \text{ and } d_X(M_i, M_j) \geq d_X(M_k, M_m) \\
 & d_{purpose}(M_i, M_j) < d_{purpose}(M_k, M_m) \text{ and } d_X(M_i, M_j) < d_X(M_k, M_m)
 \end{aligned}
 \Rightarrow \alpha_{ijkm} = 1$$

...	...
$d_{purpose}(M_i, M_j)$	$d_X(M_i, M_j)$
...	...
$d_{purpose}(M_k, M_m)$	$d_X(M_k, M_m)$
...	...

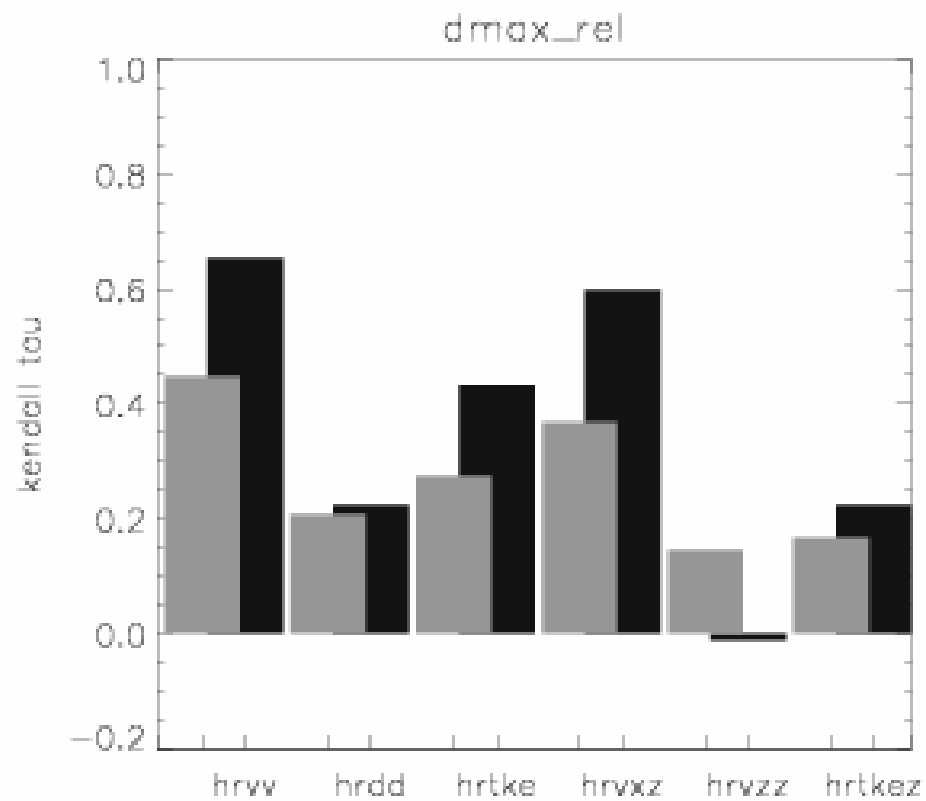
...	...
$d_{purpose}(M_k, M_m)$	$d_X(M_k, M_m)$
...	...
$d_{purpose}(M_i, M_j)$	$d_X(M_i, M_j)$
...	...

else

$$\Rightarrow \alpha_{ijkm} = -1$$

$$\tau_{kendall} = \frac{\sum_{ijkm} \alpha_{ijkm}}{N_c(N_c - 1)}$$

The highest the index, the most similar the rankings are.



Model to model

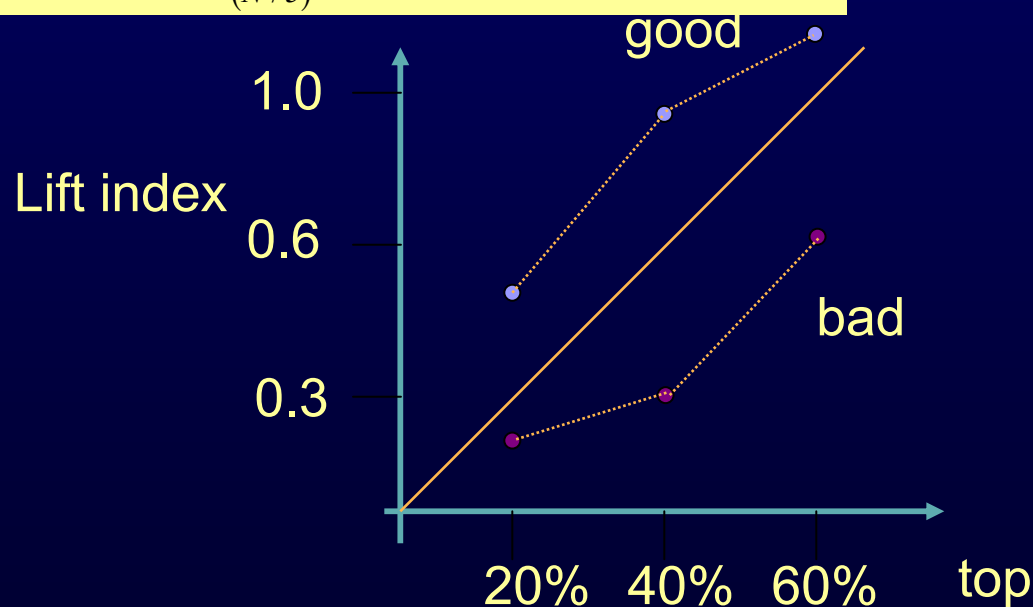
Model to observation

Lift curve

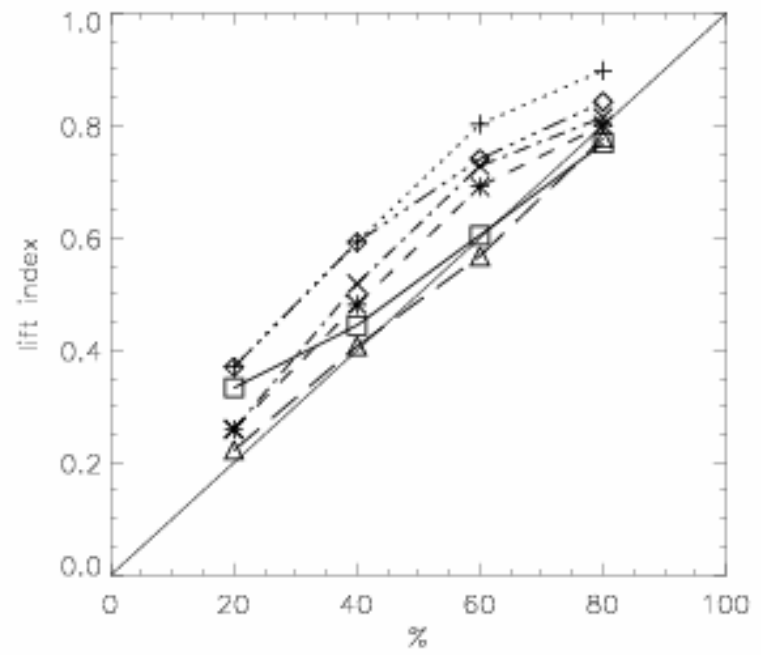
The lift index is the percentage of the top 20% (40%, 60%, ...) model's couples for the $d_{purpose}$ ranking, which are also in the top 20% (40%, 60%, ...) of the d_x ranking.

$$m_{ij} = \begin{cases} 1 & \Leftrightarrow [Rnk(d_{purpose}(M_i, M_j)) < N/5] \cap [Rnk(d_x(M_i, M_j)) < N/5] \\ 0 & \text{else} \end{cases}$$

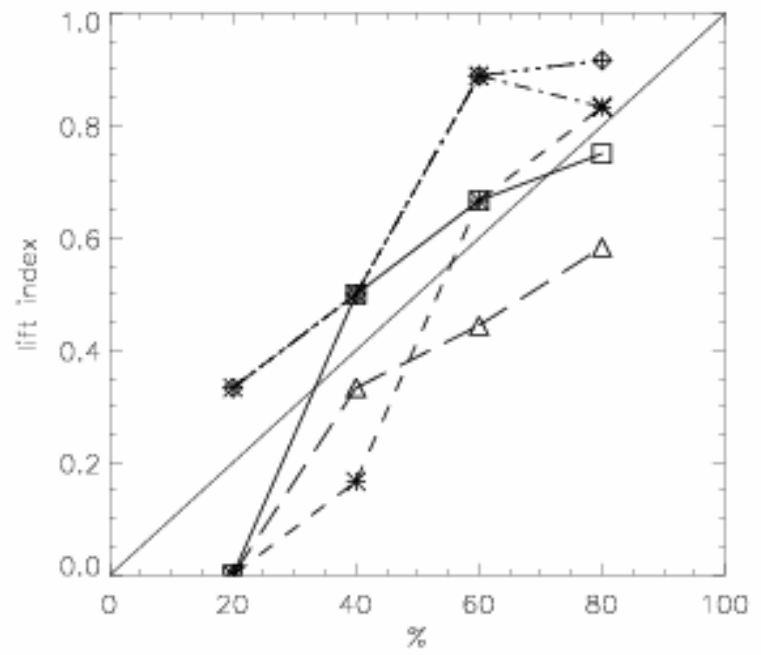
$$l(d_{purpose}, d_x, 20\%) = \frac{\sum_{ij} m_{ij}}{(N/5)}$$



dmax_rel Model



dmax_rel Observations



- ++hrvv
- * - - -*hrdd
- x ······xhrtke
- ◇ ······◇hrvxz
- △ - - -△hrvzz
- - - -□hrtkez

First conclusion:

hrvv (hit rate involving horizontal wind speed) seems to be the best metric for the purpose.

hrvzz ((hit rate involving vertical velocity from profiles) seems to be the worst metric for the purpose.

Horizontal wind speed more important than vertical velocity
for the maximum of concentration ?

In any case, this conclusion is valid only for this configuration (obstacles well spaced), and this distribution of measurements

Separation values

How to find a K such that, given H the following is true (or at least highly probable)?

$$d_{purpose} \leq H \Leftrightarrow d_X \leq K$$

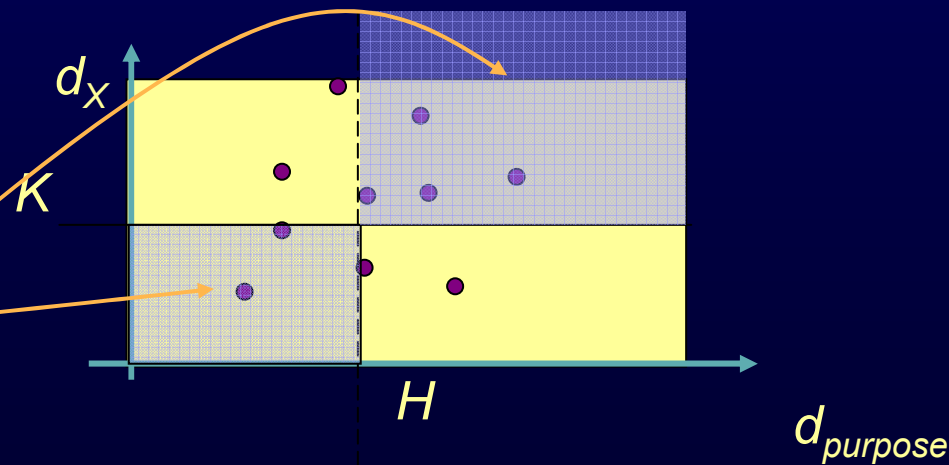
I defined a function s such that

$$\text{if } [d_{purpose}(M_i, M_j) \leq H \cap d_X(M_i, M_j) \leq K] \text{ and } [d_{purpose}(M_i, M_j) > H \cap d_X(M_i, M_j) > K] \Rightarrow m_{ij} = 1$$

$$\text{else } m_{ij} = 0$$

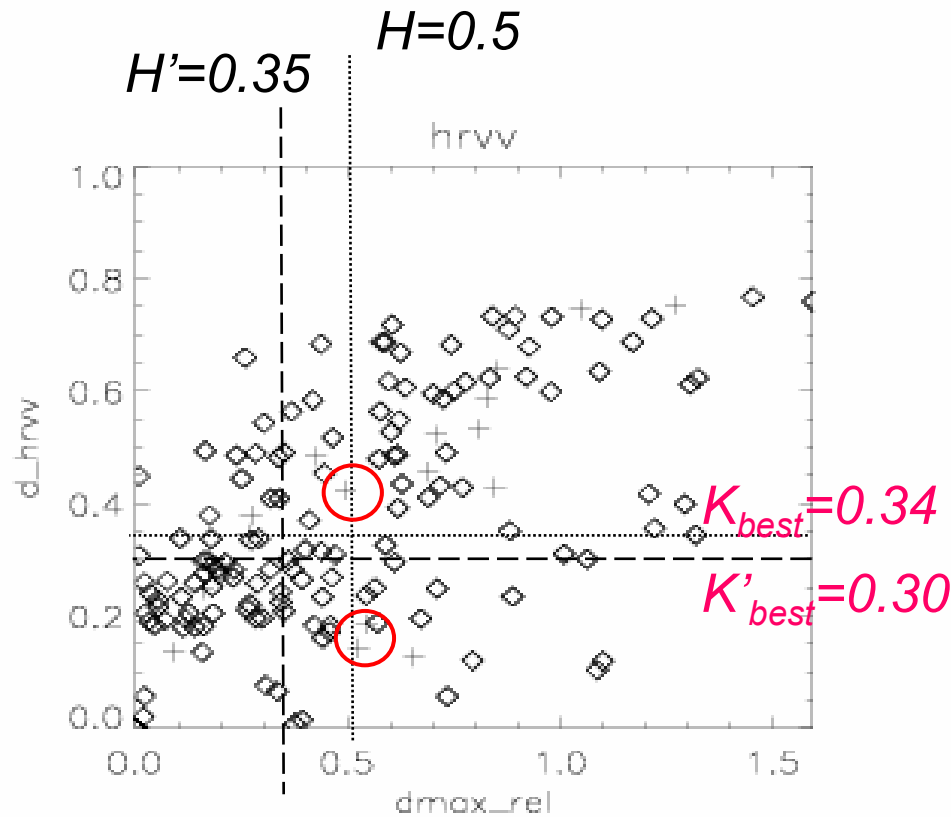
$$s(K) = \frac{\sum_{ij} m_{ij}}{N_c}$$

Percentage of model's couples in the sectors



Then I looked for the value of K that maximize $s(K)$

$$K_{best} \Rightarrow \max[s(K_{best})]$$



$H = 0.5$
 $K_{best} = 0.34$
 $s(K_{best}) = 0.77$
 $s(K_{best}, Observ) = 0.59$

$H' = 0.35$
 $K'_{best} = 0.30$
 $s(K'_{best}) = 0.72$
 $s(K'_{best}, Observ) = 0.70$

One of the two is true in 77% of the cases

$$d_{hrvm}(M_i, M_j) \leq 0.34 \Rightarrow d_{max_rel}(M_i, M_j) \leq 0.5$$

$$d_{hrvm}(M_i, M_j) > 0.34 \Rightarrow d_{max_rel}(M_i, M_j) > 0.5$$

One of the two is true in 59% of the cases

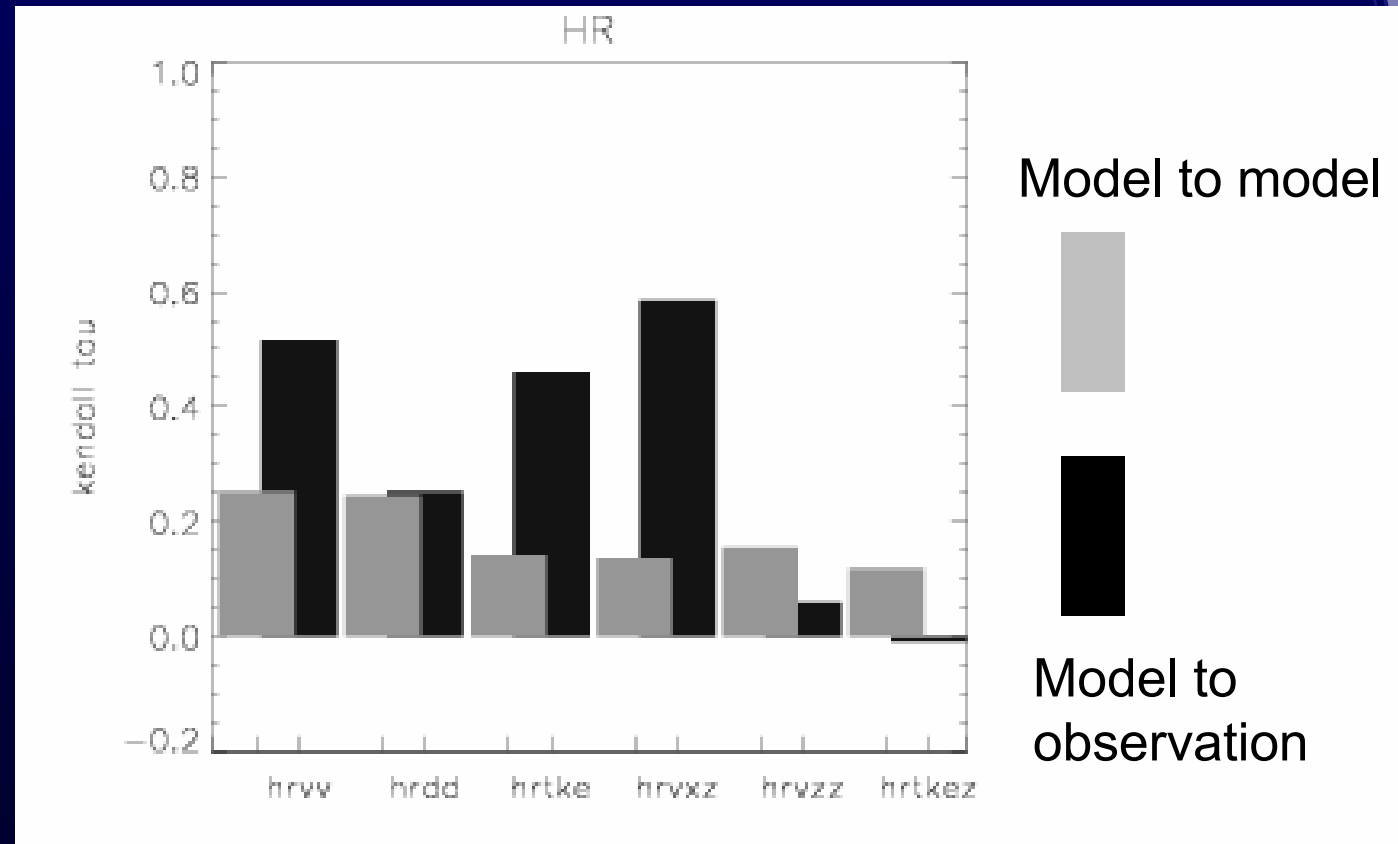
$$d_{hrvm}(M_i, O) \leq 0.3 \Rightarrow d_{max_rel}(M_i, O) \leq 0.35$$

$$d_{hrvm}(M_i, O) > 0.3 \Rightarrow d_{max_rel}(M_i, O) > 0.35$$

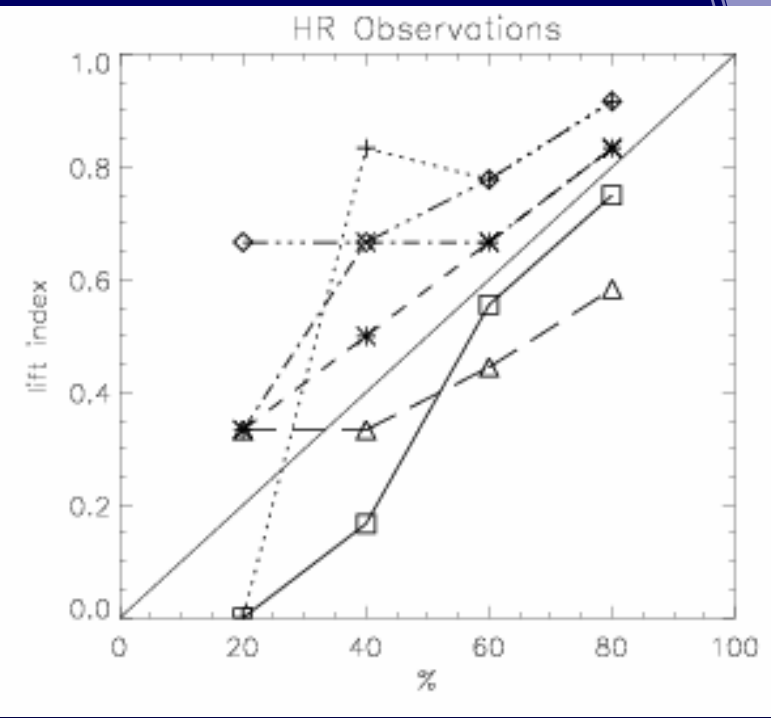
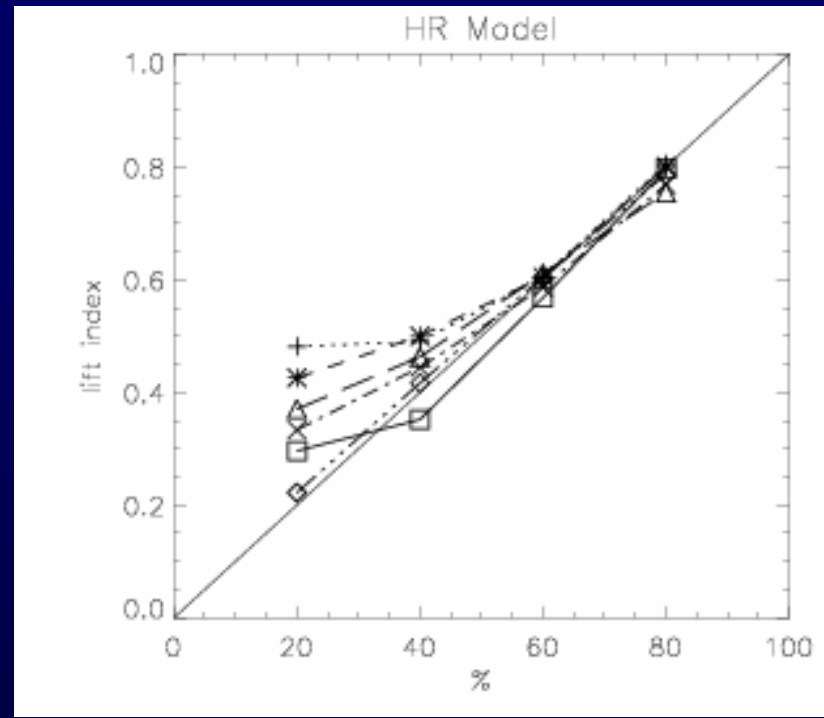
Test 2 (a bad one).
 Purpose: Hit Rate.
 Accuracy allowed: 66%

$$d_{purpose}(M, R) = 1. - HitRate(M, R), H = 0.34$$

Same strategy as before.
 Kendall tau

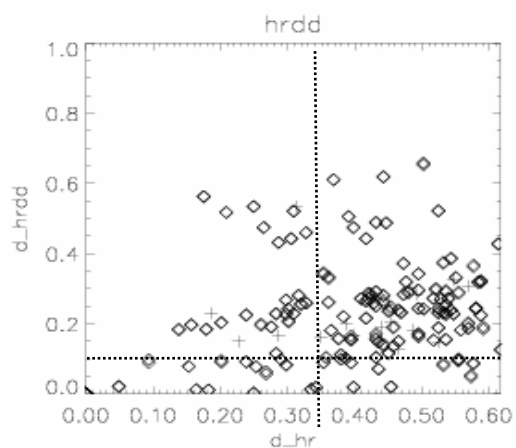
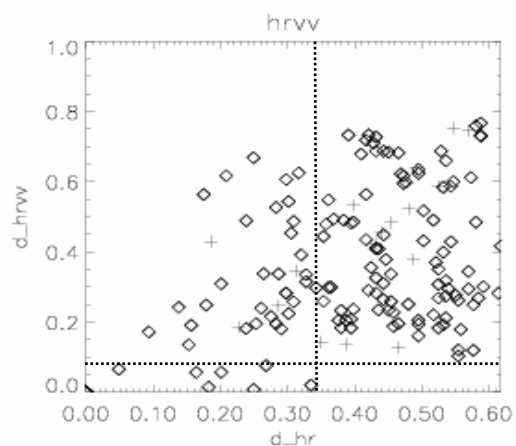


Lift Curve



- + +hrvv
- * - - - *hrdd
- x - - - xhrtke
- ◇ - - - ◇hrvxz
- △ - - - △hrvzz
- - - - □hrtkez

For the hit rate of concentration it is difficult to find the best surrogate metrics. Also looking for K_{best} seems to be more difficult.



Possible reasons:

- The metrics analysed are not appropriate. New metrics should be derived.
- From the behavior of the models at the measurements points it is not possible to decide if the model is fit for purpose or not. New measurements are needed.

Summary

- Is the model I'm using "appropriate" to simulate the case in study (a priori)?
 - Formulate and apply a model evaluation guidance and protocol for cloud models
- Model evaluation: model-measurements or model-model; is there any "choice"?
 - Use metrics and apply fit-for-purpose validation criteria
- Is a "good" comparison between model and measurements for a certain case a guarantee that the model will perform well in a "similar" case?
 - Probably yes if you follow the first two steps

Final comments – Towards a cloud modeling etiquette???

- Instead of claiming that the model results compare well with measurements report how well they compare.
- Do quantitative model intercomparison.
 - Adopt common used metrics or formulate new ones and define acceptance criteria.