CCA MODELING & MONITORING

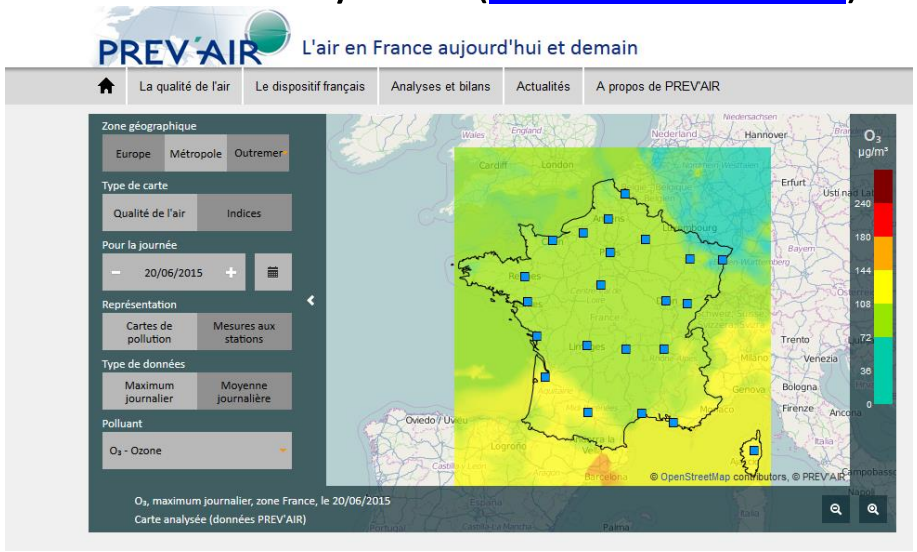
Evaluation of the re-analysis validation methodology for France

Laure Malherbe, Charline Pennequin, INERIS
laure.malherbe@ineris.fr

FAIRMODE technical meeting, 24-25 June 2015, Aveiro, Portugal

Context: analysed maps

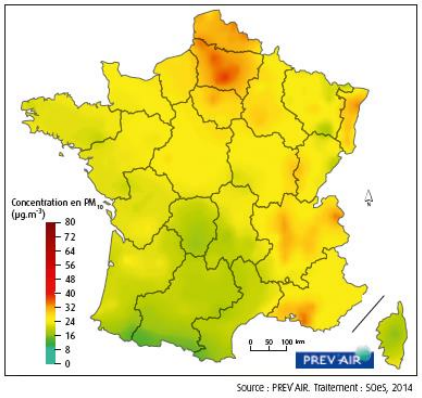
- Maps combining modelling and monitoring data are produced every day by the PREV'AIR system (www.prevair.org).



Daily maximum one-hour concentration of ozone 20/06/2015

CHIMERE + up-to-date (NRT) monitoring data

- They are also produced in retrospect for the annual national AQ assessment report and other research projects.



Source : PREV'AIR. Traitement : SOeS, 2014

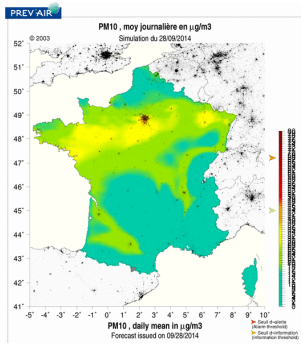
Average of hourly concentrations of PM_{10} over winter 2013

CHIMERE + validated monitoring data

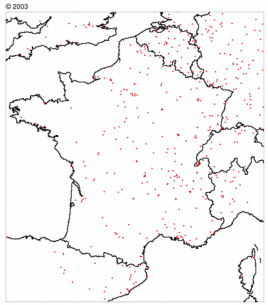
Context: analysed maps

-Meteorology
- Emissions
- Landuse
- Boundary conditions

CHIMERE



D-1, 27 September 2014, daily mean
0.1° x 0.15°



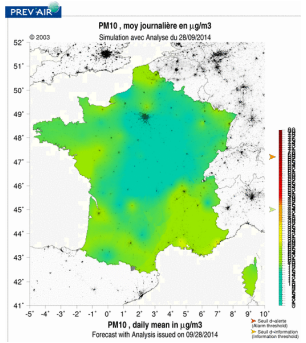
Monitoring data (France + Europe)

ANALYSIS

Combination of background observations with CHIMERE data

Geostatistical approach: **external drift kriging**

Example, PM_{10}
Map produced on the 28th of September for the 27th



Analysed map

D-1, 27 September 2014, daily mean

The kriging is done for each hour (input data: hourly values) or each day (input data: average daily values). It is implemented with R: *RGeostats* (Renard, 2010) and *gstat* (Pebesma, 2004) packages.

Current validation method

- The quality of the maps is currently evaluated by **cross-validation.**
  - **Leave-one-out cross-validation:** one station is removed from the input data set and the concentration is estimated at this point using the remaining stations.
  - **N-fold cross-validation**: The set of stations is split into N (e.g. 5) subsets. One subset of stations is removed and the concentrations at those stations are estimated using the N-1 remaining subsets.

  The cross-validation function (*gstat.cv*) included in the *gstat* package is used. It makes use of the variogram model fitted with the complete set of stations.

- Cross-validation is performed for each hour (each day)**.** Annual statistical scores are then computed.

New validation approach

- In the current cross-validation procedures, both leave-one-out and n-fold, a station is removed from the input dataset only once → the final result is **one estimated value per station**, to be compared to the actual measurement.

- In the proposed **methodology based on Monte-Carlo**, a subset of stations (e.g. 20%) is randomly removed, concentrations at those stations are estimated by kriging and this procedure is repeated a large number of times (n).

  A station can be selected for validation several times and should be selected at least once ($1 \leq k \leq n$). The final result is ***k* estimated values per station**, to be compared to the actual measurement.

Test of the Monte-Carlo approach

- The methodology has been tested for:
  - the French domain,
  - **PM_{10}** ,
  - the whole year **2012**, on an **hourly basis**.
- Input data:
  - hourly time series of PM_{10} concentrations measured at rural and suburban or urban background stations in France and surrounding countries (source: French national AQ database and Airbase v8)
  - hourly time series simulated by CHIMERE CTM with a spatial resolution of approximately 4km
- Monte-Carlo parameters:
  - **20% of stations removed for validation** at each random selection (function *sample* of R)
  - Number *n* of random selections: ***n* = 200**, ***n* = 300**, and ***n* = 500**

Test of the Monte-Carlo approach

Questions :

- Should the *n* samples be selected once for all the year or should they be selected independently every hour?

  FAIRMODE procedure: both options seem possible. → Second option retained in these tests (easier to implement in our calculation chain).

> In particular, this procedure is based on the performances of a number of re-analyses performed randomly by selecting 20% of the stations to be used for validation. The subsets of stations can be selected "offline" or during the procedure, with the only constrain that each station has to be selected at least one time for the validation. The validation is then performed considering, for each station, the worst case, i.e. the re-analysis with the highest RMSE.

- A constraint is that each station should be selected at least once -> this implies that the *n* selections are redone until this condition is fulfilled. This automatic check has not been introduced yet. Is there an easy solution to ensure this condition?

Test of the Monte-Carlo approach

Questions:

- For a given hour, should a unique variogram be fitted with the complete set of stations and used in the *n* kriging calculations or should the variogram be recalculated for each of the *n* selections using the partial set (80%) of stations ? → second option chosen, considered as more penalizing.

- For a given station *i*, the result is a time series made of multiple estimated values for each hour:

  - $k_{i,1}$ $(1 \leq k_{i,1} \leq n)$
  - $k_{i,2}$ $(1 \leq k_{i,2} \leq n)$
  - $k_{i,3}$ $(1 \leq k_{i,3} \leq n)$
  - …
  - $k_{i,8784}$ $(1 \leq k_{i,8784} \leq n)$

Which values should be retained for comparison to the observations?

FAIRMODE procedure : select the worst case. Other cases will also be considered in this exercise for comparison purpose.

Test of the Monte-Carlo approach

Questions:

➢ How is this worst case identified?

1. A set of **n Monte Carlo** re-analyses has to be performed.

For each re-analysis:

a) Randomly select 20% of the stations to be used as **validation stations** so **do not use them to perform the re-analysis**
b) Compute for each station **i** (at least) in each re-analysis **j** the **RMSE(i,j)**

2. Compute for each station **i** the maximum of **RMSE(i,j)**. Let be **vect_max(i)** the number of the re-analysis associated to the maximum RMSE for station **i**.
3. Create a CDF file to be used in the Delta Tool by selecting for each station **i** the **vect_max(i)** instances, i.e. the time-series of the station **i** computed during the re-analysis **vect_max(i)**. In this way, for each station, the worst case is selected.
4. Use the delta tool as if the CDF file was the CDF file of a single model.

Does reanalysis *j* correspond:

- to one estimation for a given time (hour in these tests)? In that case *RMSE(i,j)* is just the square error (*SE(i,j)*).
- or to a full time series ? In that case *RMSE(i,j)* is computed over the whole year. Only possible if the *n* samples are exactly the same for each hour ($k_{i,1} = k_{i,2} = k_{i,3} = \ldots = k_{i,8784} = k_i$), thus allowing the constitution of k_i full time series.

→ First option is considered since the requirements for the second option are not met in these tests.

Test of the Monte-Carlo approach

- Output data:
  - For each station and each test (*n*=200, *n*=300, *n*=500), 7 time series of hourly values

Date	Obs	CTM	CV_LOO	CV_Nfold	MC_P50	MC_P90	MC_max
2012010101	15	7.6	20.0	24.0	20.0	27.1	33.1
2012010101	12	7.9	16.0	23.2	18.8	20.8	22.5
…	…	…	…	…	…	…	…
2013010100	…	…	…	…	…	…	…

Obs	**Measured value**	
CTM	CHIMERE (interpolation at the station)	
CV_LOO	Leave-one-out cross-validation	
CV_Nfold	5-fold cross-validation	
MC_P50	Monte-Carlo validation, estimated value corresponding to the median square error	added for comparison
MC_P90	Monte-Carlo validation, estimated value corresponding to the 90th percentile of the square error	added for comparison
MC_max	Monte-Carlo validation, estimated value with maximum square error (worst case)	

Processing of the evaluation results

- Only French stations with annual data coverage ≥ 85% have been kept for calculating scores (213 stations).

- Calculation of usual scores:

  For each station and each type of estimation:
  - RMSE: Root Mean Square Error
  - R: Correlation coefficient
  - NMB: Normalized Mean Bias
  - NMSD: Normaized Mean Standard Deviation
  - Taylor Diagram

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(O_i - M_i)^2}$$

$$R = \frac{\sum_{i=1}^{N}(M_i-\bar{M})(O_i-\bar{O})}{\sqrt{\sum_{i=1}^{N}(M_i-\bar{M})^2}\sqrt{\sum_{i=1}^{N}(O_i-\bar{O})^2}}$$

$$NMB = \frac{BIAS}{\bar{O}} \text{ where } BIAS = \bar{M} - \bar{O}$$

$$NMSD = \frac{(\sigma_M-\sigma_O)}{\sigma_O}$$

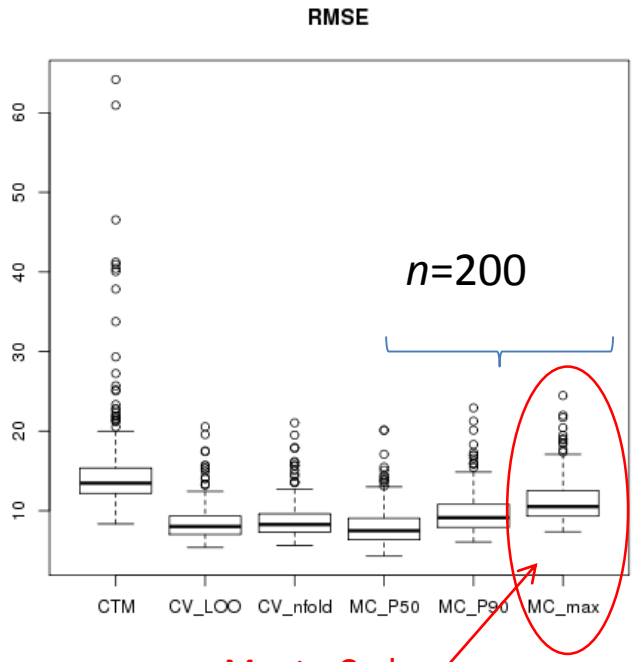
Guidance Document on Model Quality Objectives and Benchmarking, Viaene et al., 2015)

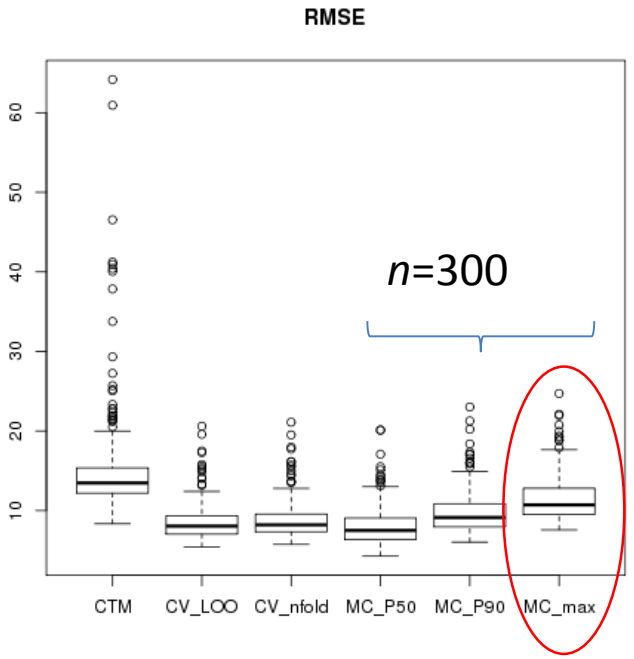
- Use of the Delta Tool

  (online updated version)

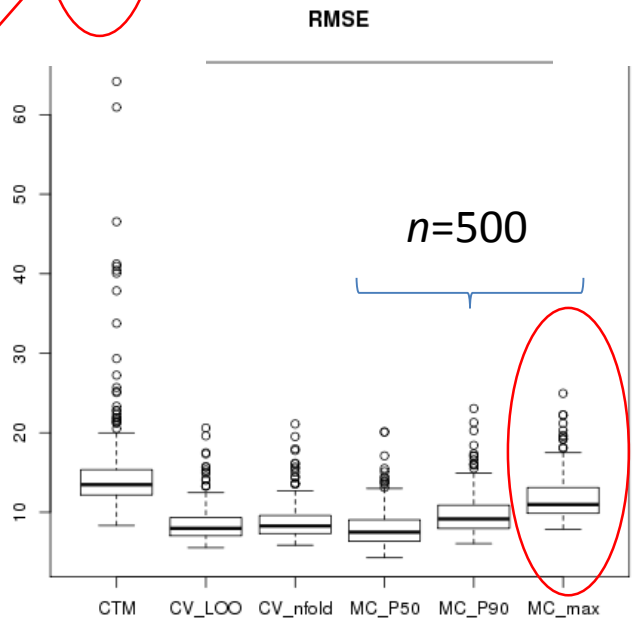
Processing of the evaluation results

Boxplots of the RMSE calculated for each type or evaluation and the 213 French stations





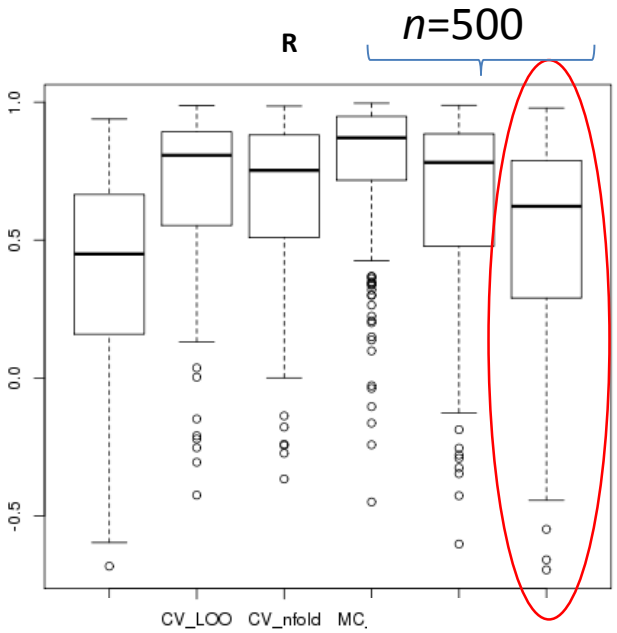
Monte-Carlo, worst case



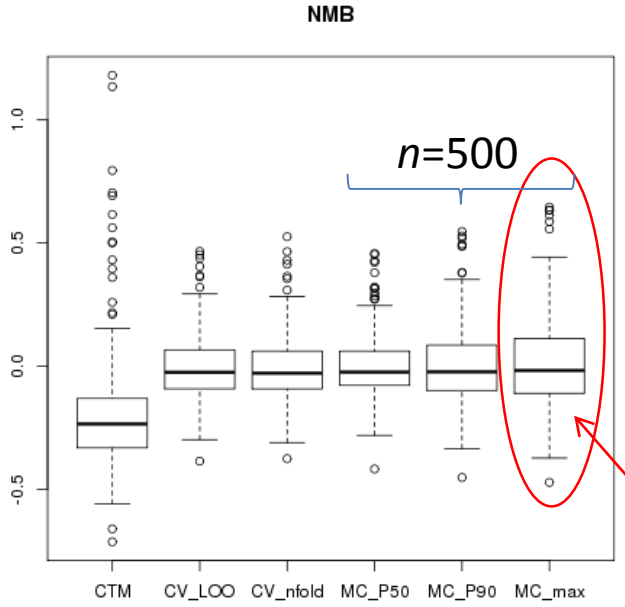
- No significant difference according to the number of subset selections.
Same observation for the other scores

Processing of the evaluation results

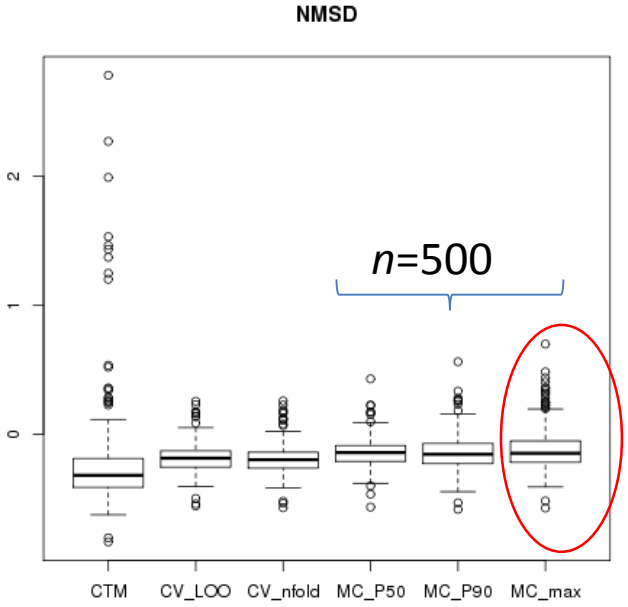
Boxplots of the correlation, the NMB and the NMSD calculated for each type or evaluation and the 213 French stations



- Best scores for the Monte-Carlo estimates corresponding to the median error
- Worst scores (RMSE, R, NMB) for the Monte-Carlo estimates corresponding to the maximum error

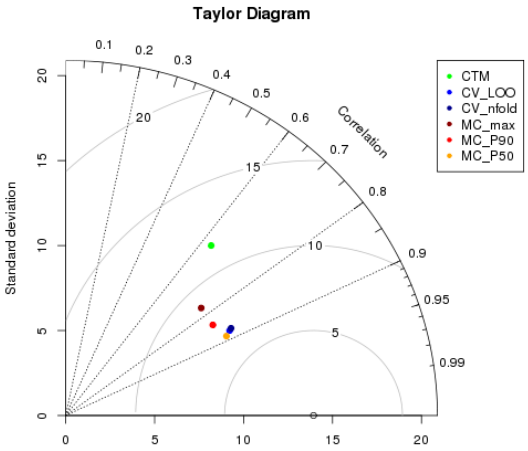


Monte-Carlo, worst case

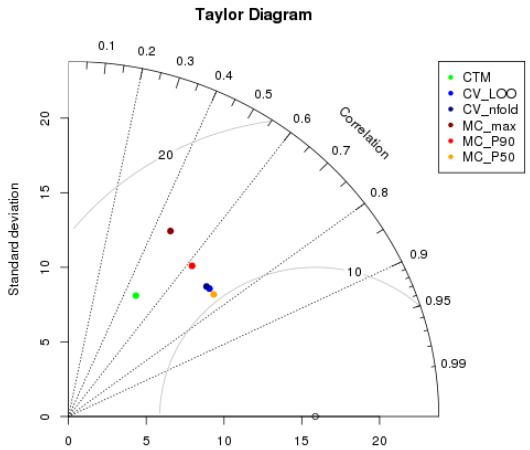


Processing of the evaluation results

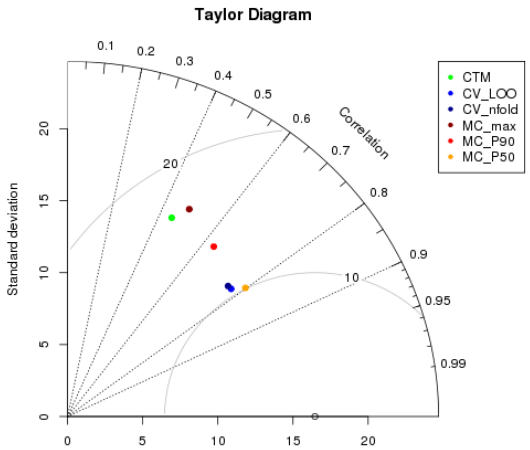
FR01001



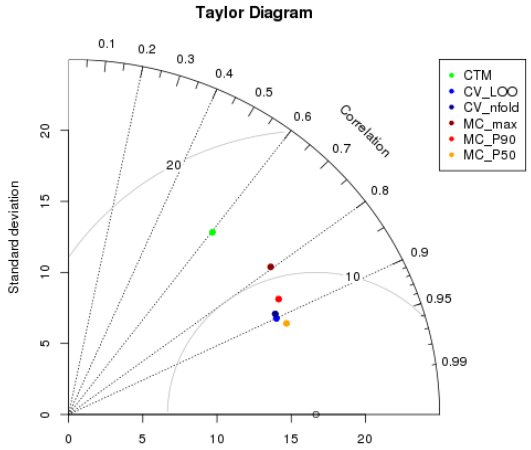
FR02005



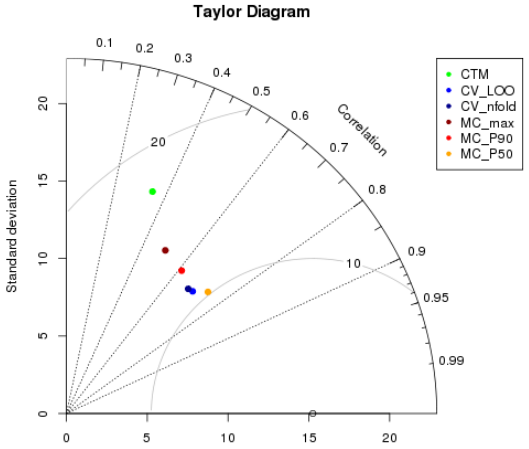
FR03043



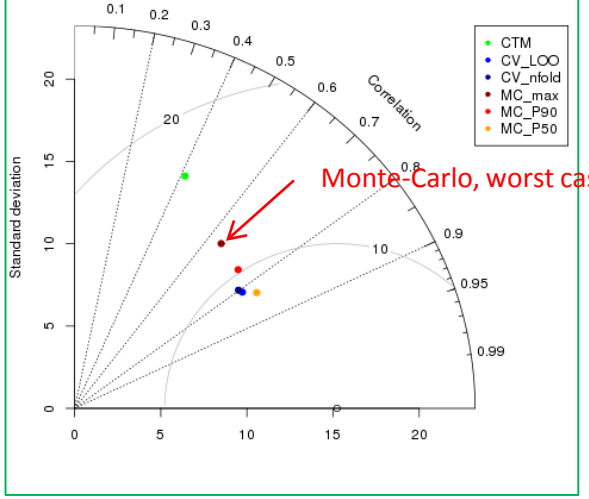
FR11027



FR31001



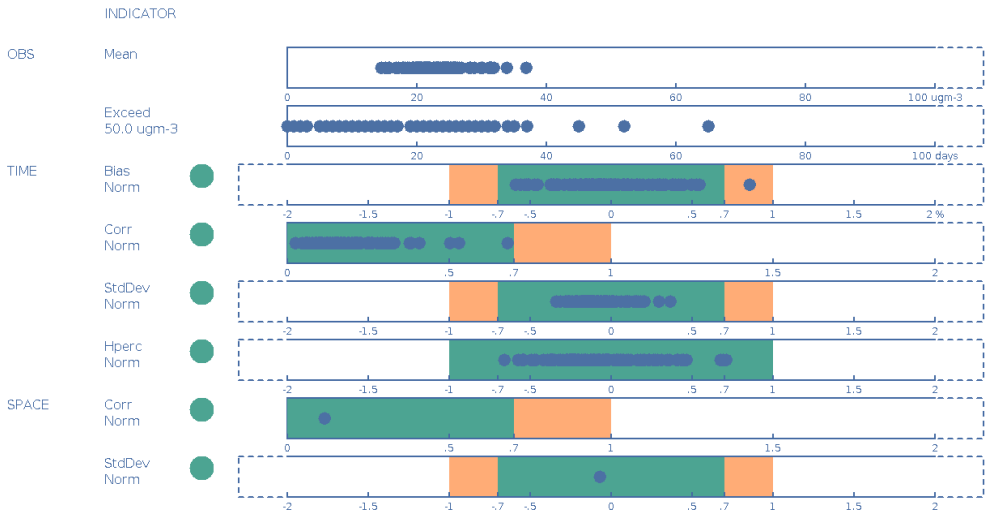
All stations together



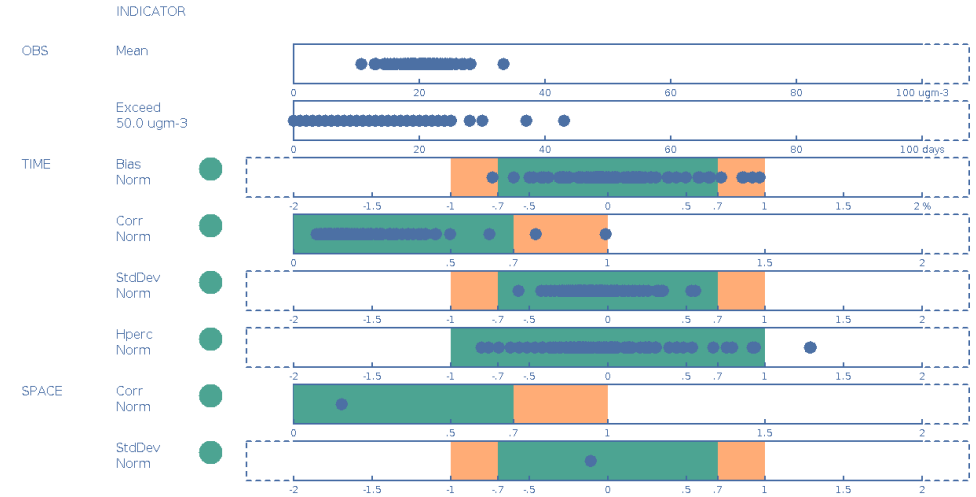
n=500

Processing of the evaluation results

1st half of the stations

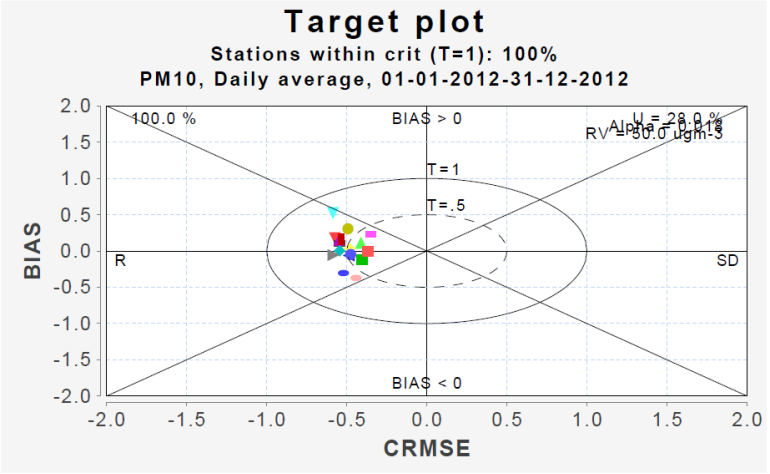


2nd half of the stations

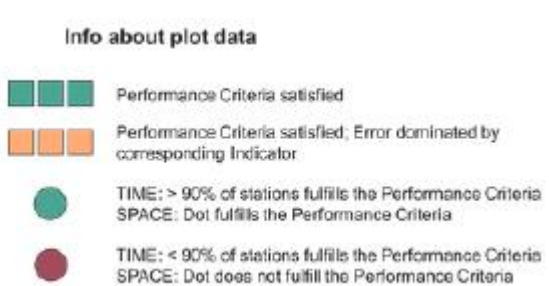


n=500

Delta tool output in the worst case



Target plot for stations located in the South-West of France



Preliminary conclusions

- About the approach:
  - **From a methodological point of view:** more detailed specifications could be helpful but no special difficulty was encountered.

    After the procedure is extensively tested by FAIRMODE community, some aspects of the approach could be detailed or revised:

    - Could an interval of values be recommended for the number *n* of simulations?
    - Does it make a difference if the *n* selected subsets are different for each time step or are the same for the whole year ?
    - In the present tests, performance criteria were satisfied. However, could the « worst case » be too penalizing? Consider a high percentile of the error instead of the maximum?
  - **From a technical point of view:** the implementation requires attention but does not pose any particular problem.

    Calculations were performed on the CCRT* Airain supercomputer (*Research and Technology Computing Centre of the CEA). About 4 to 10 hours needed for one year depending on *n*.

Preliminary conclusions

- Next steps:
  - The influence of the different parameters of the methodology will be further investigated.
  - The analysis of the results with the Delta tool will be continued.
  - The added value of the Monte-Carlo approach in relation to the usual leave-one-out or n-fold cross-validation will be further examined.
  - Tests will be performed for other years and pollutants.