ML4AQ meeting – First results with KFAN and ML

PETETIN Hervé

10/02/2019

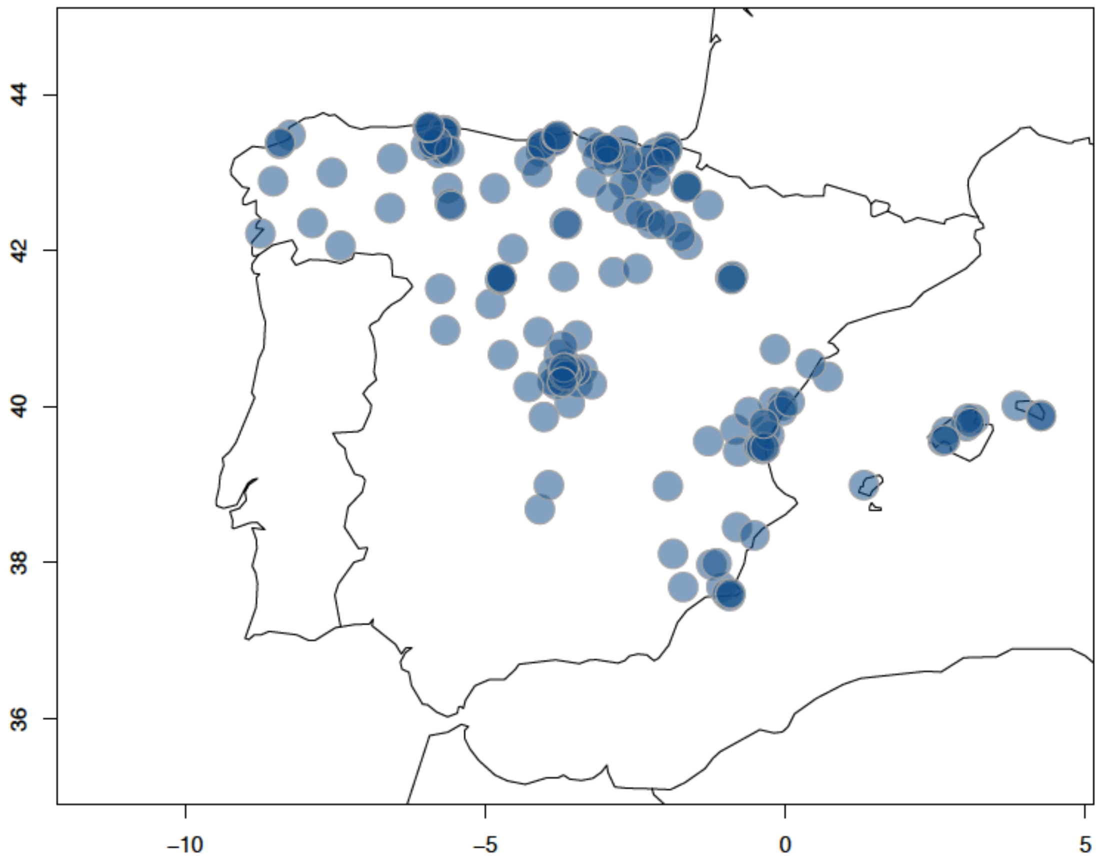
What has been done?

- Modification of the MOS scripts to work in CAMS50-like operational conditions (i.e. daily 4-days forecasts)
- Test of several MOS approaches :
  - MA<N> : moving average on N previous days
  - KF<s/d> : optimcal <static/dynamic> Kalman filter
  - AN<X> : analogs with configuration X
  - KFAN<X> : analogs with configuration X in Kalman space
  - ML<X> : machine learning with algorithm X
- Entire IP domain (only stations with >75% data retained)
- Many changes of my MOS script (e.g. operational-like mode, various MN4 issues to handle, additional flexibility for parallelisation)
- Experiments :
  - MONARCH b007 (2015, without the 2 bugged months)
  - NB : Problem with CAMS50 : no meteorological variables!
- Most of my R scripts are translated to python (results shown today are with the R version)

MONARCH alone (IP domain)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

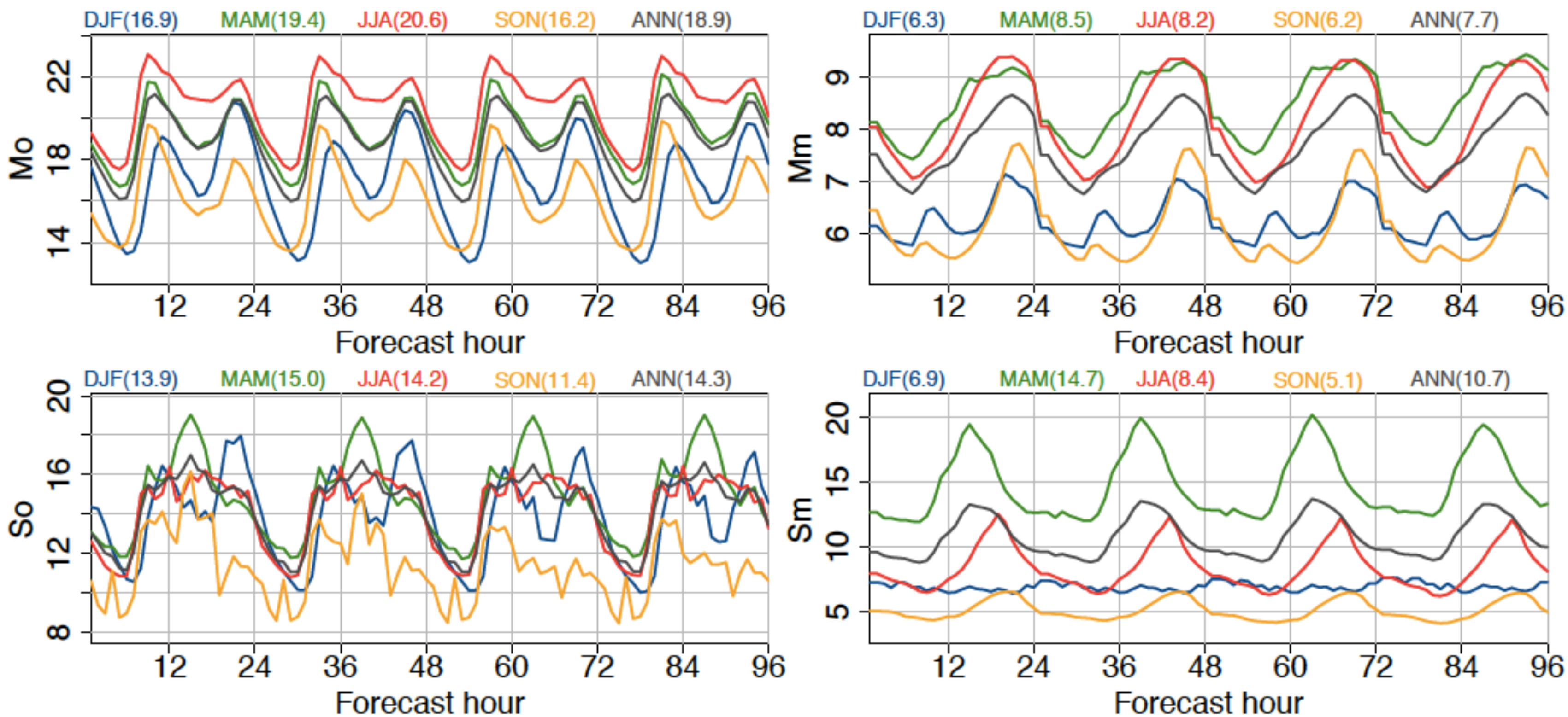
Statistical performance of MONARCH (PM10/IP)

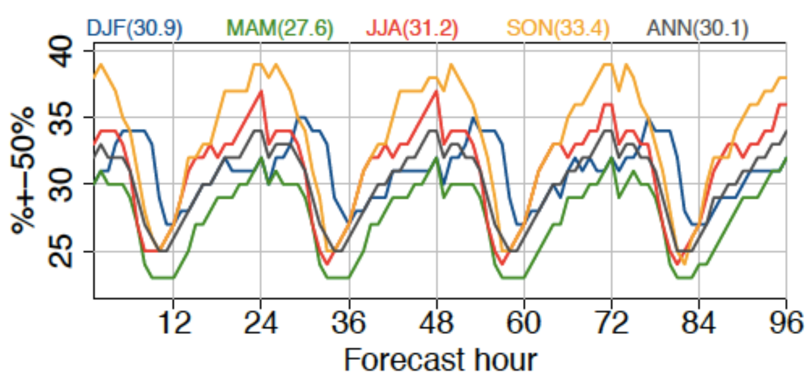


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

➔ 166 stations

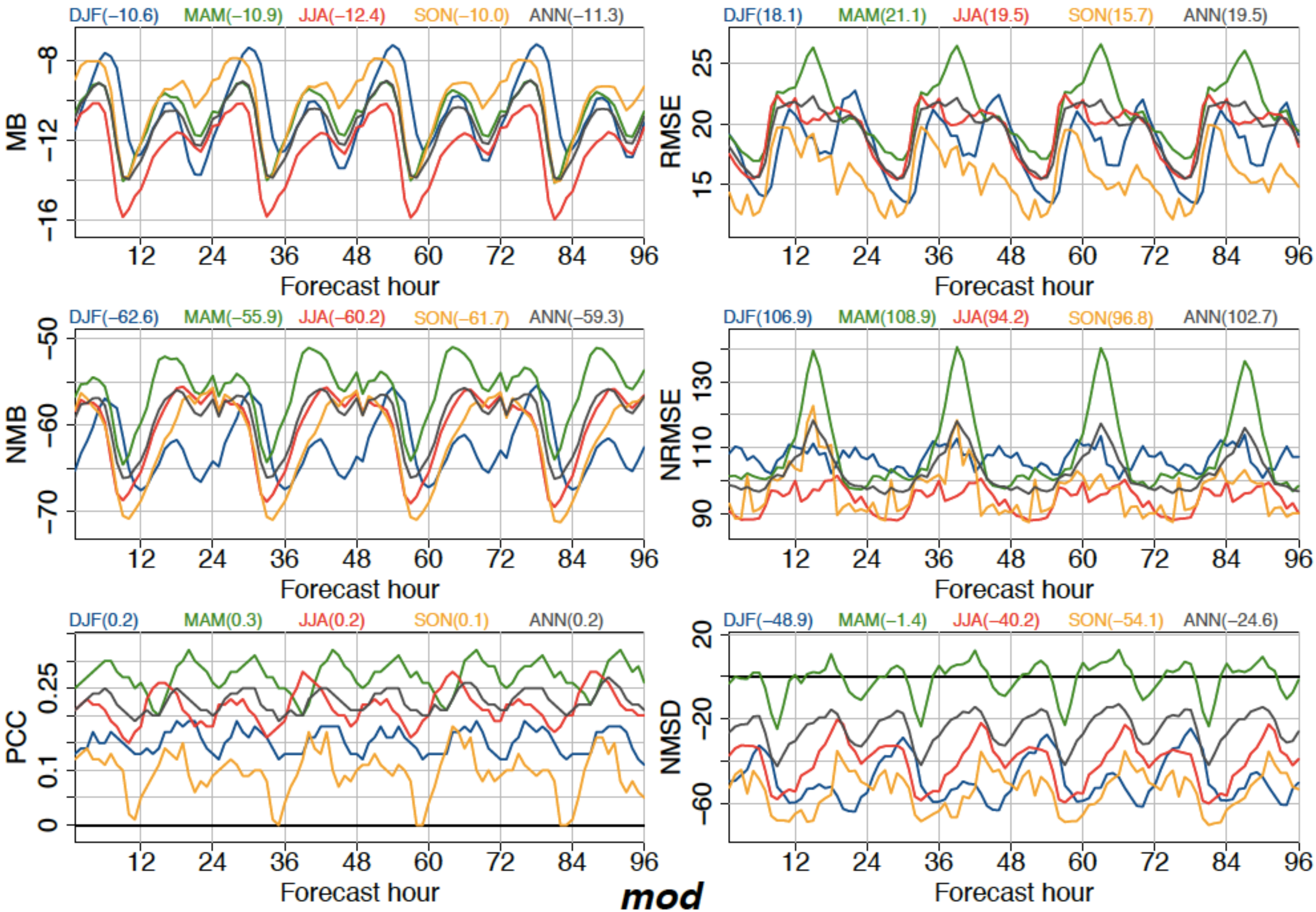
Statistical performance of MONARCH (PM10/IP)



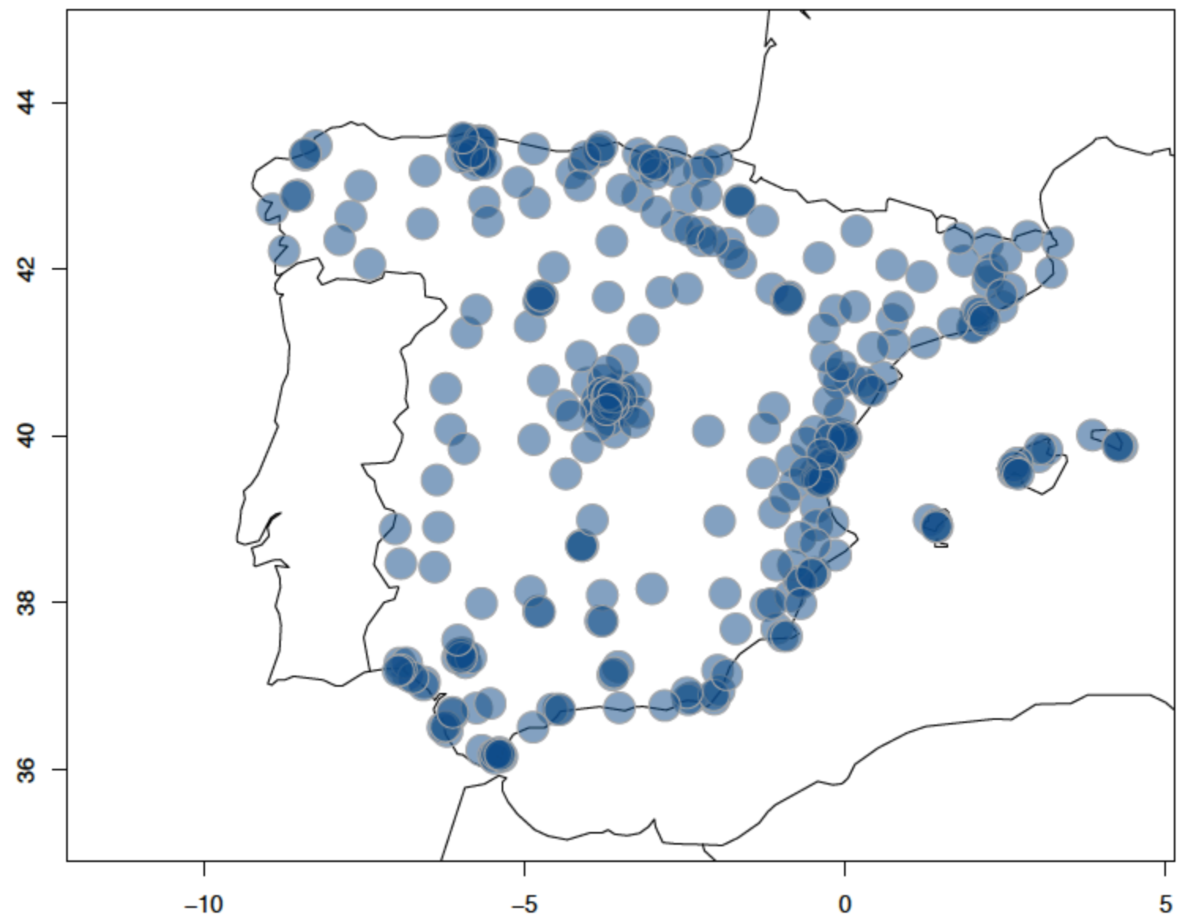


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Statistical performance of MONARCH (PM10/IP)

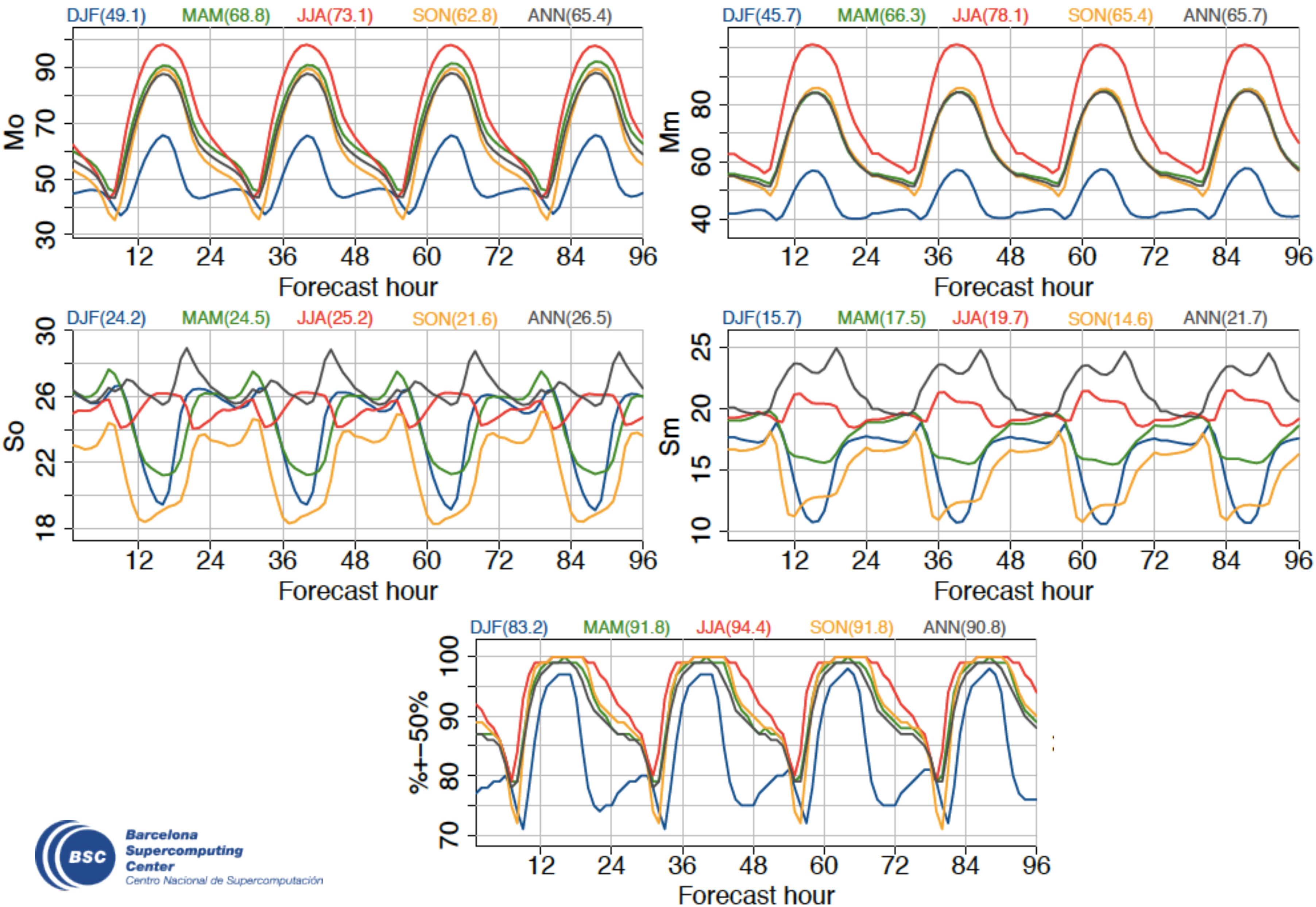


Statistical performance of MONARCH (O3/IP)



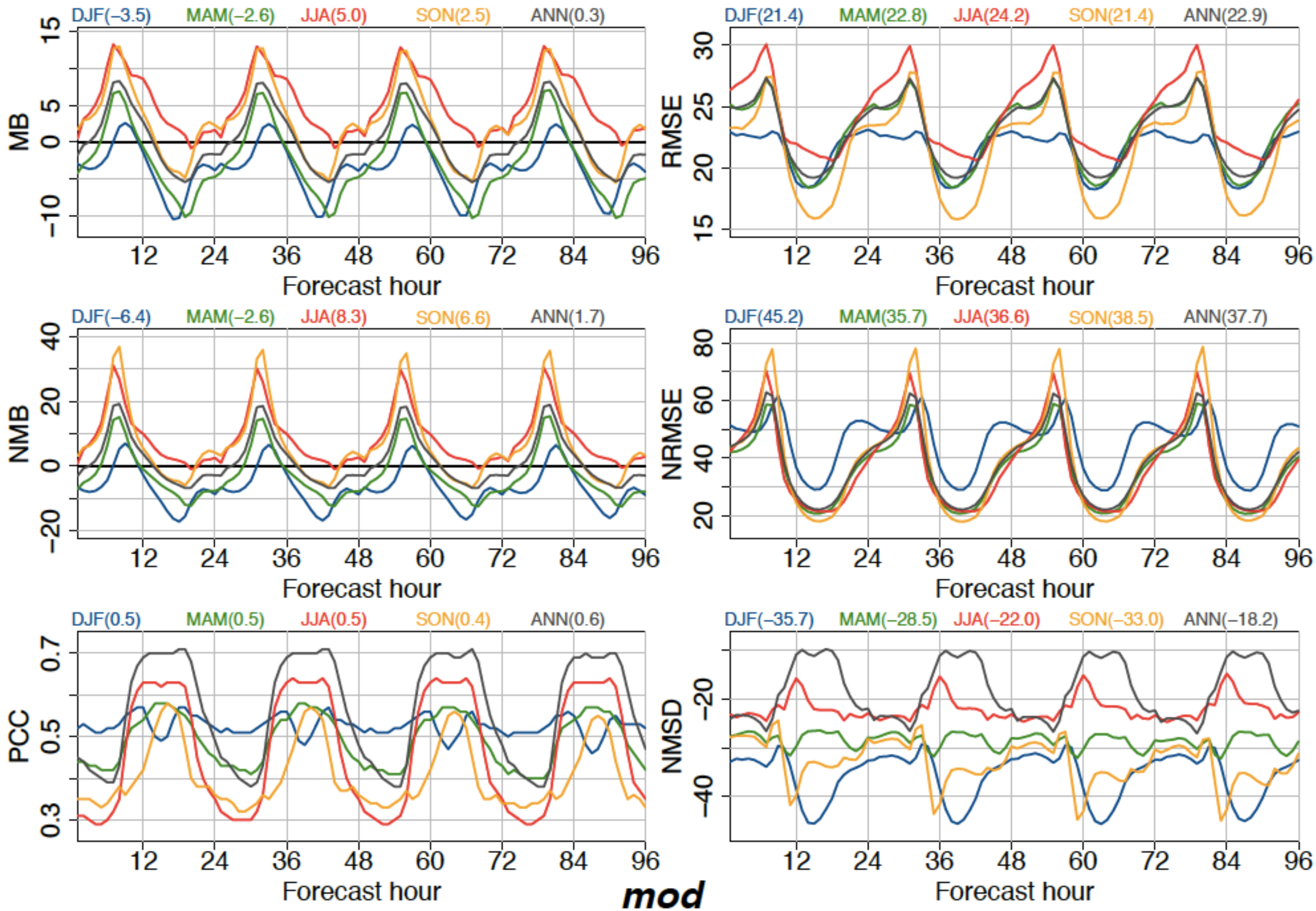
➔ 312 stations

Statistical performance of MONARCH (O3/IP)

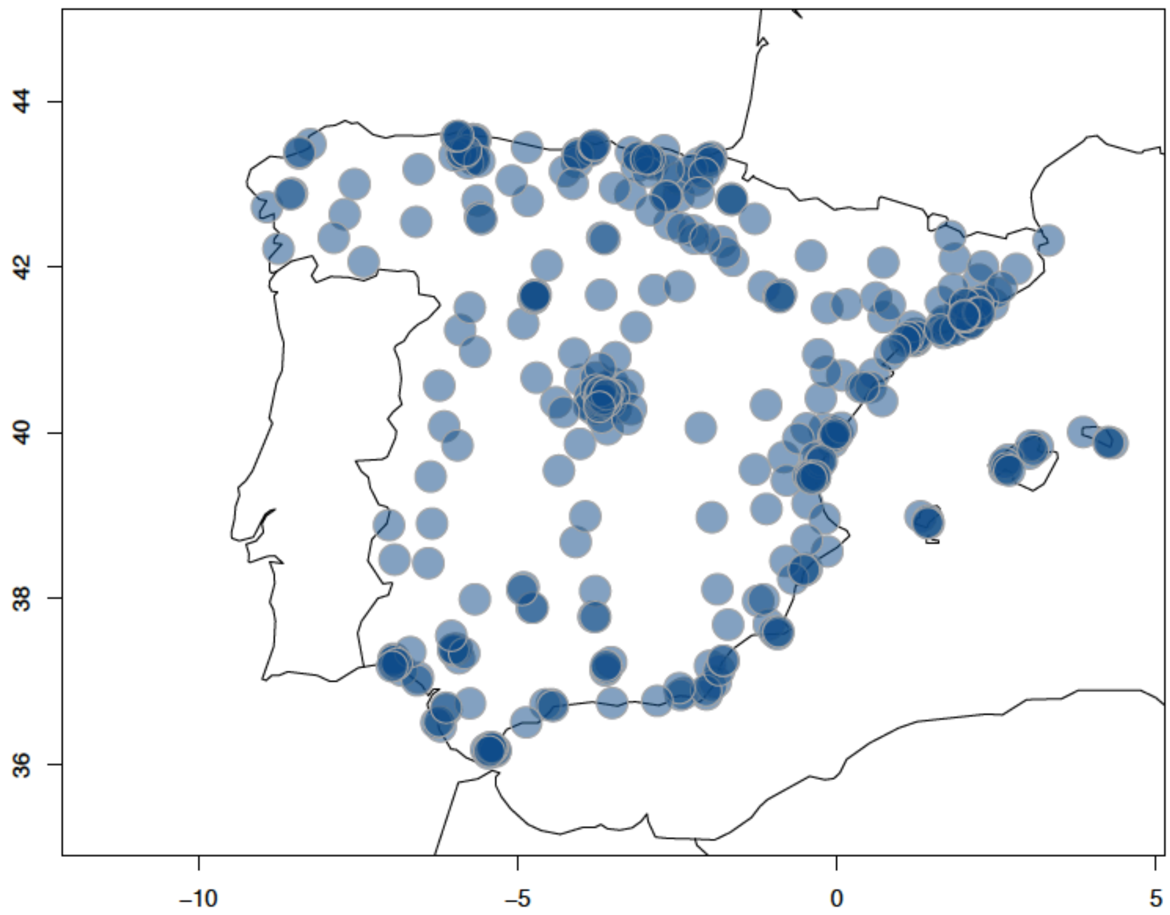


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Statistical performance of MONARCH (O3/IP)



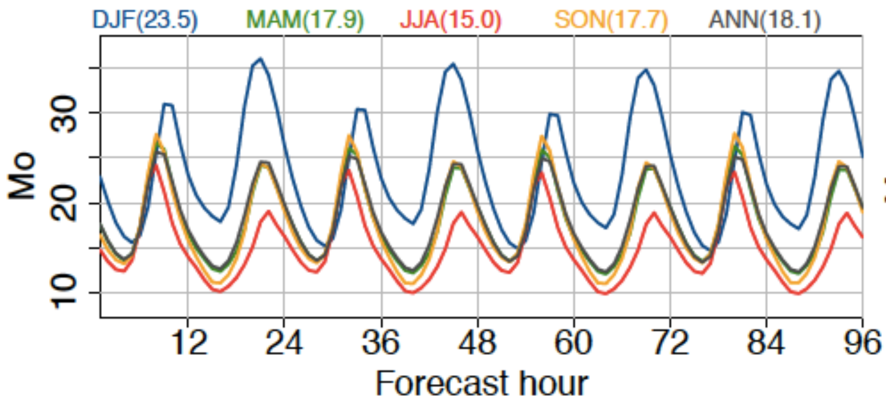
Statistical performance of MONARCH (NO2/IP)

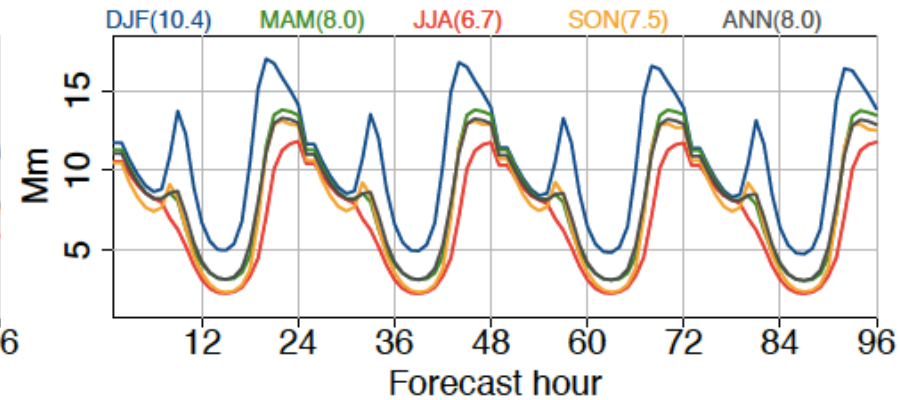


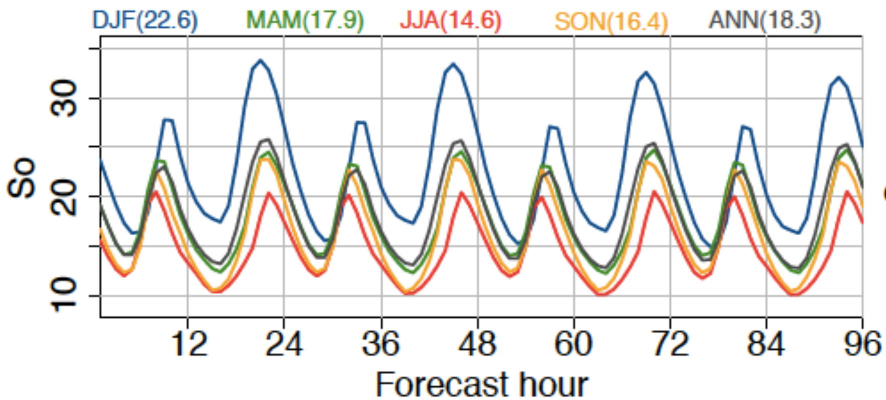
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

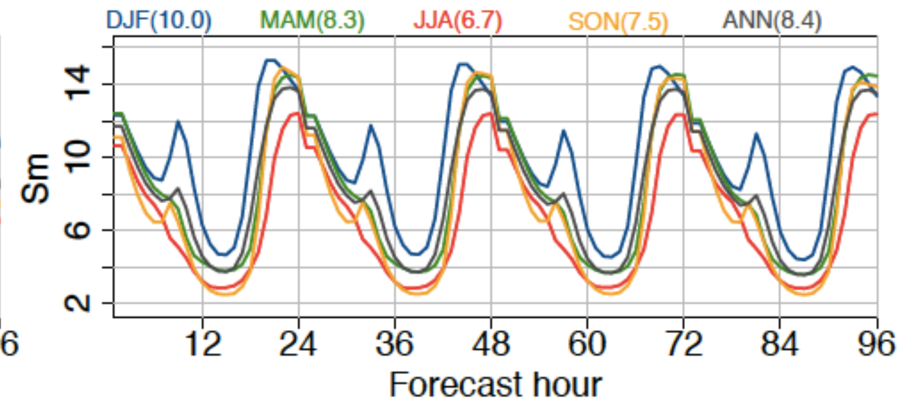
➔ stations

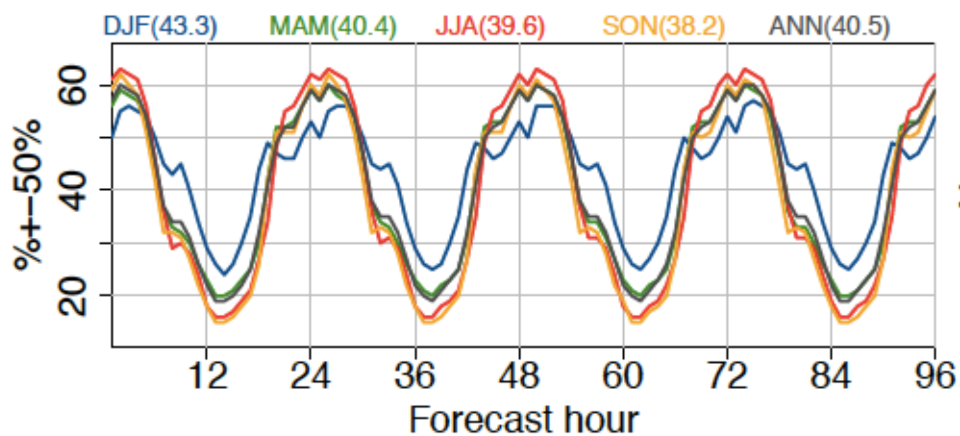
Statistical performance of MONARCH (NO2/IP)





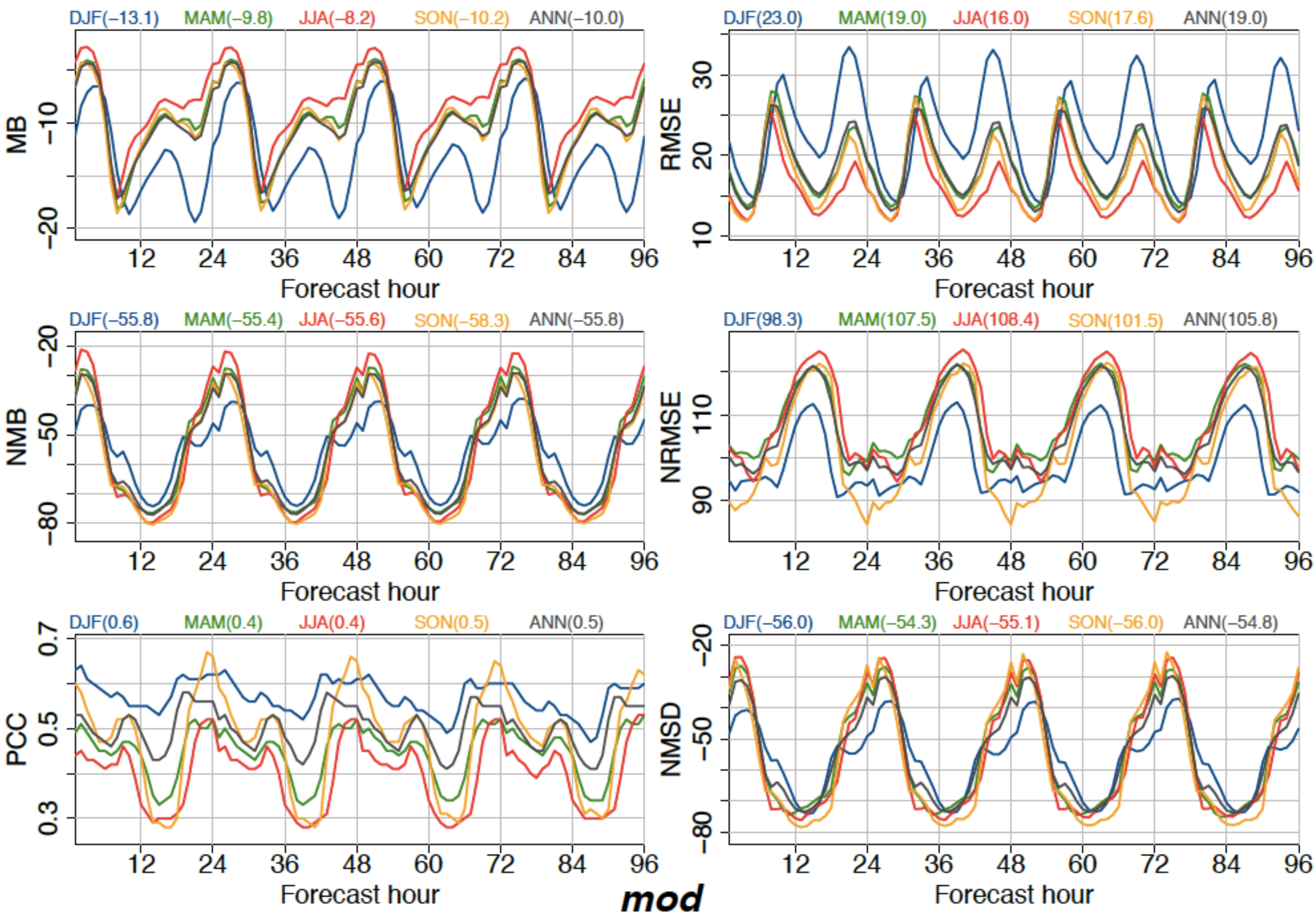






Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Statistical performance of MONARCH (NO2/IP)



Statistical performance of MONARCH - Overview

Annual statistics (seasonal range when substantial) on IP domain :

	NMB (%)	NRMSE (%)	PCC	NMSD (%)	Ndaily
PM10	-60	103	0.23 (0.10; 0.27)	-25 (-1; -55)	31,000
O3	2 (-3; 8)	38 (36; 45)	0.57 (0.41; 0.53)	-18 (-22; -36)	53,000
NO2	-56	106	0.50 (0.40; 0.57)	-55	65,000

In terms of statistics, no strong differences between station types except for NO2 :
RUR versus URB stations

➔ better NMB (-30% versus -60%)
➔ worst PCC (0.33 versus 0.49)
➔ better NMSD (-33% versus -54%)
➔ worst NRMSE (114% versus 98%)

But statistics can show different diurnal profiles between station types (e.g. O3)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

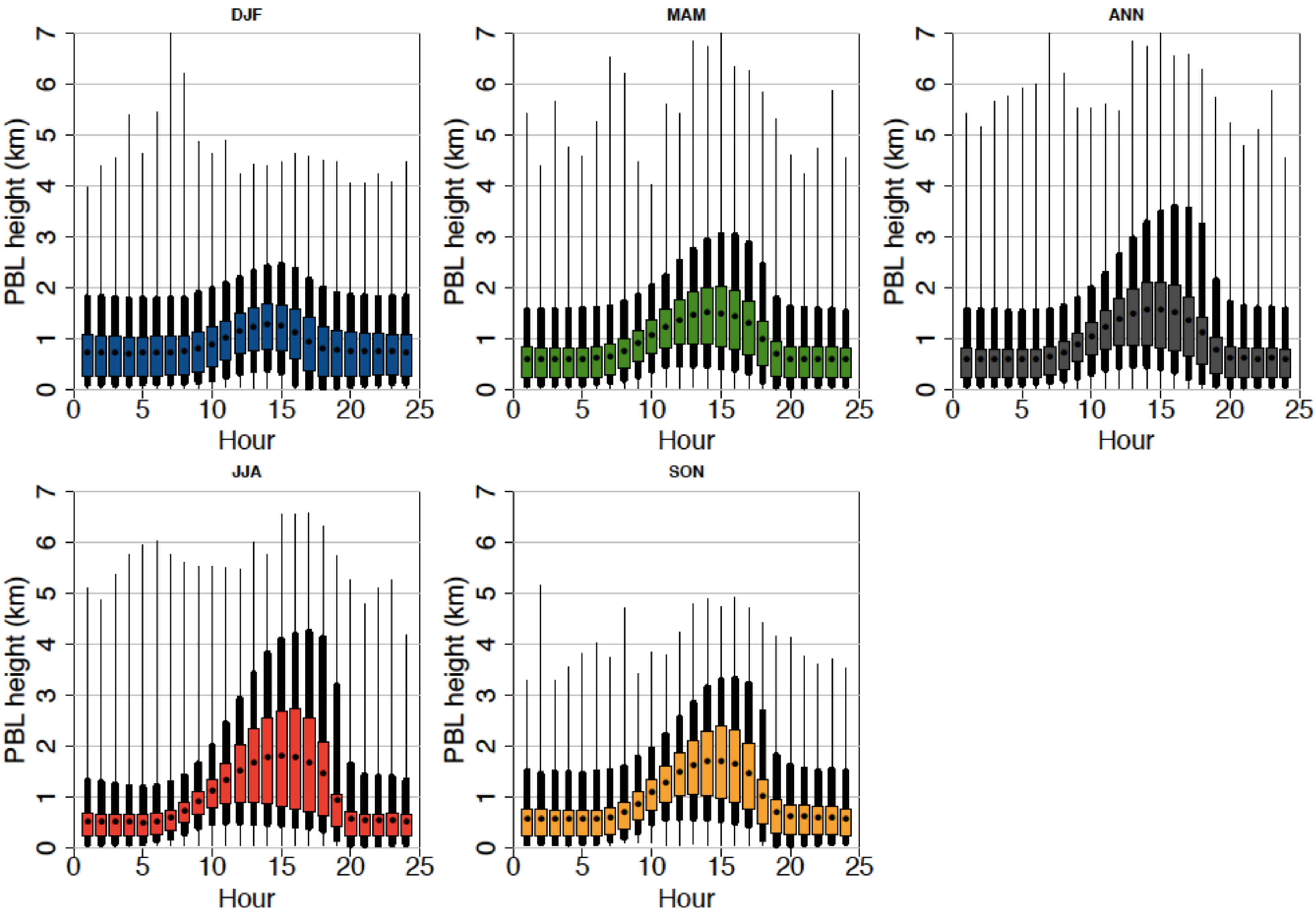
Statistical performance of MONARCH - Overview

- PM10 :
  - PM10 diurnal variability poorly represented
  - Overall strong negative bias (in particular during morning)
  - Strongest errors during morning transition
  - Lowest correlations in winter/fall
- O3 :
  - O3 diurnal variability reasonably well represented for all seasons
  - Overall strong positive bias
  - Strongest errors during morning transition whatever the season (strong positive bias) and late evening in winter (more random errors)
  - Highest (lowest) correlations during afternoon (early morning)
- NO2 :
  - NO2 diurnal variability quite well represented except morning peak (too low) and late evening (too persistent peak)
  - Overall strong negative bias (in particular during morning)
  - Strongest errors during early afternoon
  - Lowest correlations in summer/fall/spring
- All pollutants :
  - Underestimated variability

Statistical performance of MONARCH - Overview

- Strong underestimation of NO_2 during daytime : resolution? emissions? PBL height? vertical mixing? bug?
  - Inconsistent with Badia et al. (2017) : positive bias on rural EMEP stations (e.g. summer, nighttime), despite coarser resolution (1.4°x1°)
  - Check with CAMS50
- More specifically, important issue during morning rush hours : erroneous NO_x and PM emissions (wrong emissions and/or wrong temporal profile) and/or eventually too deep PBL and too coarse resolution
- This leads to strong negative bias on PM (that peaks during morning rush hours) and NO_2 (that peaks in early afternoon) and strong positive bias on O_3 (too low titration by NO?)

PBL height in MONARCH (averaged over IP stations)

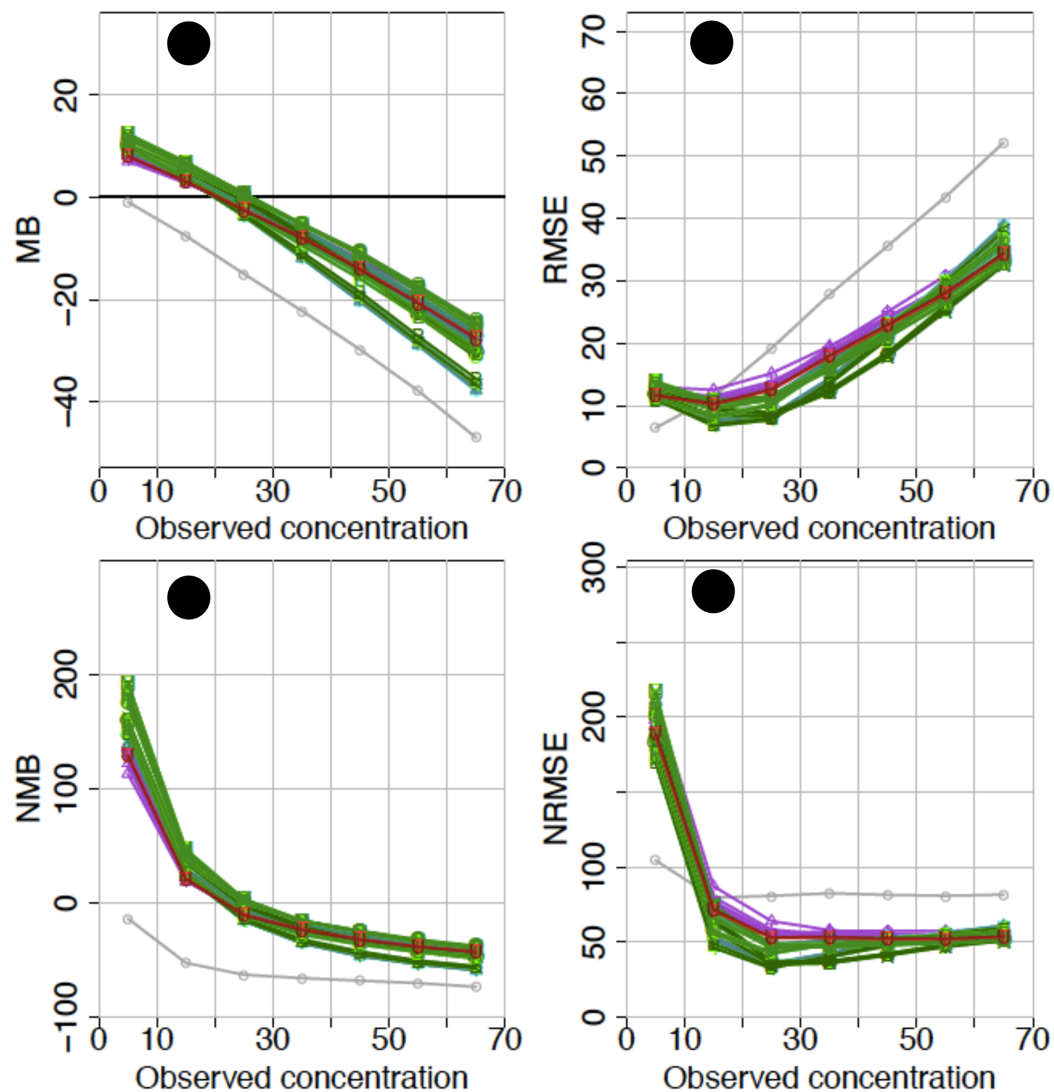


MOS correction

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Effect of MOS correction on MONARCH errors

- PM10 : negative bias and error increasing with observed concentration, but quite constant in relative
- MOS correction of PM10 ➔ increase the concentrations to correct the bias, leading to stronger errors in very low concentrations (positive bias) but lower errors elsewhere
- Very similar for NO2

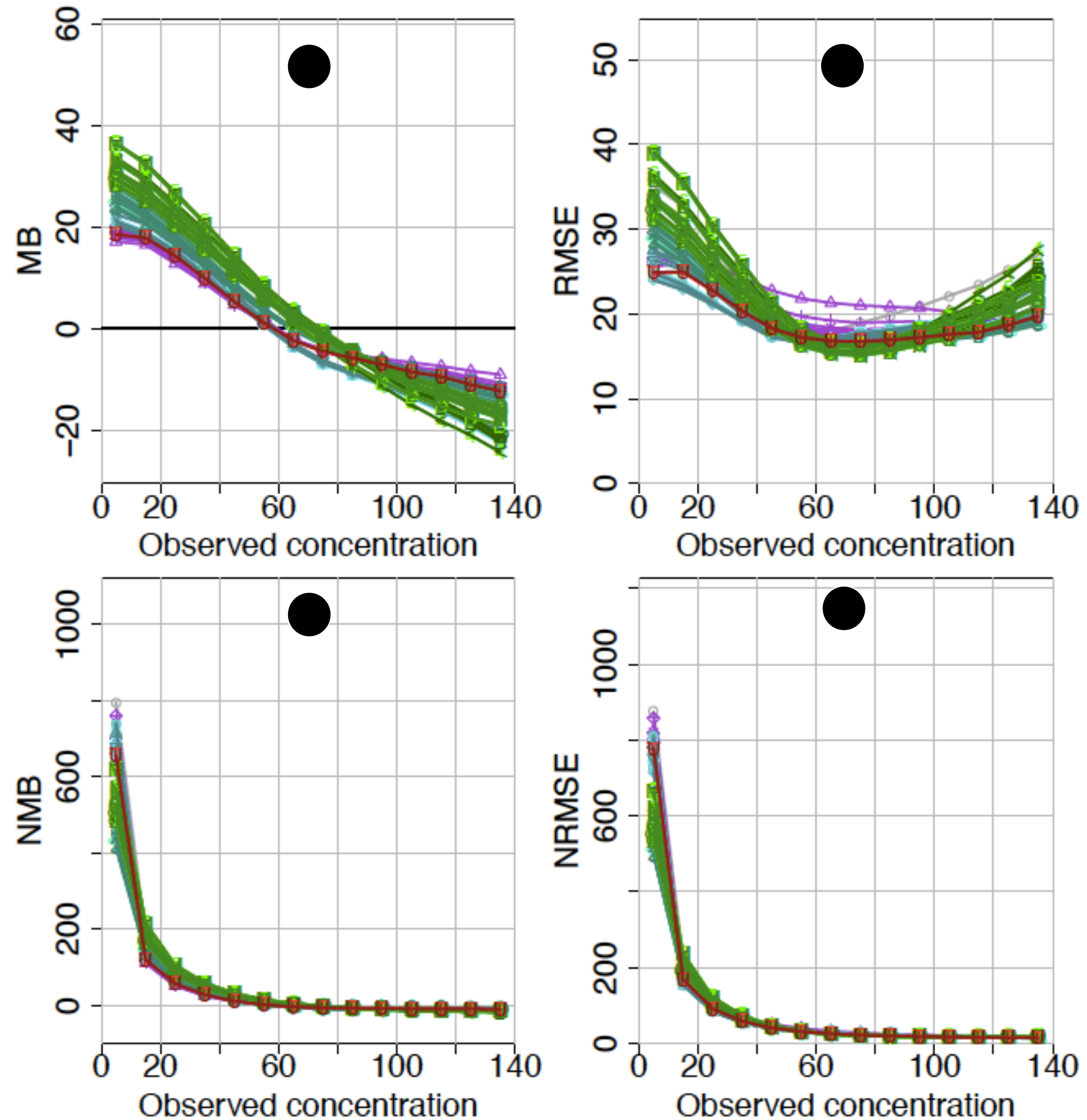


● *Most numerous data*

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

Effect of MOS correction on MONARCH errors

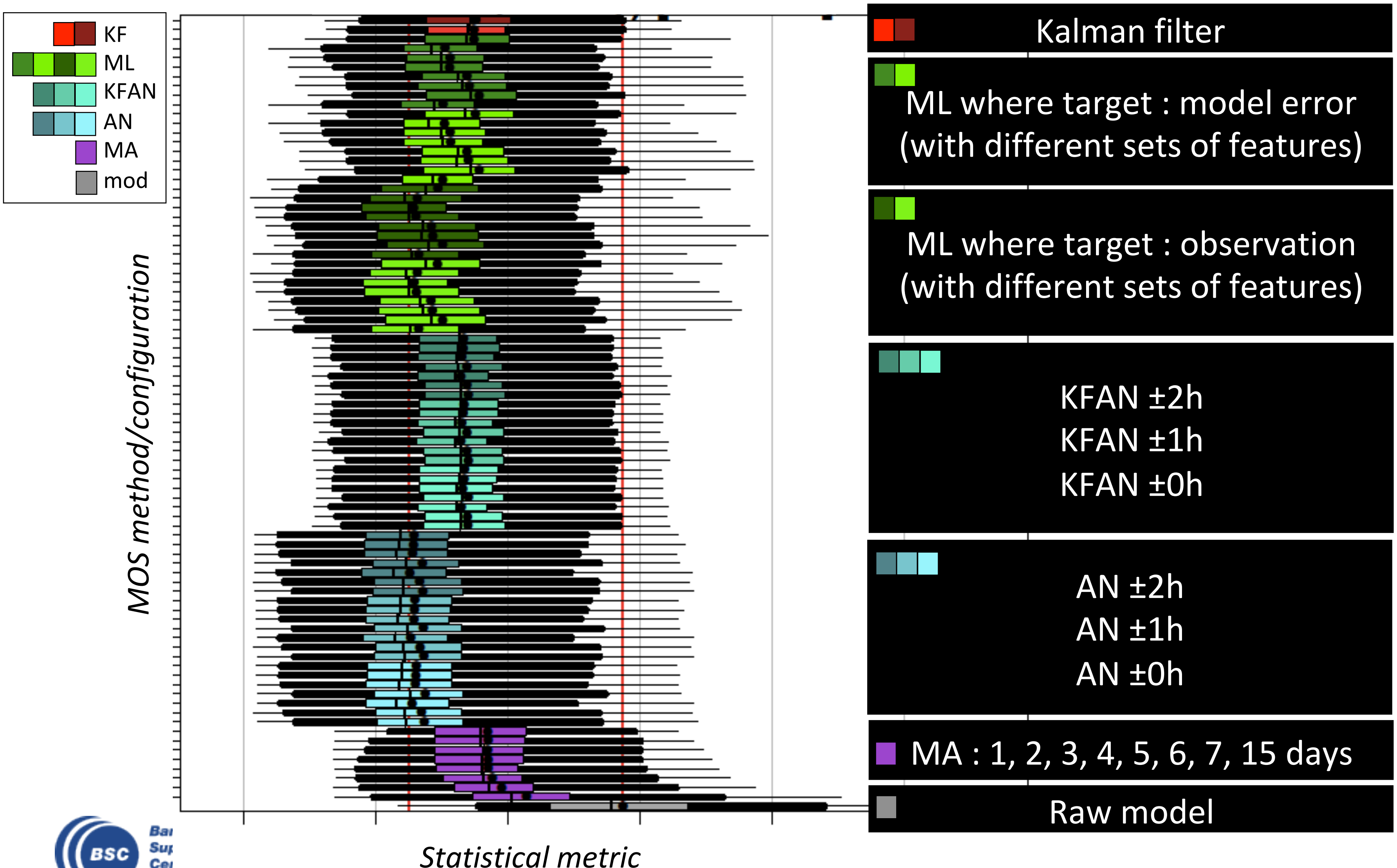
- O3: positive (negative) bias on lower (higher) concentrations
- MOS correction ➔ reduce both negative and positive bias, and usually reduce the error over the whole range of observed concentrations



● *Most numerous data*

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MOS methods and configurations



MOS-ML naming convention

Example for the GBM algorithm :

ml_gbm-<start>-<frequency>-<target>-<config_var>-<config_train>-<bagfraction>

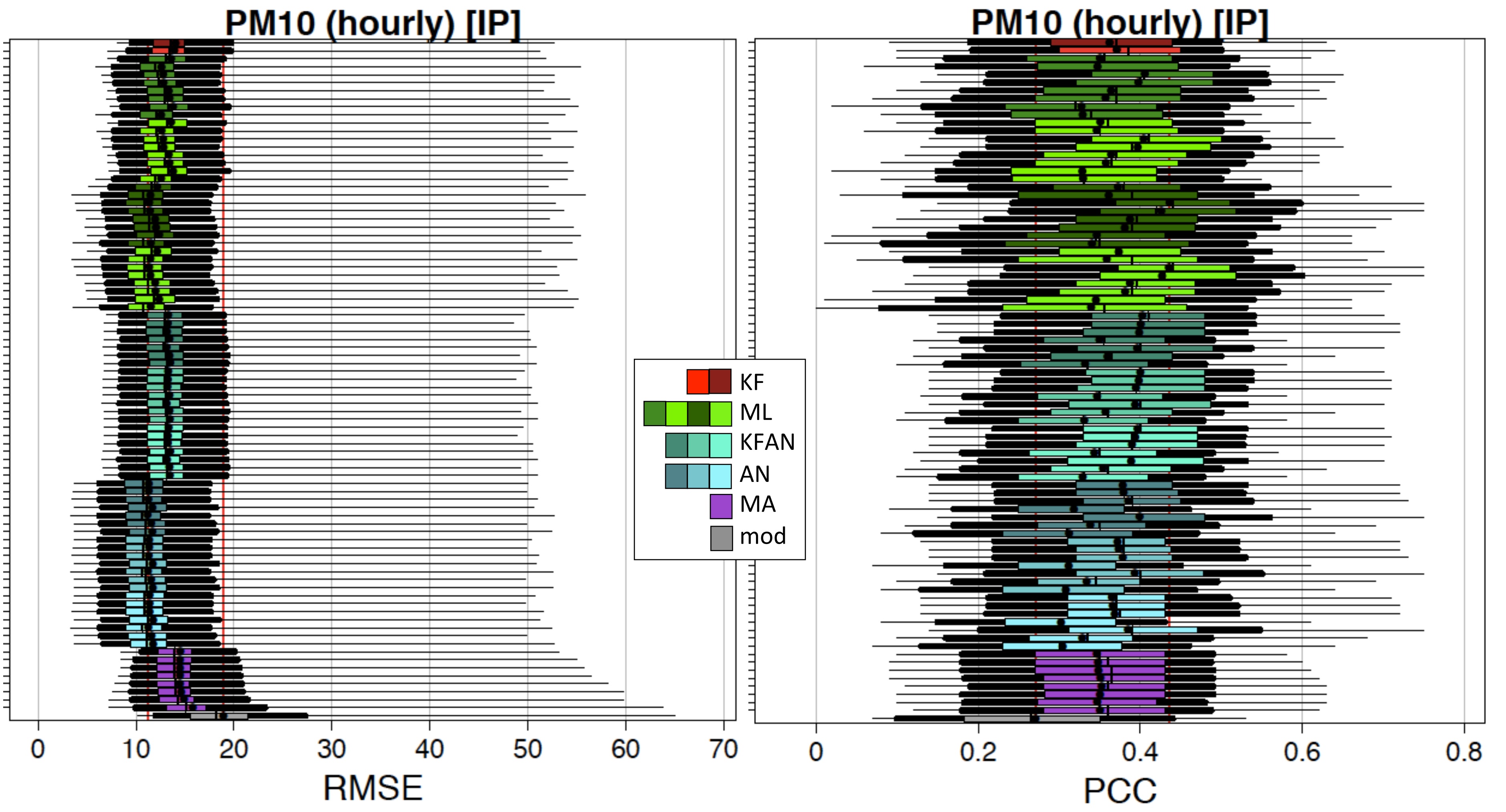
- Start : initial number of days before starting to train ML models
- Frequency : frequency (in days) at which the ML model is updated
- Target : 0 if the target is the observed concentration, 1 if the target is the error (mod-obs)
- Config_var : id of the set of features taken into account
- Config_train : id of the training configuration chosen
- Bagfraction : bag.fraction tuning parameter of the GBM algorithm

Example : ml_gbm-90-30-0-1-1-0.75

➔ training start after 90 days, is updated every 30 days, tries to predict the observed concentration, based on the set of features n°1 (modeled concentration + standard meteorological parameters) and the training configuration n°1 (default), with bag fraction of 75%

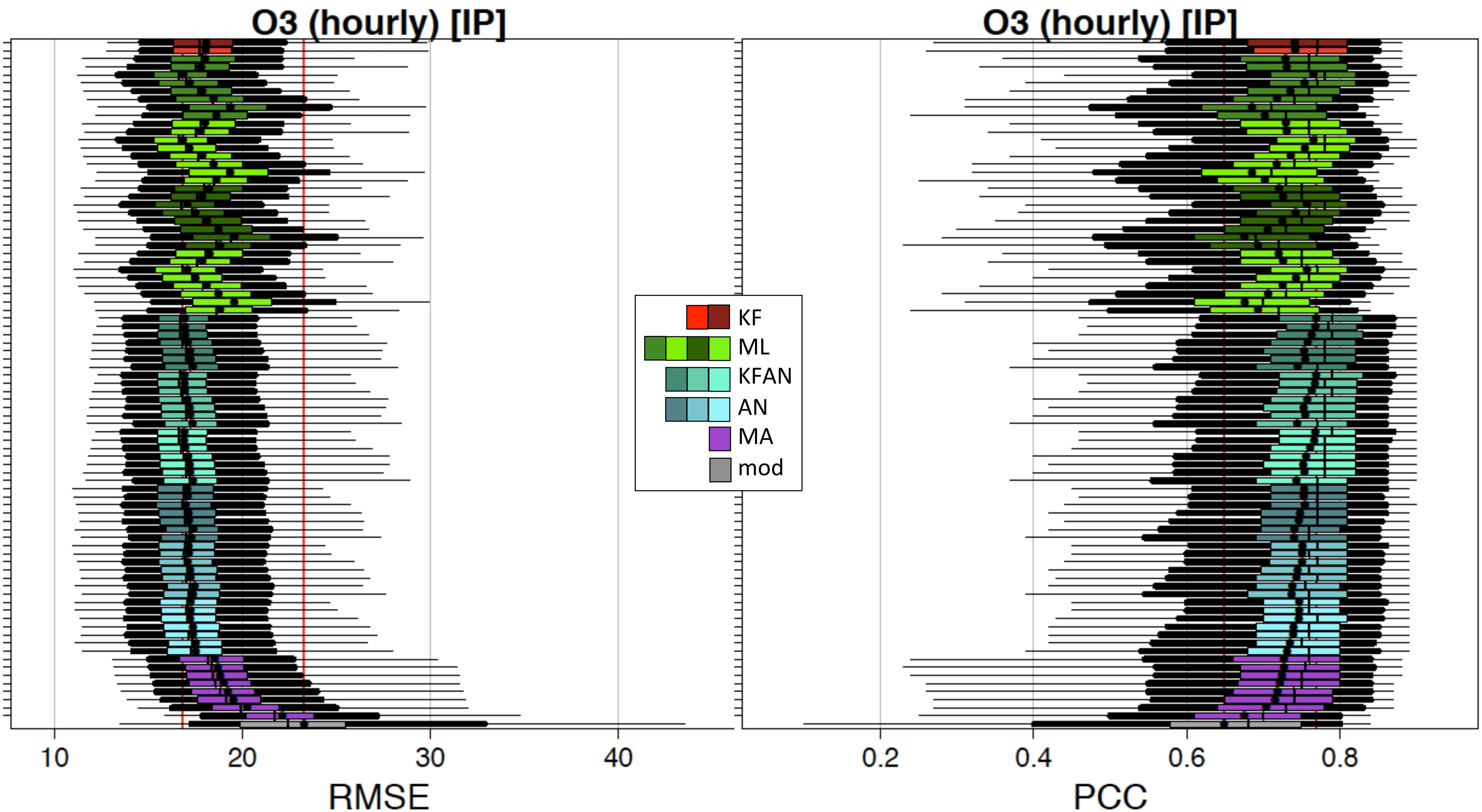
Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MOS correction on PM10

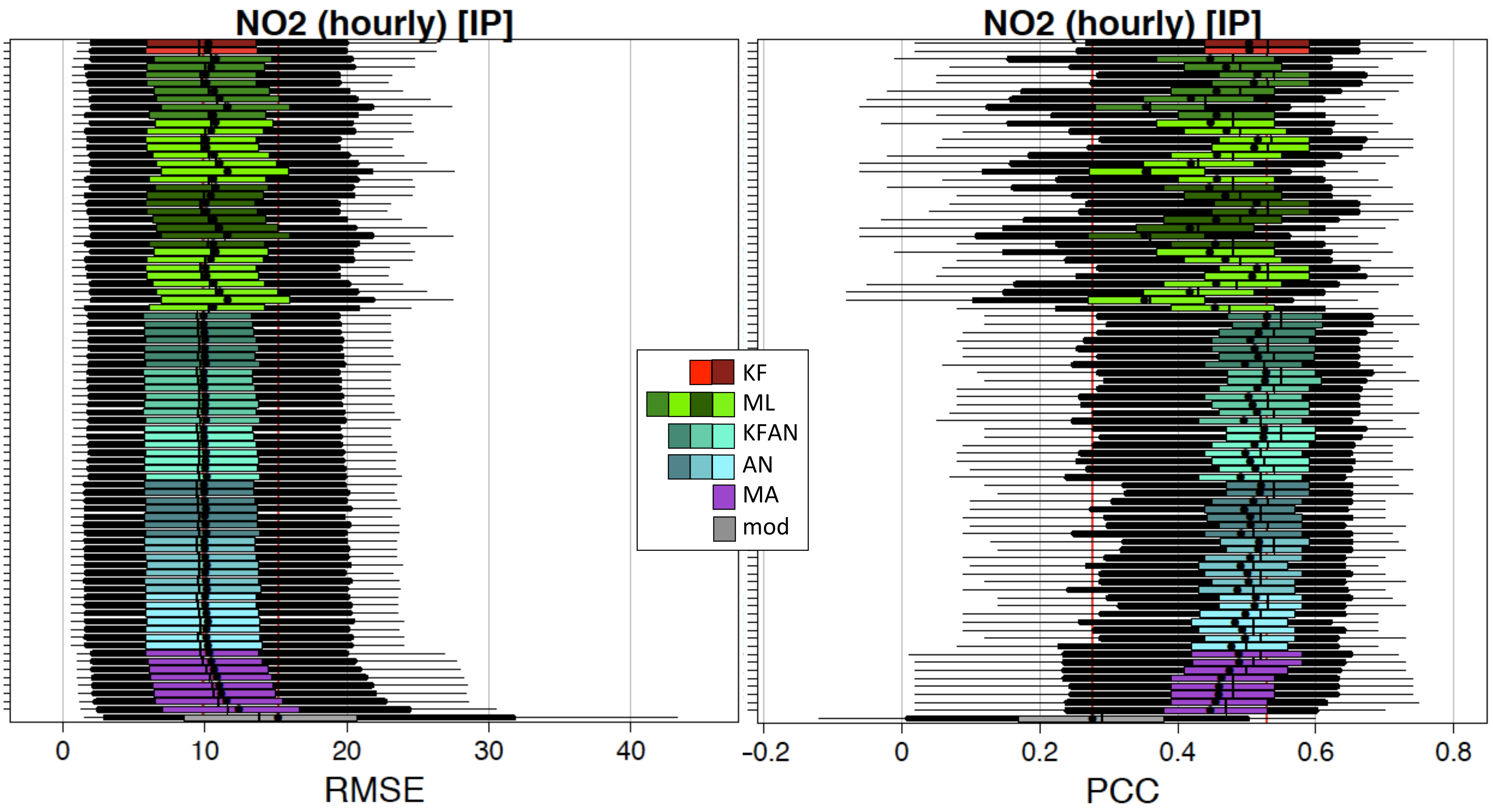


Barcelona Supercomputing Center
Centro Nacional de Supercomputación

MOS correction on O3

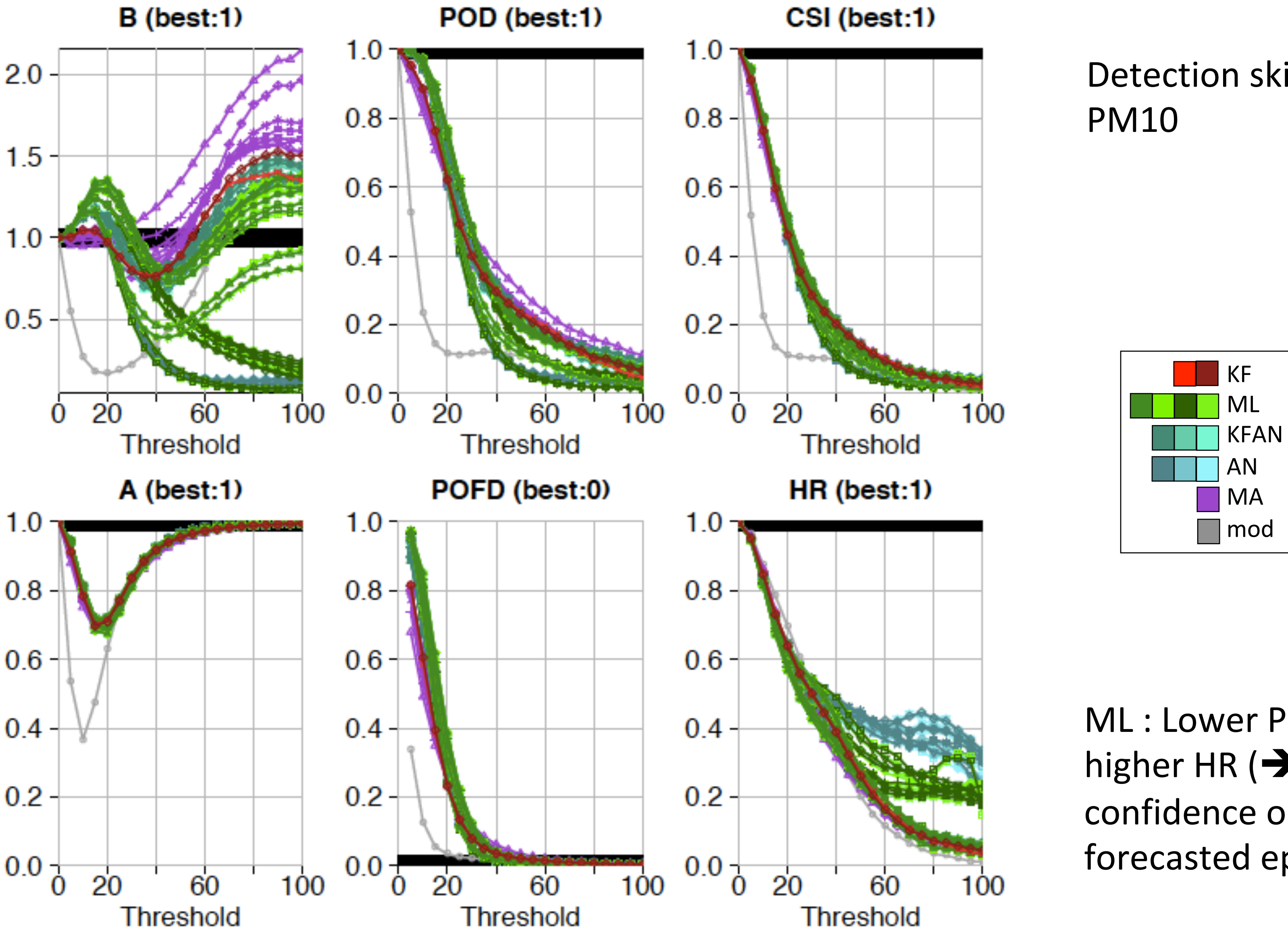


MOS correction on NO2



Barcelona Supercomputing Center
Centro Nacional de Supercomputación

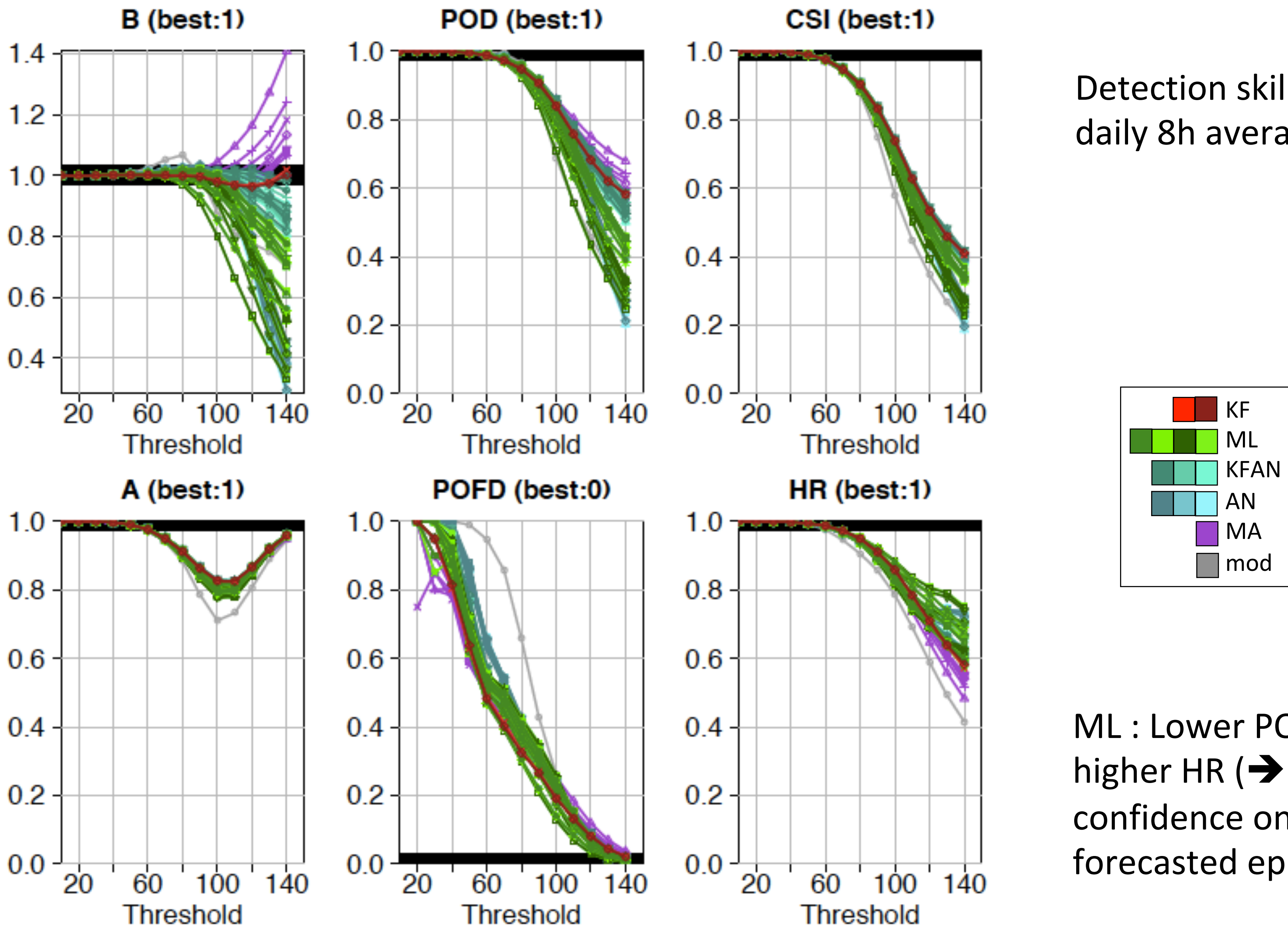
Episode detection



Detection skills – Hourly PM10

ML : Lower POD but higher HR (➔ more confidence on the forecasted episodes)

Episode detection



Detection skills – Max daily 8h average O3

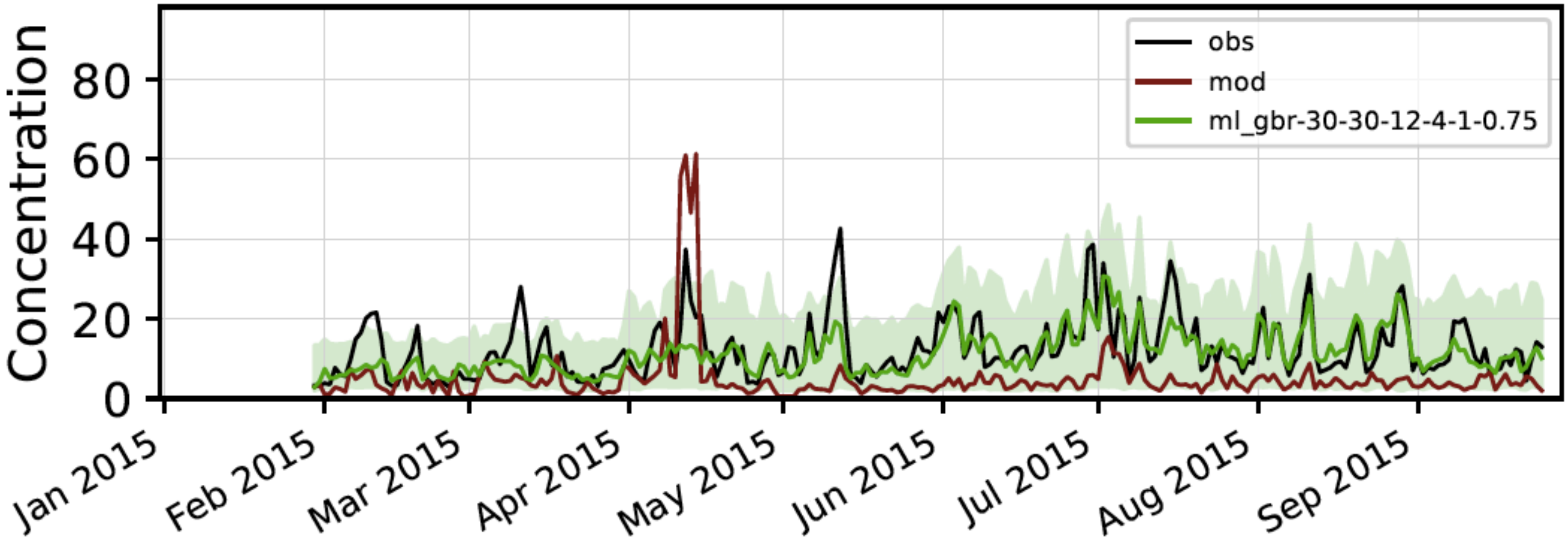
ML : Lower POD but higher HR (➔ more confidence on the forecasted episodes)

Conclusions

- Differences between the MOS methods generally quite consistent from one type of stations to another (URB, SUB, RUR) (not shown)
- Results with traditional MOS methods :
  - KF : reference
  - MA : best improvement with windows larger than 5 days, but lower performance compared
  - AN : often better than KF for both RMSE and PCC (despite very short dataset)
  - KFAN : even better than AN but surprisingly only for O3 and NO2 and not for PM10
- MOS-ML :
  - Often gives the best statistical results (again, despite very short dataset)
  - Tested with different sets of features, best results when many different features are taken into account : *model[day0], mod[day-1], mod[day-2], meteorology, meteorological gradients, hour of the day, error[day-1], error[day-2]*
  - Better results on PM10 when the target is the observed concentrations rather than the model error (no big difference for O3 and NO2)
- But better statistical results does not imply better skills for detecting episodes! MOS methods often smooth the variability, thus reducing the ability to detect extreme concentrations

On-going and planned work

- Run the new python script over all stations
- Investigate the effect of various data preprocessings (e.g. standardization, log-transformation), test other ML algorithms
- Quantile regression to get the prediction intervals (available only for GBM in python)



On-going work

****************************** hourly:

Proportion of observations with 5-95th ML predictions :	77.84% (N=6552 points) (after 2015-01-01)
Proportion of observations with 5-95th ML predictions :	87.40% (N=5808 points) (after 2015-02-01)
Proportion of observations with 5-95th ML predictions :	89.37% (N=5136 points) (after 2015-03-01)
Proportion of observations with 5-95th ML predictions :	89.73% (N=4392 points) (after 2015-04-01)
Proportion of observations with 5-95th ML predictions :	90.55% (N=3672 points) (after 2015-05-01)
Proportion of observations with 5-95th ML predictions :	91.43% (N=2928 points) (after 2015-06-01)
Proportion of observations with 5-95th ML predictions :	89.90% (N=2208 points) (after 2015-07-01)
Proportion of observations with 5-95th ML predictions :	90.44% (N=1464 points) (after 2015-08-01)
Proportion of observations with 5-95th ML predictions :	84.17% (N=720 points) (after 2015-09-01)

****************************** daily:

Proportion of observations with 5-95th ML predictions :	82.05% (N=273 points) (after 2015-01-01)
Proportion of observations with 5-95th ML predictions :	92.15% (N=242 points) (after 2015-02-01)
Proportion of observations with 5-95th ML predictions :	93.93% (N=214 points) (after 2015-03-01)
Proportion of observations with 5-95th ML predictions :	94.54% (N=183 points) (after 2015-04-01)
Proportion of observations with 5-95th ML predictions :	94.77% (N=153 points) (after 2015-05-01)
Proportion of observations with 5-95th ML predictions :	95.90% (N=122 points) (after 2015-06-01)
Proportion of observations with 5-95th ML predictions :	94.57% (N=92 points) (after 2015-07-01)
Proportion of observations with 5-95th ML predictions :	93.44% (N=61 points) (after 2015-08-01)
Proportion of observations with 5-95th ML predictions :	86.67% (N=30 points) (after 2015-09-01)

Barcelona Supercomputing Center
Centro Nacional de Supercomputación

EXCELENCIA SEVERO OCHOA

Thank you

herve.petetin@bsc.es

Episode detection metrics

	Event observed	Event not observed
Event forecasted	a	b
Event not forecasted	c	d

- **Error** : ERROR = (b+c)/(a+b+c+d)   complement of **Accuracy** : A=(a+d)/(a+b+c+d)
  - ➔ *How many events or non-events are erroneously classified?*
- **Probability of detection** : POD = a/(a+c)
  - ➔ *How many observed events have been well predicted by the model?*
- **Probability of false detection** : POFD = b/(b+d)
  - ➔ *How many observed non-events are erroneously classified as events by the model?*
- **Hit rate** : HR = a/(a+b)   complement of **False Alarm Ratio** : FAR = b/(a+b)
  - ➔ *Over all events forecasted by the model, how many are indeed observed?*
- **Critical success index** : CSI = a/(a+b+c)
  - ➔ *If we ignore the (numerous) non-events, how many events are correctly detected?*
- **Bias** : B=(a+b)/(a+c)
  - ➔ *Are we forecasting the correct number of events? (no matter when they occur or if the are correct)*

MONARCH alone

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Statistical performance of MONARCH - O3

- URB stations :
  - O3 dynamics reasonably well represented for all seasons
  - Overall strong positive bias
  - Strongest errors during morning transition whatever the season (strong positive bias) and late evening in winter (more random errors)
  - Highest (lowest) correlations during afternoon (early morning)
  - Underestimated variability, in particular during winter/fall afternoon

- RUR stations :
  - Some differences in O3 dynamics, notably in summer
  - Mainly moderate negative biases
  - Idem but different share between random and systematic errors
  - Idem
  - Stronger underestimation of the variability, in particular during night whatever the season and also afternoon in winter/fall

Barcelona
Supercomputing
Center
Centro Nacional de Supercomputación

Statistical performance of MONARCH – PM10

- URB stations :
  - PM10 dynamics poorly represented
  - Overall strong negative bias (in particular during morning)
  - Strongest errors during morning transition
  - Lowest correlations in winter/fall
  - Underestimated variability, in particular during winter/fall, better in spring

- RUR stations :
  - Idem
  - Idem (slightly lower bias)
  - Idem
  - Idem
  - Idem

Statistical performance of MONARCH – NO_2

- URB stations :
  - NO_2 dynamics quite well represented except morning peak (too low) and late evening (too persistent peak)
  - Overall strong negative bias (in particular during morning)
  - Strongest errors during early afternoon
  - Lowest correlations in summer/fall/spring
  - Underestimated variability, in particular during early afternoon

- RUR stations :
  - Idem
  - Idem (slightly lower bias)
  - Idem
  - Idem
  - Idem

Conclusions

- Statistical results in short (at annual scale) :
    - PM10 – URB/RUR : -60% bias, 100-110% error, 0.25 correlation
    - O3 – URB/RUR : ±10% bias, 30-40% error, 0.5-0.6 correlation
    - NO2 – URB : -60% bias, 100% error, 0.5 correlation
      RUR : -30% bias, 110% error, 0.3 correlation
- To be confirmed : maybe underestimated NOx and PM emissions during morning rush hours, which would explain :
    - Underestimated PM
    - Too low nitration of O3 ➔ strong positive bias
    - Too low NO2 and accumulation of this negative bias during morning ➔ strongest negative bias in early morning
    - NB : May also be at least partly due to the dilution in too coarse grid cells, and/or too deep PBL in the morning