- European Centre for Medium-Range Weather Forecasts -
Third International Workshop on Verification Methods
Reading 31 Jan – 2 Feb. 2007

A multi-model multi-analysis limited area ensemble: calibration issues

M. Marrocu

CRS4, Parco Scientifico e Tecnologico – POLARIS
Edificio 1, 09010 Pula (CA), Italy

email: *marino@crs4.it*

P. A. Chessa

Servizio Agrometeorologico della Sardegna,
Viale Porto Torres 119, 07100, Sassari, Italy

email: *chessa@sar.sardegna.it*

MUSE
a *Multimodel-multianalysis ensemble*

4 LAMs : BOLAM - MM5
RAMS1 – RAMS2

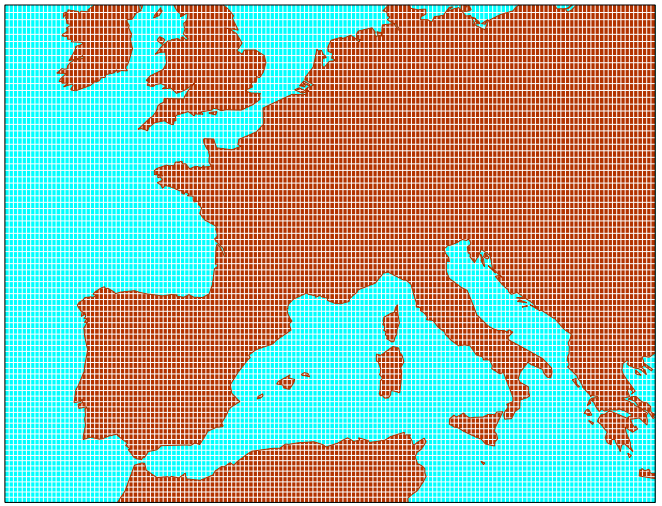
2 I.C & B.C.: AVN 12Z - ECMWF 12Z

Area: 13.5W-34N / 24.5E-54.5N

Spatial Resolution: 0.25°

Fct time range: +72h (by 6 h steps)

Integration period: 15/10/2002 to 15/04/2003 (183 days)



Thanks to C. Dessy, G. Ficca, C. Castiglia, I. di Piazza

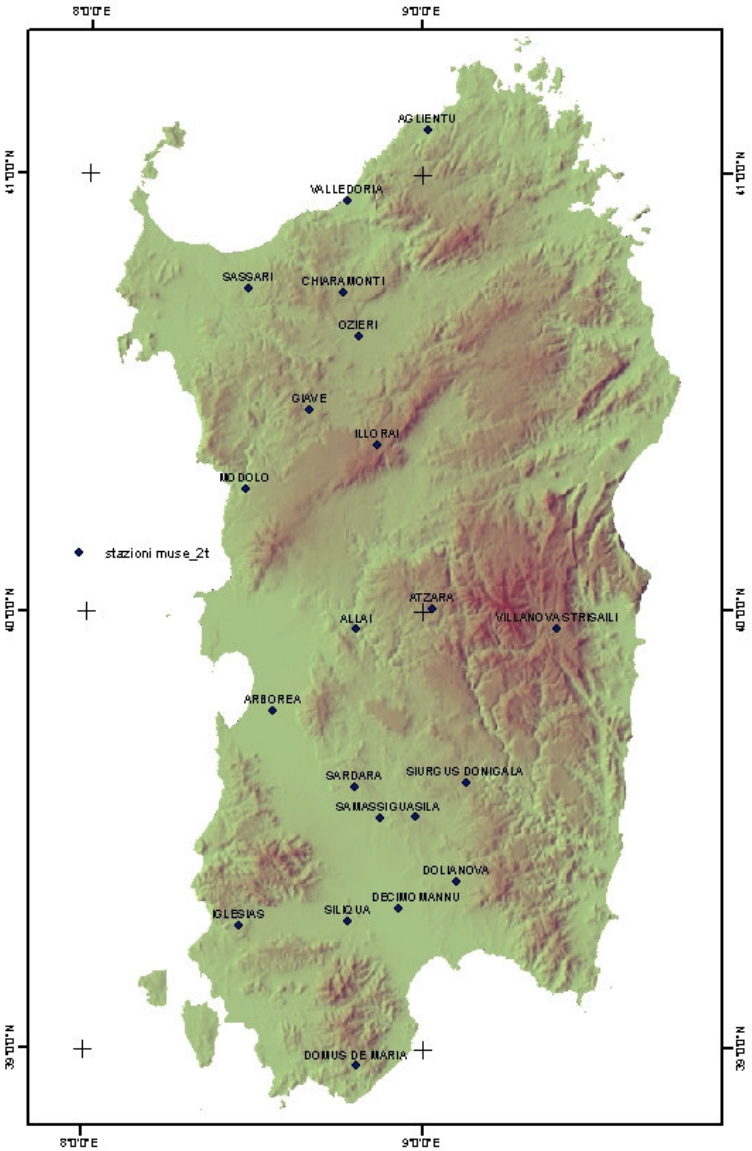
OPERATIONAL IN MARCH 2007

Measured data

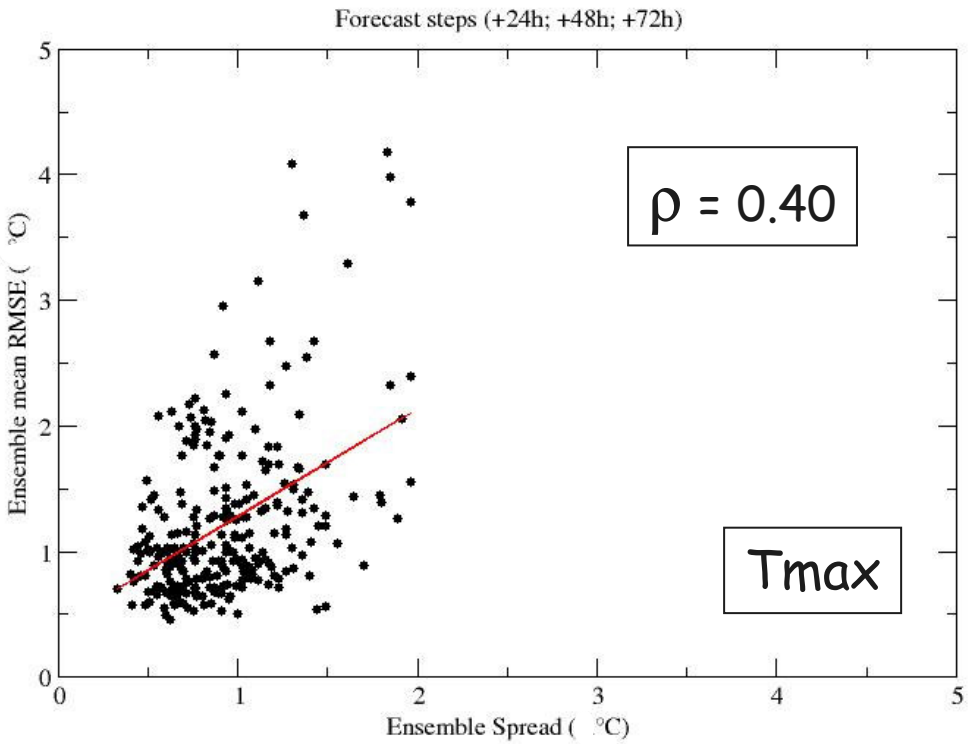
The calibration assessment is done for a continuous variable with a relatively simple PDF. Namely, the 2m temperature.

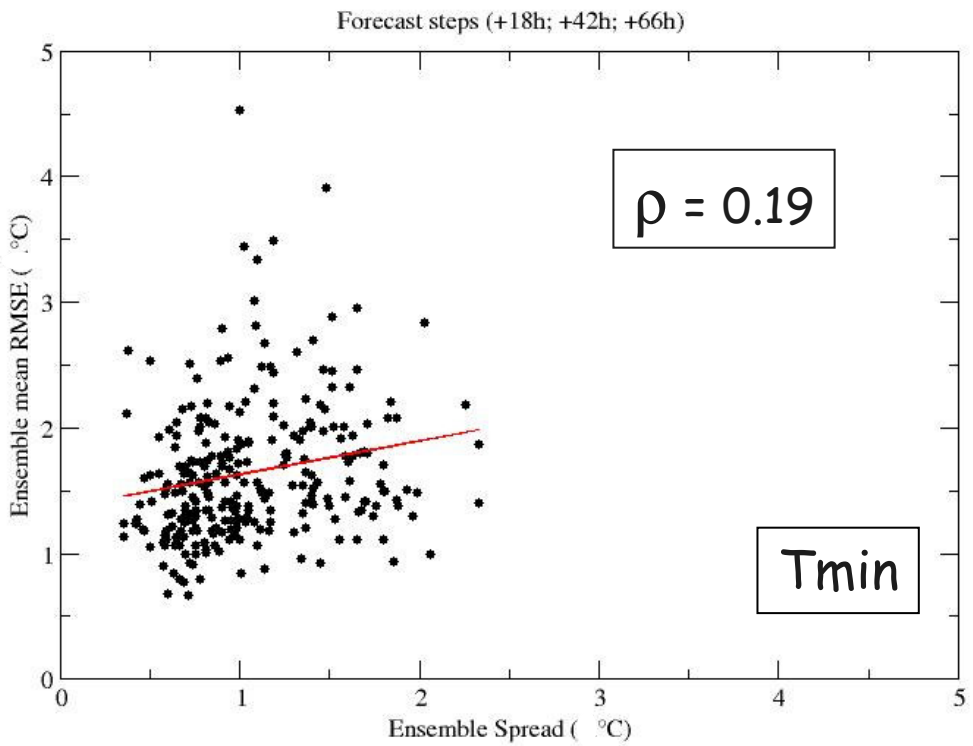
For the 186 days, all 6-hourly measured data were collected from 21 ground meteorological stations located in Sardinia.

These stations were singled out from the whole network (about 60 stations), because they were the sole having no missing data.



Spread-skill relationship

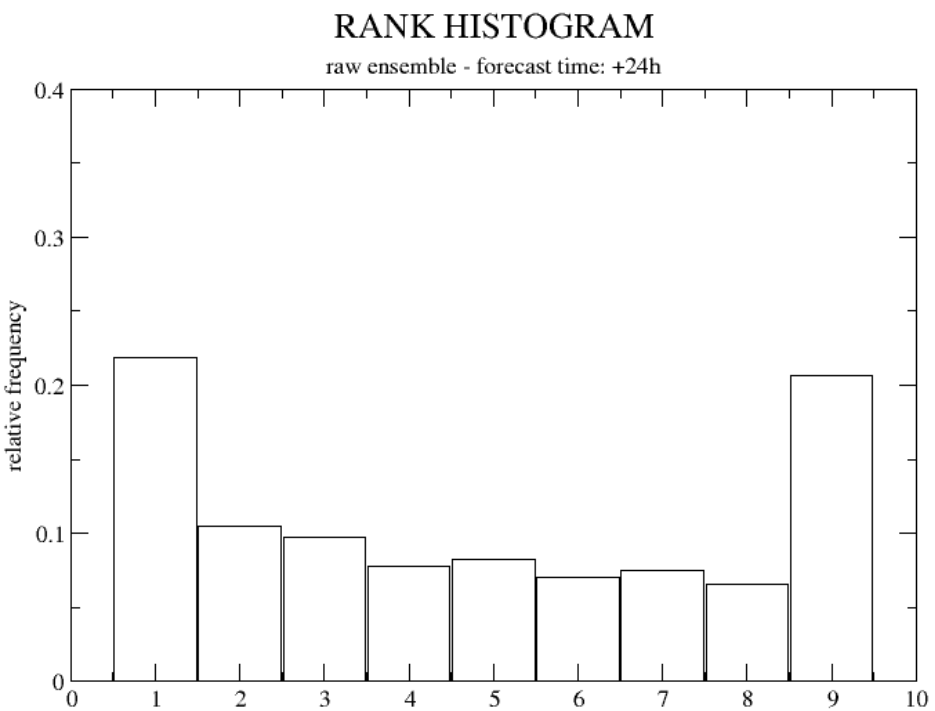


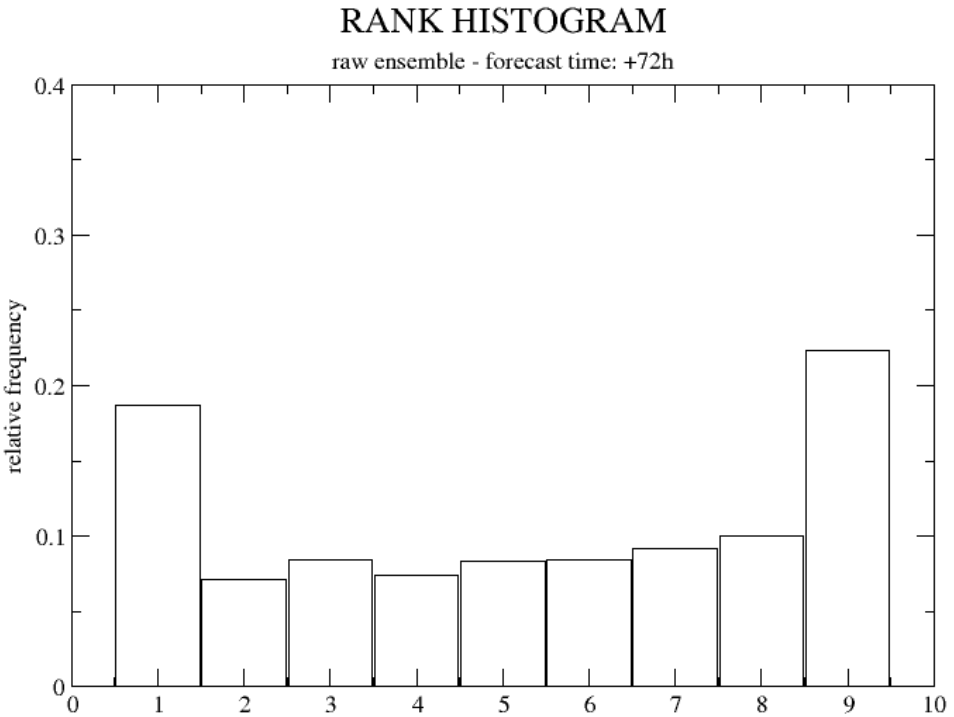


NOTE.

The variability of the spread-skill relationship across the forecast time steps reflects on the RMSE of the deterministic forecasts and on the calibration outcomes.

Why calibrate ?





The ensemble is under-dispersive and the single forecasts are clearly not equi-probable. Calibration should reduce the under-dispersion, provide a suitable weight for each member and, hopefully, increase the sharpness of the resulting distribution.

Calibration methods - 1

- Bayesiam Model Averaging (BMA)

$$p(o|f_1,.....,f_K)=\sum_{m=1} w_m G_m(o|f_m) \qquad \text{where} \qquad G(o|f_m)=\mathcal{N}(a_m+b_m f_m,\sigma_m{}^2)$$

w_m and σ_m are estimated by maximum likelihood and in a further step the variance is refined minimizing the Continuous Ranked Probability Score,

$CRPS=\frac{1}{K}\sum_{j=1}^{K}\int_{-\infty}^{+\infty}\{F_j(z)-H(z-t_j)\}^2\,dz$, over the training period. F(z) is the Cumulative Distribution Function of G while H is the Heaviside function.

Ref.: A. E. Raftery et al. - MWR 2005

Calibration methods - 2

- Ensemble model output statistics (EMOS)

The EMOS PDF is expressed as:

$$\mathcal{N}(\alpha + \beta_1 f_1 + ... + \beta_K f_K ; \gamma^2 + \delta^2 S^2)$$

(ensemble spread)

the coefficients are calculated minimizing the CRPS over the training period

- Modified ensemble model output statistics (EMOS$^+$)

CRPS minimization iterated: after each step models associated to negative β_i are drop out from the next iteration. The process stops when all β_i left are positive. Id est: ensemble retains only forecasts providing a skilful contribution.

Ref.: T. Gneiting et al. - MWR 2005

Calibration methods - 3

- Dressing kernel

The covariance of the stochastic values η to be added to the *dynamical* forecasts f, is calculated in a way that renders the, seasonally averaged, variance of the dressed ensemble and that of the observation, indistinguishable. That is to say that:

with $F_{dress} = f + \eta$ then $\overline{\eta^T \eta} = \overline{(\overline{f_i} - o_i)^T (\overline{f_i} - o_i)} - \overline{\sigma_i^2}$

(sample mean and variance of true forecast PDF)

(observations)

The means are taken over all forecast-observation occurrences in the training period.

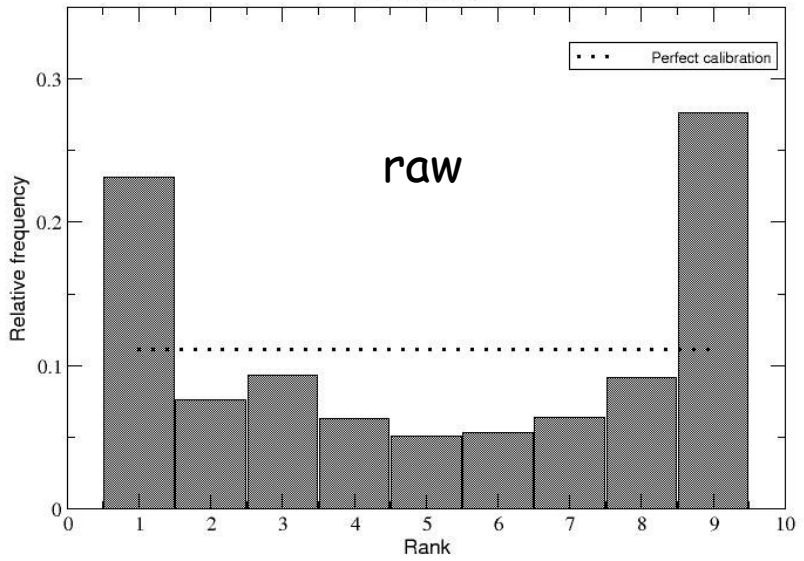
The number of perturbations to be added to each dynamical forecast was set to 32.

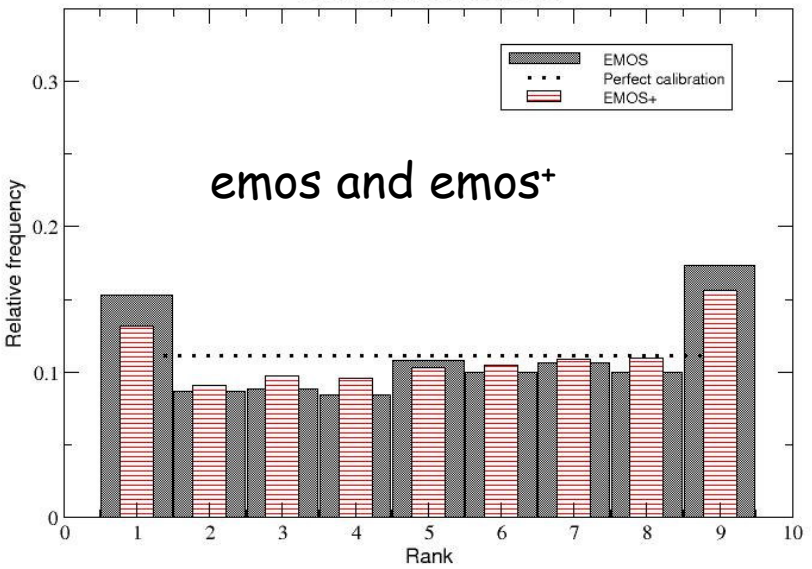
Ref.: Wang and Bishop - QJRM 2005

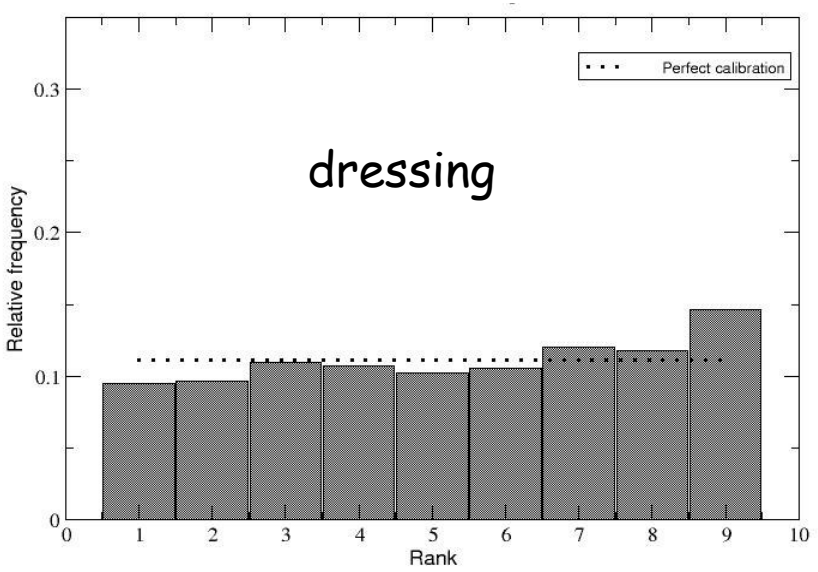
Training period

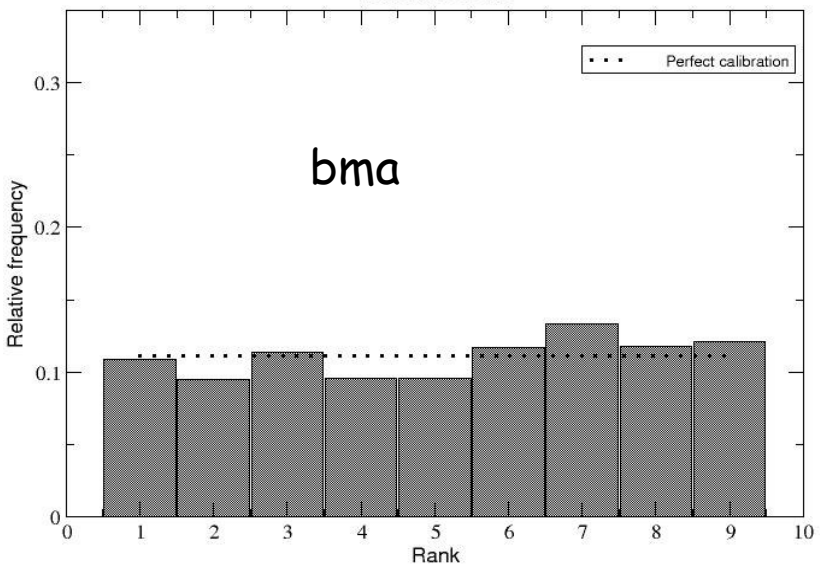
- The training period is a sliding-window varying from time step to time step.To define it, a few quantities used to evaluate the calibration quality (the rank histogram, the PDF's coverage and width, the RMSE for the related deterministic forecasts) were used.
- In practice the chosen interval length is such that a longer training period do not bring any improvement on the calibration scores.
- In this case this happens between 60 and 90 days. In the following results are based on a 90 days training period.
- In order to test the robustness of the techniques and its independence from the training set, all the calculation were also accomplished swapping training and testing periods. The final results did not change.

Calibration: rank histograms (+66h)

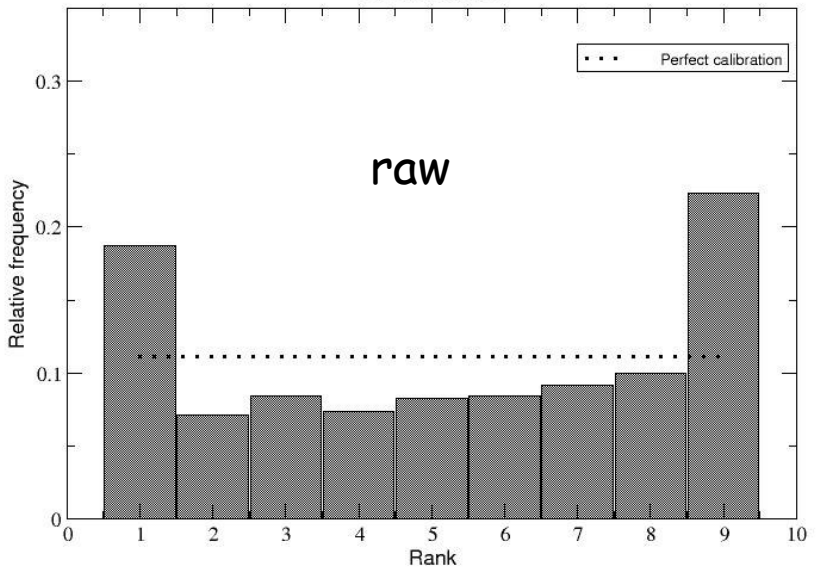


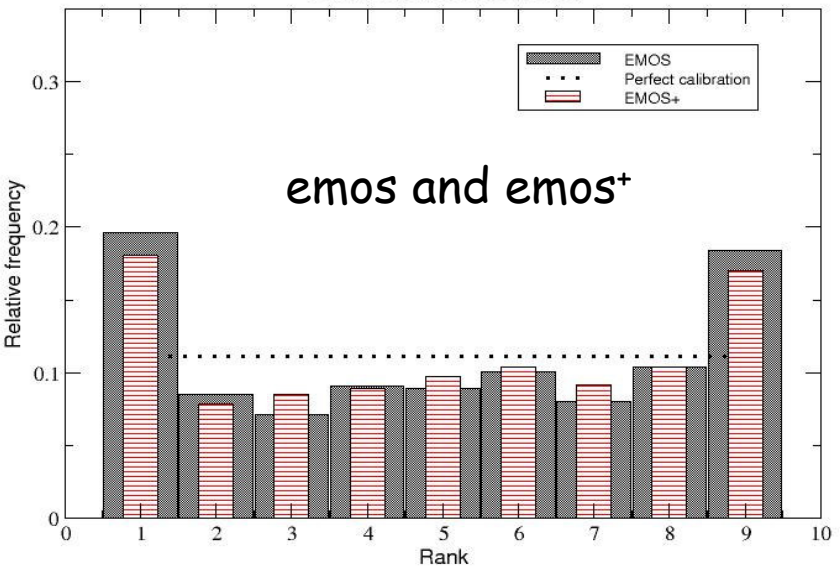


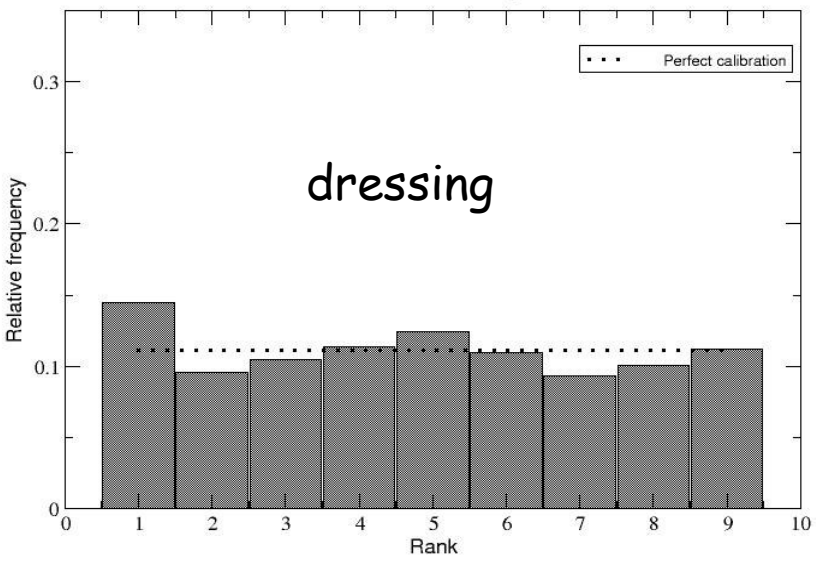


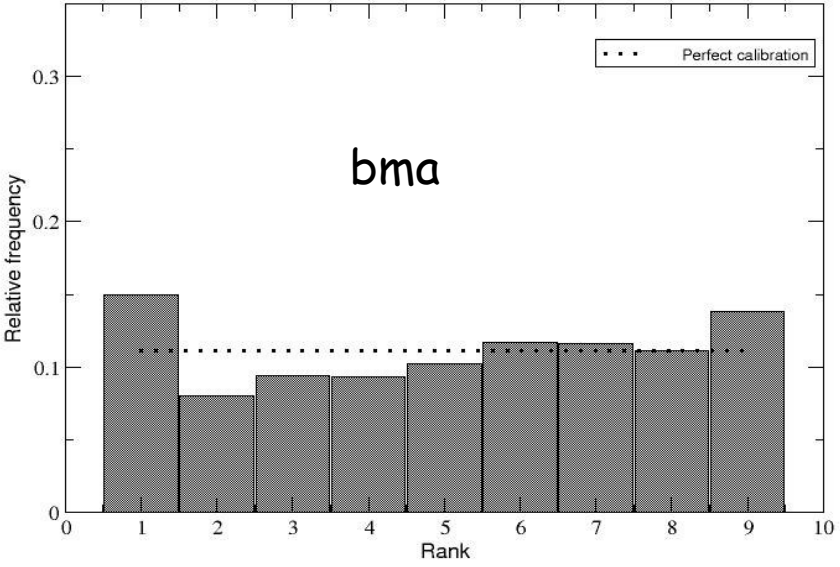


Calibration: rank histograms (+72h)

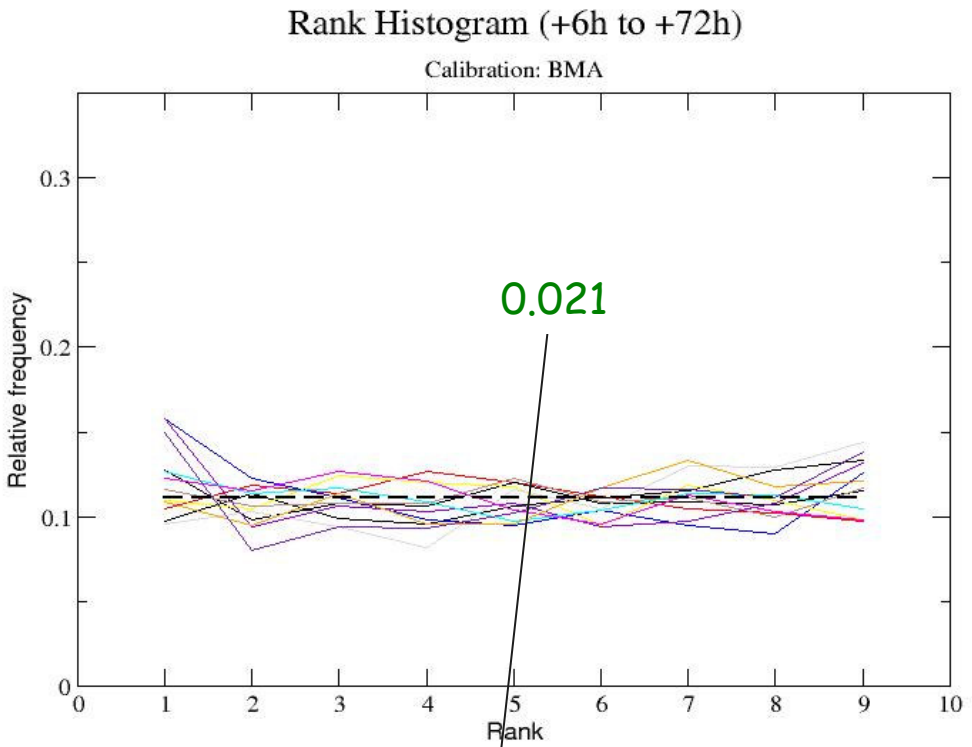


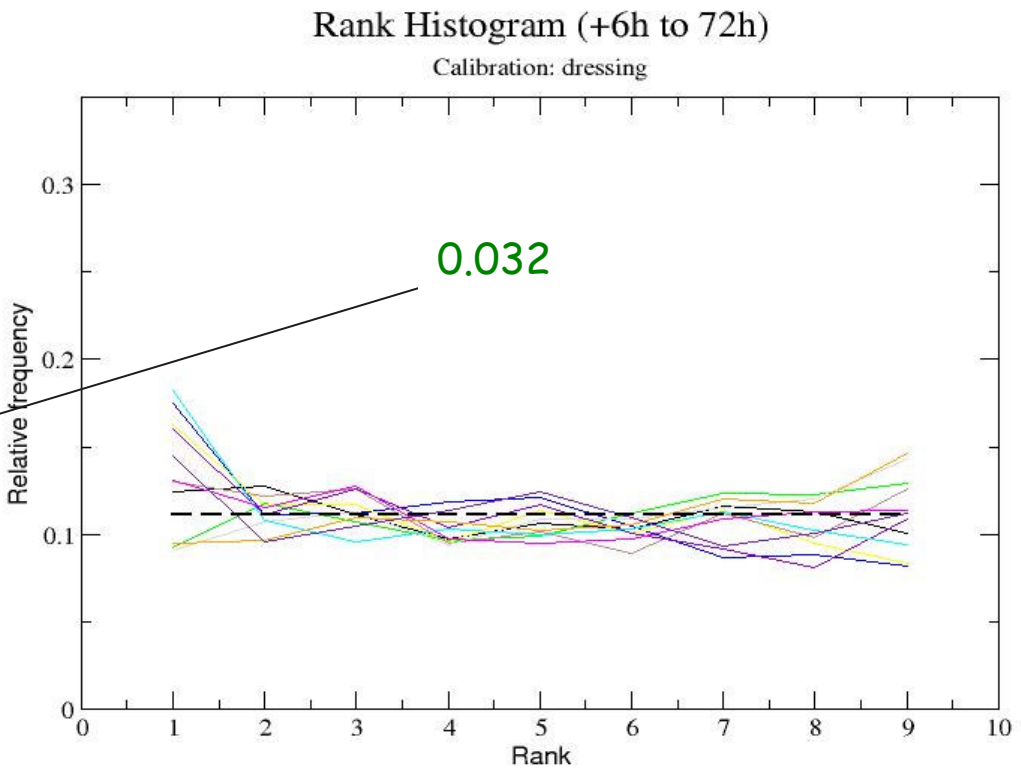






Calibration: rank histograms (all steps)

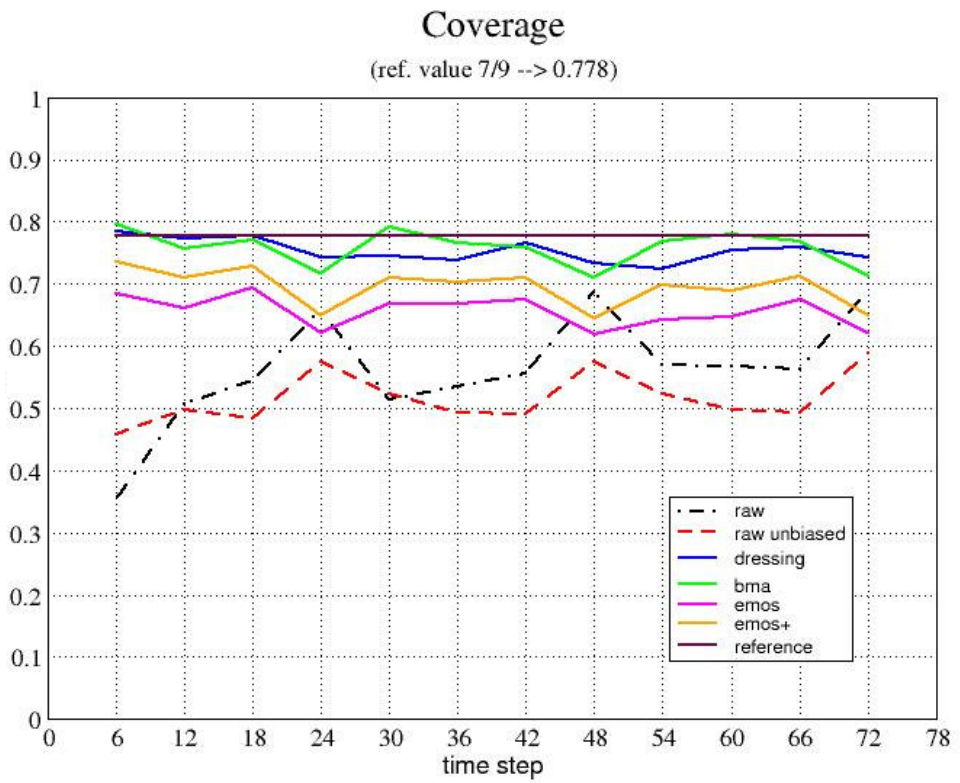


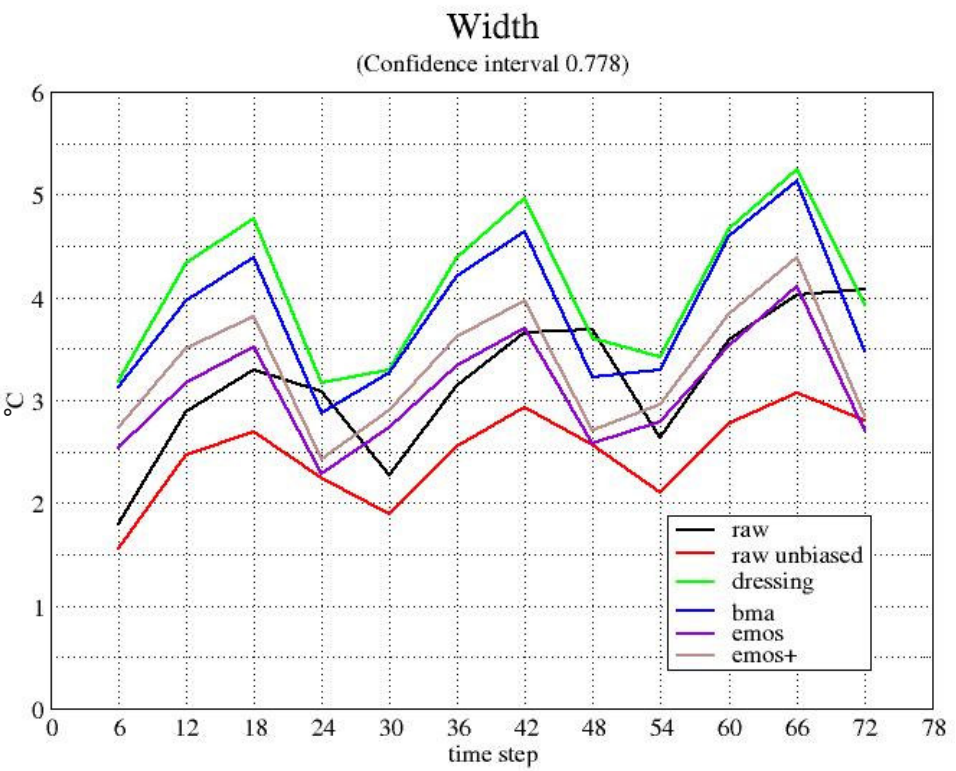


"Root mean square error"

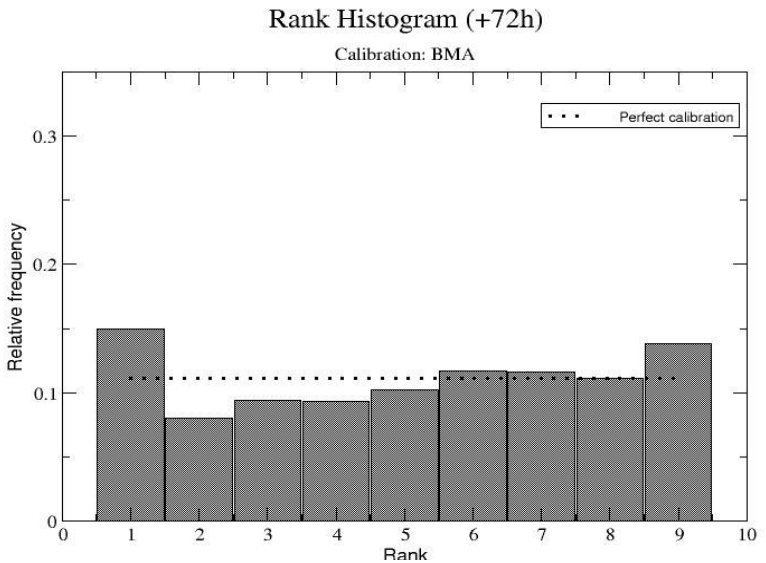
outliers

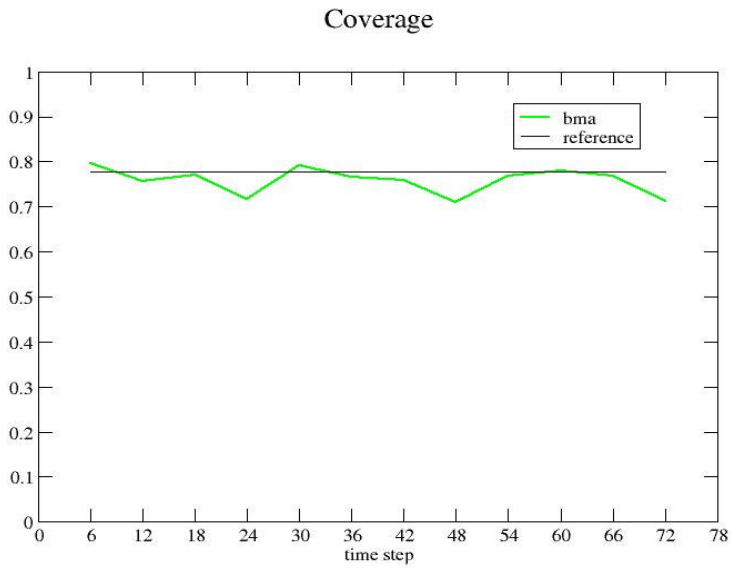
Calibration: coverage and width

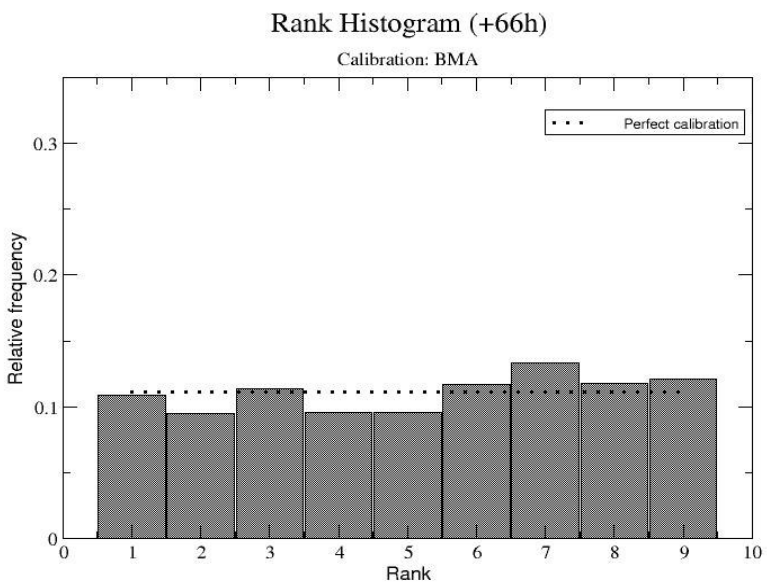


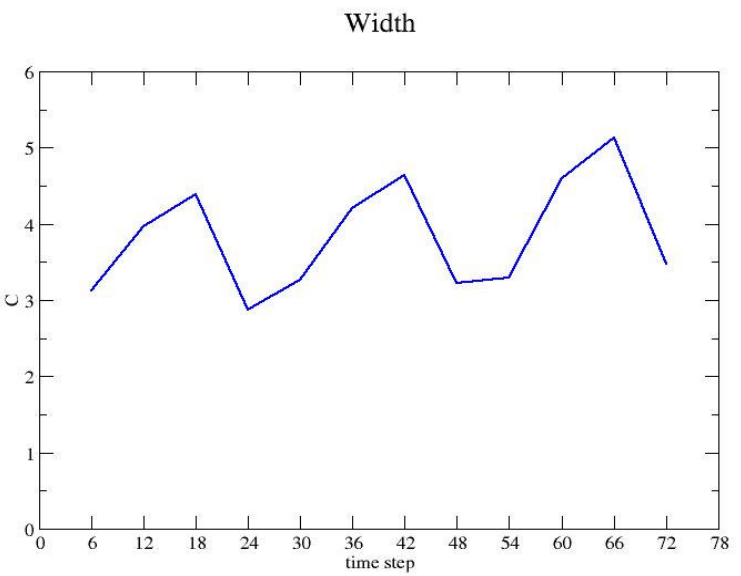


Calibration: coverage and width

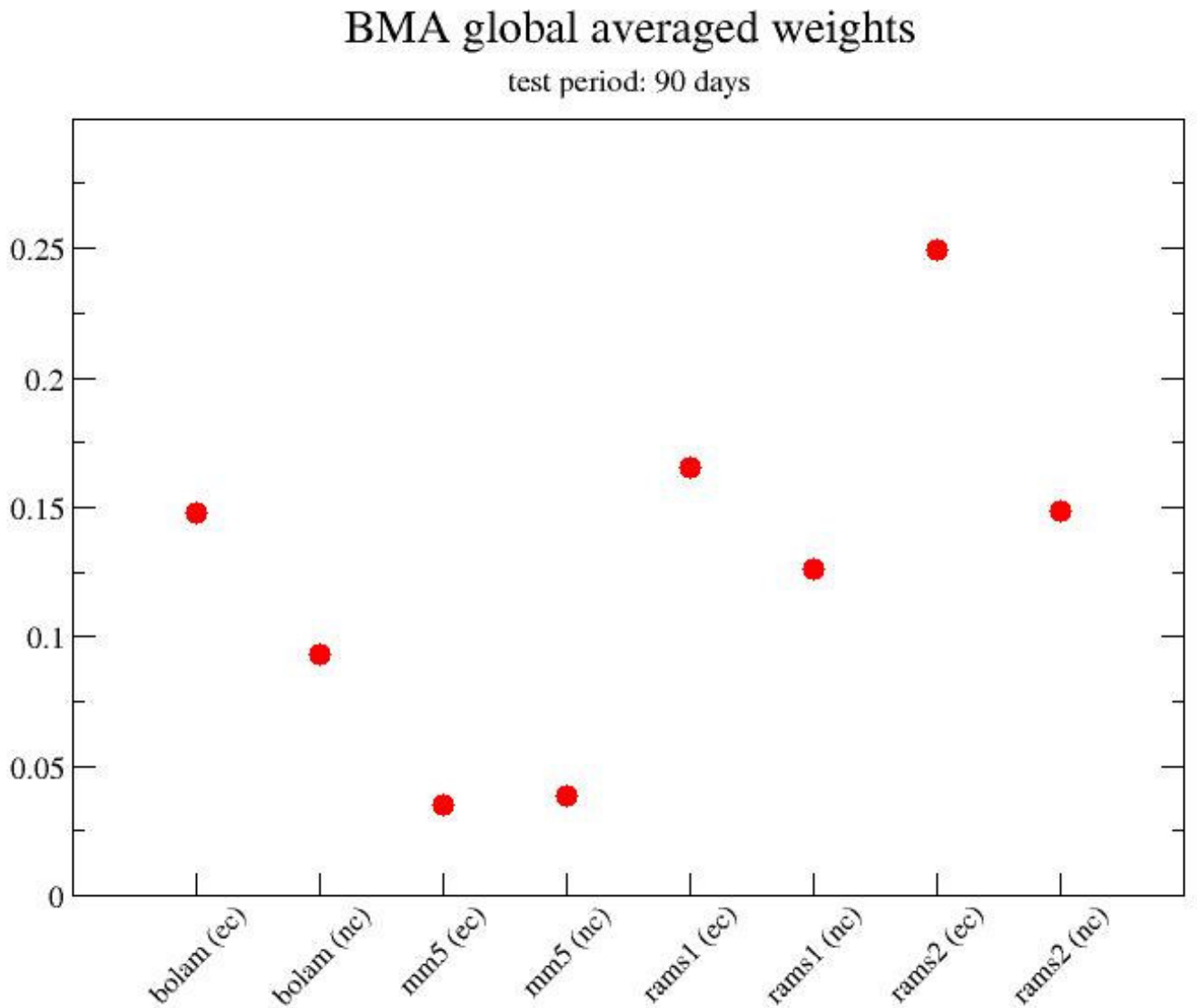








BMA weights


BMA global averaged weights
test period: 90 days
0.25
0.2
0.15
0.1
0.05
0
bolam (ec)
bolam (nc)
mm5 (ec)
mm5 (nc)
rams1 (ec)
rams1 (nc)
rams2 (ec)
rams2 (nc)

Expectation values

The expectation value of the PDFs for BMA, EMOS and EMOS$^+$, and the "dressed" ensemble mean are deterministic forecasts on their own.

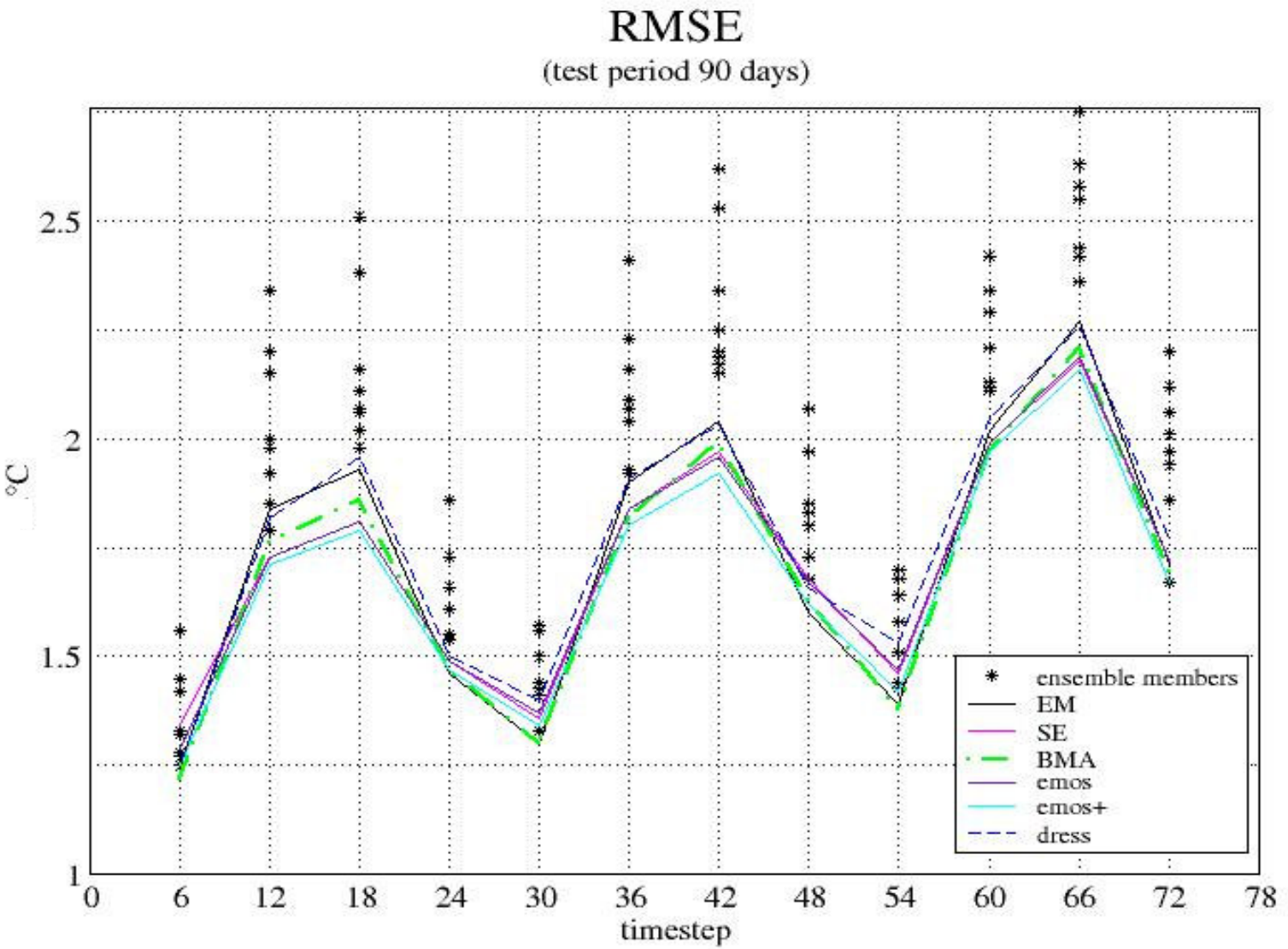
For instance for BMA is: $\mu_{BMA} = \sum_{m=1}^{K} w_m (a_m + b_m f_m)$

Scores like RMSE and MAE have been calculated for all of them and compared to the likes of: each ensemble member, the "unbiased" ensemble mean and the "super-ensemble".

Why so ?

The hope was to unveil a behaviour so good to gain for free, and for a system which inherently lacks it, a reference (control) forecast directly from the calibration method.

Deterministic forecasts



Conclusions

- Calibration for 2m temperature works well both with BMA and DRESSING. (Easy the extension to temperature at pressure levels and to other continuous variables as MSLP, geopotential, etc.)
- BMA shows more consistent results than DRESSING across the forecast time steps, especially for the external intervals (outliers). Moreover, BMA weights are directly interpretable in terms of probabilities.
- Deterministic scores for the expectation values of calibration methods, the "dressed" ensemble mean, the "unbiased" ensemble mean and the super-ensemble are similar. All of them outperform, on average, the best model. Therefore, once a calibration method is chosen, it is argued that the expectation value can be used as reference/control forecast for the ensemble.

Future work

- Calibration is going to be implemented on MUSE (needs a good amount of computer power)
- SPITLOMS: a ECMWF special project (SAR – CRS4 – Italian MetService) aimed at exploring the potential of longer and more structured training periods.
- Calibration for wind and precipitation is going to be shortly addressed (need a careful analysis of the underlying PDF and probably, for precipitation, longer training sets).