Zentralanstalt für Meteorologie und Geodynamik

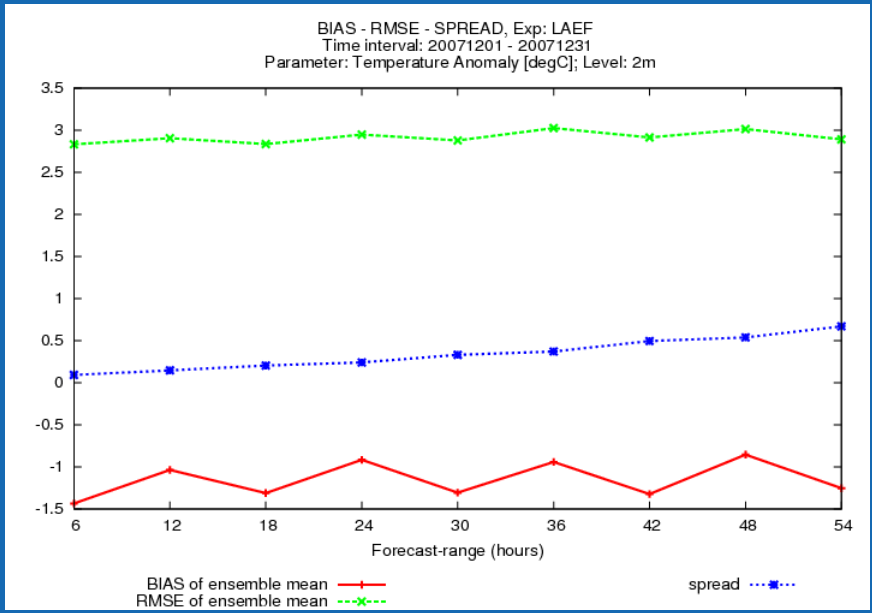
On the potential added value of calibrating LAM-EPS

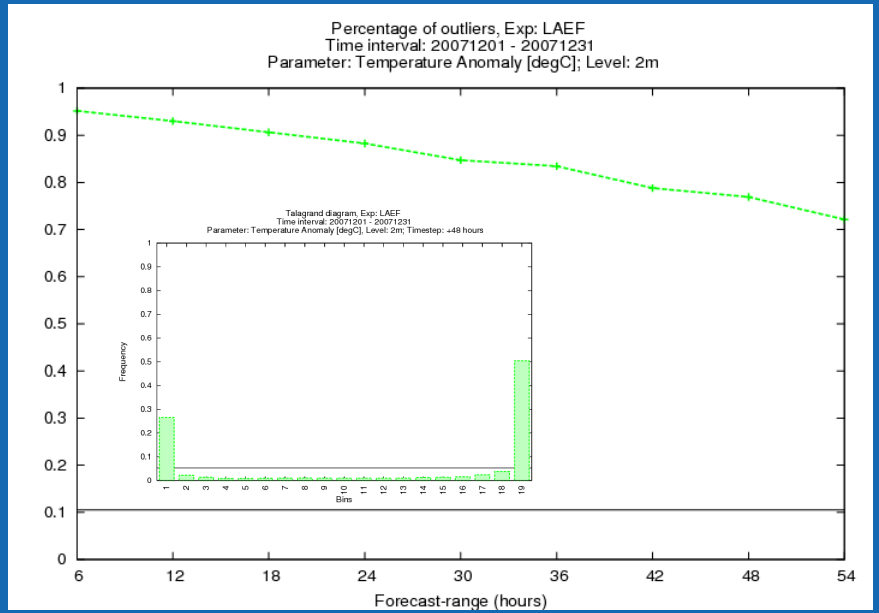
Alexander Kann
Central Institue for Meteorology and Geodynamics
Vienna, Austria

Overview

- Why we need a statistical post-processing of LAEF
- Methodological aspects
  - Bias correction (First moment calibration)
  - Full calibration (Second moment calibration)
- Raw Forecast vs. Bias Corrected Forecast vs. Calibrated Forecast: Verification results
- The impact of weighting on bias correction
- The impact of training size on calibration
- The impact of spread re-scaling on calibration
- Calibrating ECMWF or LAEF?
- Concluding remarks and ongoing activities

Need for statistical post-processing of LAEF





Bias, RMSE of Ensemble Mean and Ensemble Spread of T2M, verified against observation: The system has a cold bias and significantly lacks spread.

Percentage of outliers and Talagrand diagram for T2M. 70% - 95% of observation fall outside the EPS system‘s range, with a slight cold bias.

Bias correction method

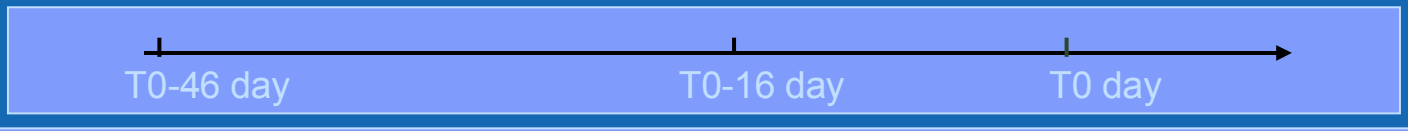
Aim: Increase the skill and utility by maximizing

- ✓ *Reliability* (Forecast Probability = Observed Relative Frequency; Statistical Consistency: Mean Square Error of EPS Mean = Ensemble Variance)
- ✓ *Sharpness* (Ability to distinguish between events and non-events; narrow, sharp PDF is better, probabilities closer to 0% or 100%)

→ Challenge: Maximize Sharpness while ensuring Reliability

Bias Assessment: adaptive (Kalman filter type) algorithm

Implementation of decaying averaging for the first moment bias (from Bo Cui, NCEP, 2006)



decaying averaging mean error = (1-*w*) * prior t.m.e + *w* * (f – a)

a) Prior estimate to startup procedure: choose T0 as current date (00Z), calculate the time mean errors between T-46 and T-16 day.

b) Update: the prior estimate of the average state is multiplied by a factor 1-*w* (<1). Then, most recent verification error (f - a) is added to the decaying average for each lead time with a weight of *w* (operational: *w* = 2%).

c) Cycling: repeat step (b) every day.

d) Carry out steps (a) to (c) for each variable, for each lead time and on every grid point.

e) The bias correction is applied on each ensemble member.

Calibration method

- Gneiting et. al., 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Mon. Wea. Rev.*, **133**, 1098 – 1118.
- Hagedorn, R., Hamill, T., Whitaker, J., 2007: Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part I: 2-meter Temperatures. Mon. Wea. Rev., in press.

- Idea of NGR (Non-homogenous Gaussian regression):

  Based on multiple linear regression, addresses forecast bias and underdispersion. NGR yields probabilistic forecasts with Gaussian PDF's for continuous weather variables. The predictive mean is a bias-corrected average of the ensemble member forecasts. The predictive variance is a linear function of the ensemble variance. Fitting the regression coefficients, the method of minimum CRPS estimation is used (optimizing CRPS for the training data).

$$PDF = N(a + bX, c + dS^2)$$

a, b: bias & general performance of ensemble mean
c = 0; d = 1: large spread-skill relationship
d = 0: small spread-skill relationship

Calibration method (cont.)

Analytical function CRPS of the coefficients a, b, c, d:

$$CRPS_{train} = \frac{1}{k}\sum_{i=1}^{k}(c + dS_i^2)^{\frac{1}{2}}\left\{Z_i\left[2\Phi(Z_i) - 1\right] + 2\varphi(Z_i) - \frac{1}{\sqrt{\pi}}\right\}$$

with:

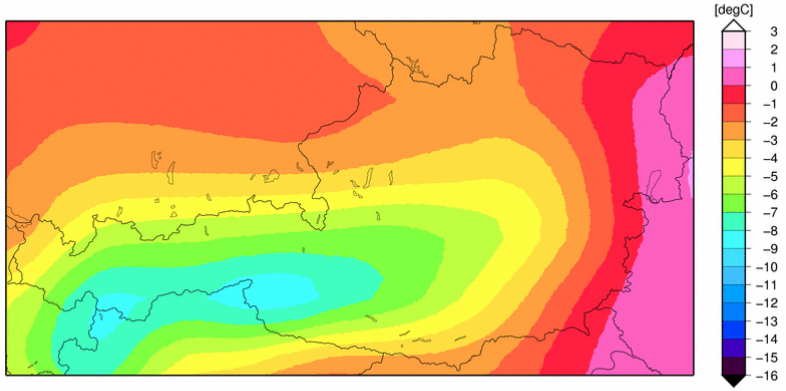
$$Z_i = \frac{Y_i - (a + bX_i)}{(c + dS_i^2)^{\frac{1}{2}}}$$

The regression coefficients are found iteratively using amoeba.f90 (Numerical recipes) routine (algorithm by Nelder and Mead).
Applying the regression coefficients on the current EPS forecast creates a predictive PDF, from which samples of any size can be generated. In our implementation, an 18 member ensemble is formed by taking the inner x% quantiles in such a way, that the new spread of the calibrated ensemble is bounded by f_{resc}*RMSE of the training data (x is found iteratively, f_{resc}=re-scaling factor: $0.5<f_{resc}<1$). This spread re-scaling has been implemented in order to avoid artificial large (due to statistical correction) ensemble spread, which lacks “synoptical” sharpness.
The calibration is done with INCA analyses on a 1km*1km horizontal grid covering Austria, a sliding 50 day training period is used. The experiment is carried out for one month (December 2007). Verification is done using station observation only.

Calibration results: Forecast fields: Ensemble Mean

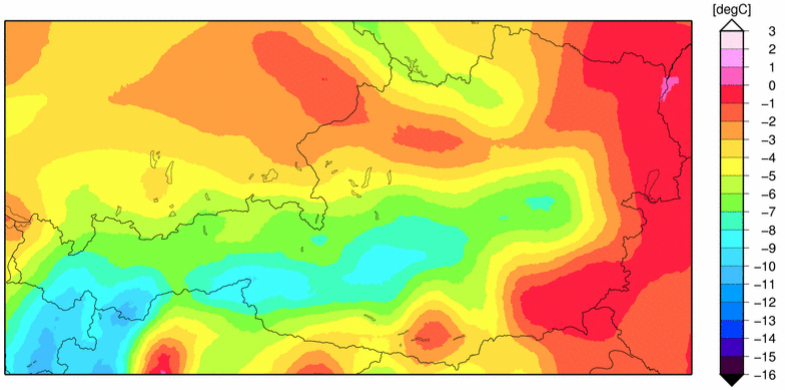
ECMWF: Uncalibrated 2m Temperature, Ensemble Mean

Forecast from: 20071216 00 UTC + 36h



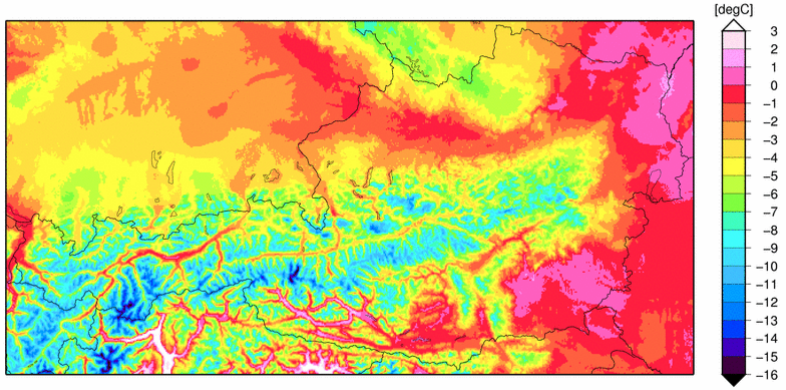
LAEF: Uncalibrated 2m Temperature, Ensemble Mean

Forecast from: 20071216, 00 UTC + 36h



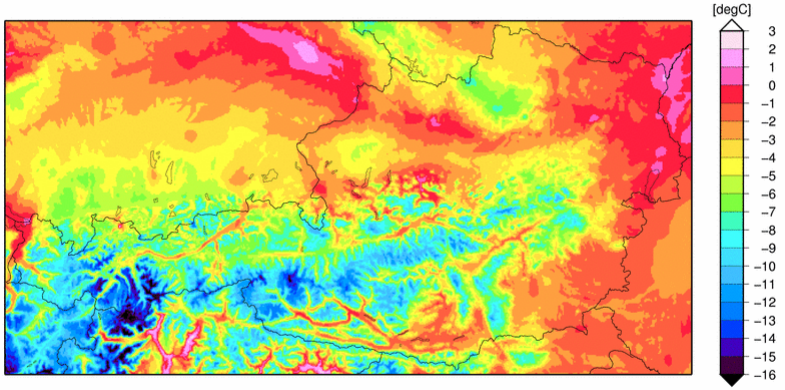
LAEF: Calibrated 2m Temperature, Ensemble Mean

Forecast from: 20071216, 00 UTC + 36h

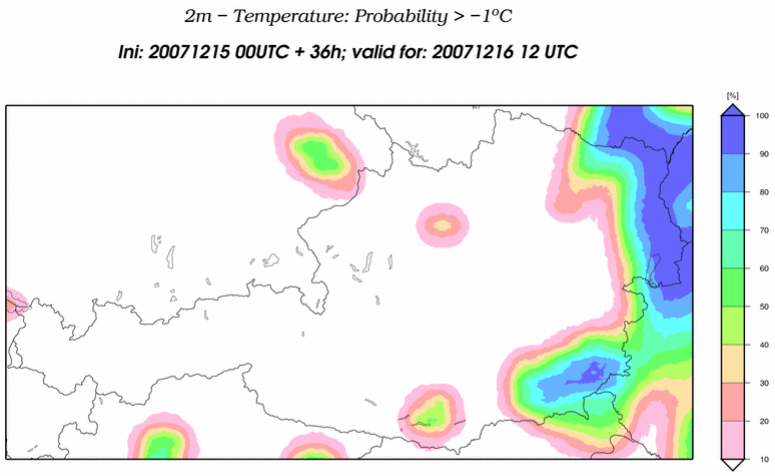


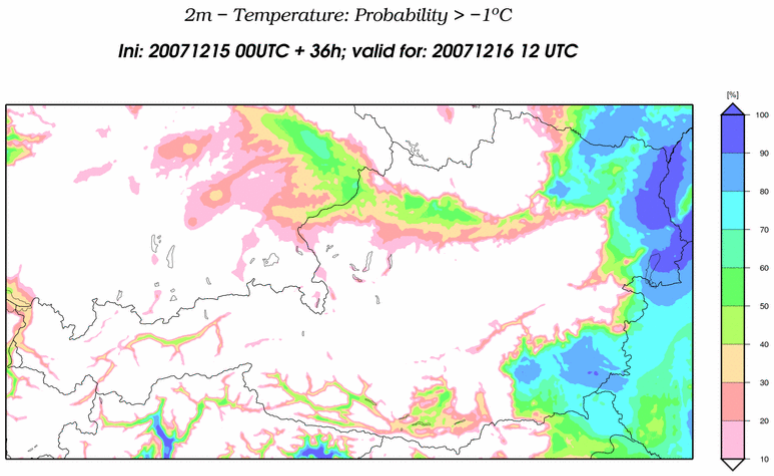
INCA: 2m Temperature

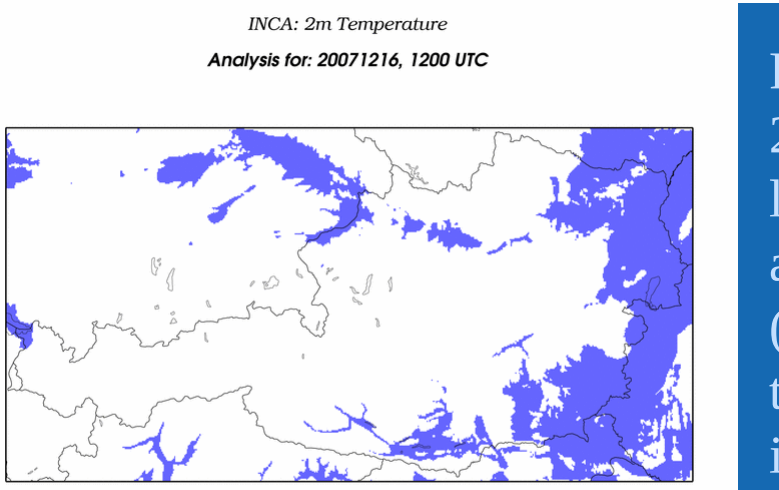
Analysis for: 20071217, 1200 UTC



Calibration results: Forecast fields: Probability Charts

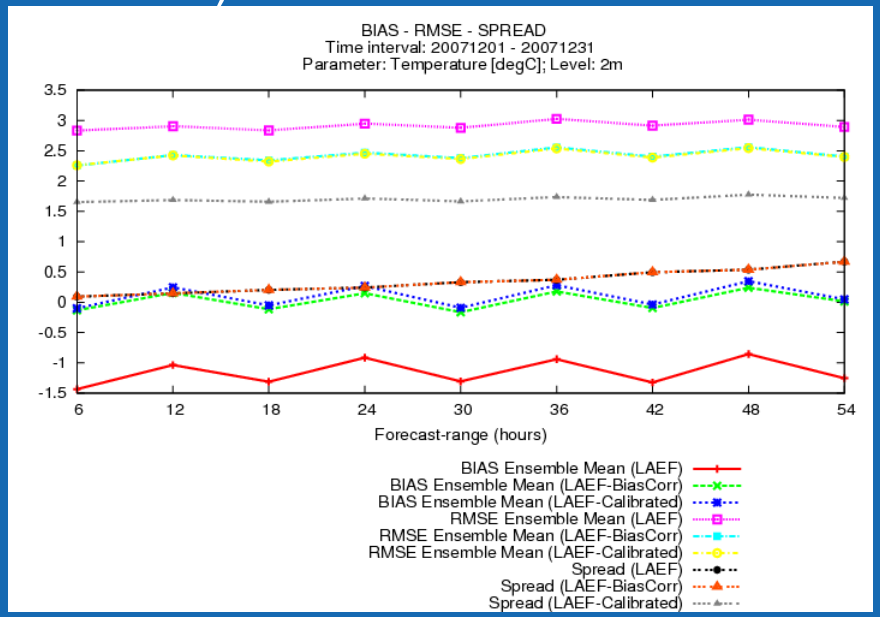


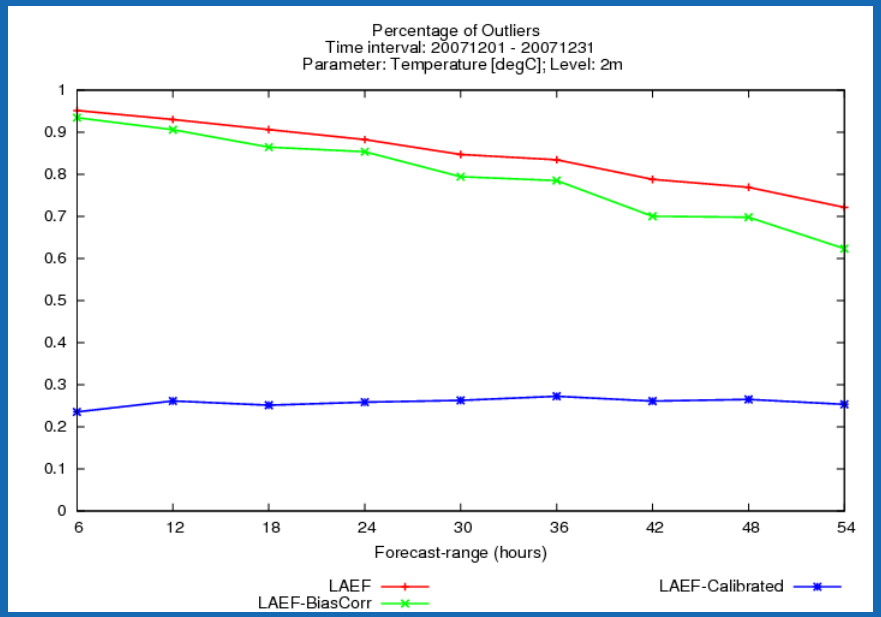




Probability plot for T2M>-1°C, forecast from 20071215, 00UTC + 36hours. Raw LAEF (top left), calibrated LAEF (top right) and INCA analysis showing areas exceeding -1°C in blue (bottom left). Although LAEF roughly covers the areas, the calibration is able to add information particularly on local scale.

Calibration results: Verification (Bias, RMSE, Spread; Outliers)

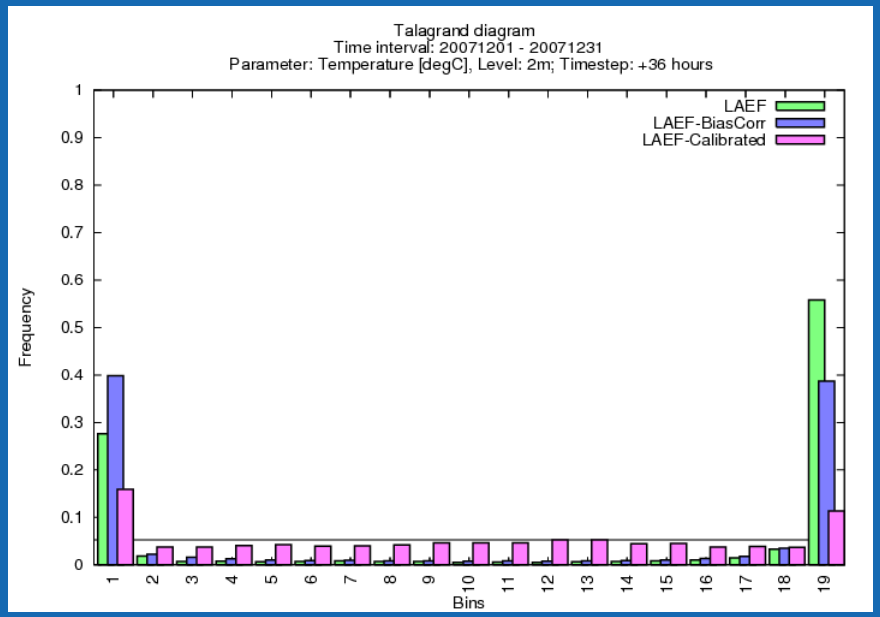




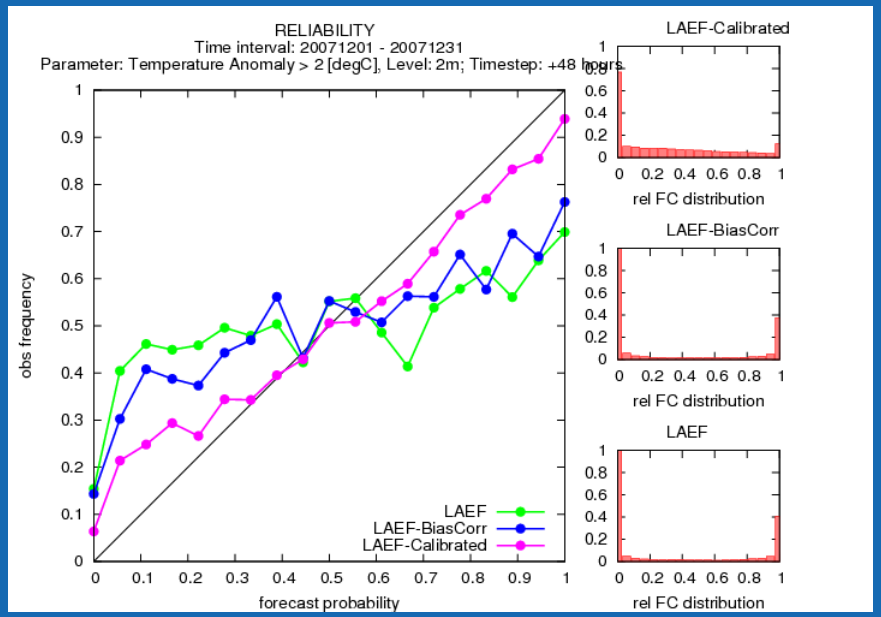
Bias, RMSE of Ensemble Mean and Ensemble Spread as a function of lead time for 2m temperature of raw LAEF, bias-corrected EPS and calibrated EPS. Bias correction leads to reduction of RMSE (from ~3K to ~2,4K), the spread is increased up to 1,5K by calibration.

Percentage of outliers as a function of lead time for raw LAEF (red), bias-corrected LAEF (green) and calibrated LAEF (blue). The bias correction shifts the PDF and therefore slightly reduces the number of outliers, but with full calibration the percentage of outliers decreases to about 30% -35%.

Calibration results: Verification (Talagramm; Reliability)

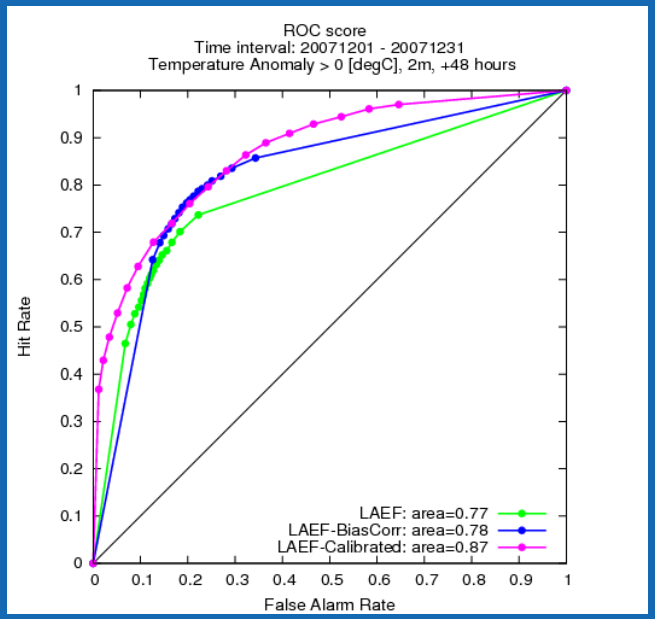


Talagrand diagram for 2m temperature, lead time +36hours. LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). The distribution becomes much flatter by calibrating, although it still remains slightly underdispersive.

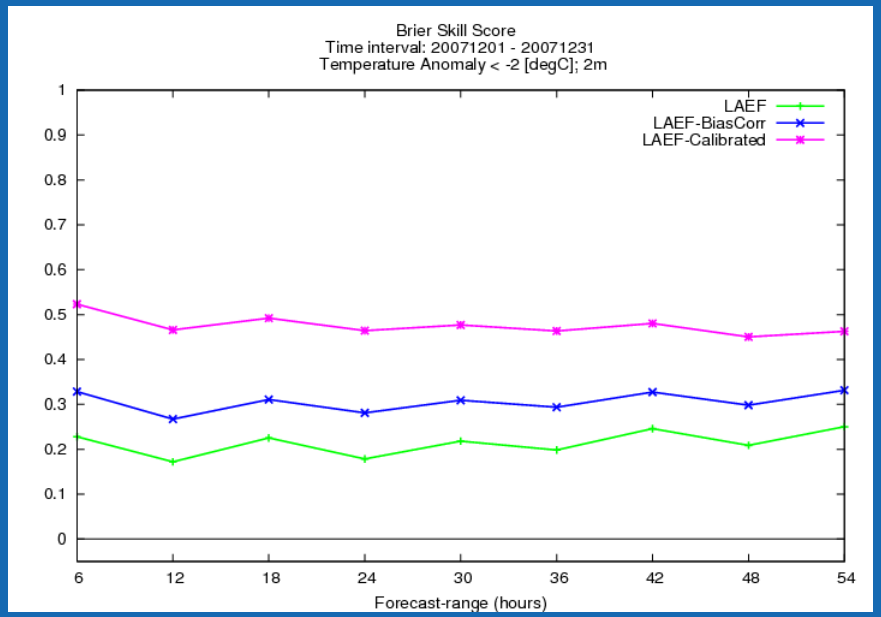


Reliability diagram for raw LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). The calibrated ensemble performs best, although they all tend to overforecast high probabilities (low probabilities are underforecast).

Calibration results: Verification (ROC; Brier Skill Score)

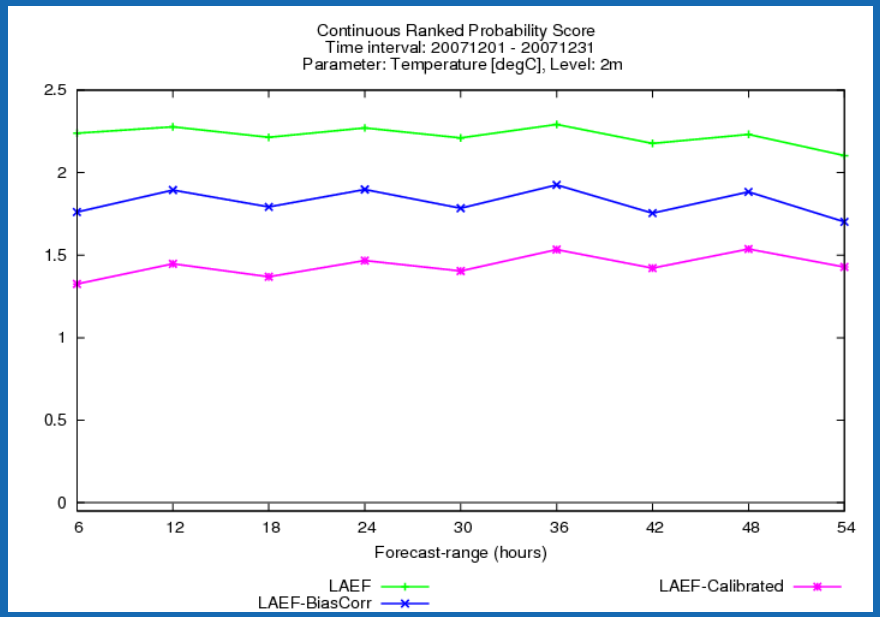


ROC curve and area under the ROC curve for 2m temperature anomaly > 0°C, lead time +48hours. LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). The area under the ROC curve for calibrated ensemble is about 10% higher than for raw or bias-corrected ensemble.

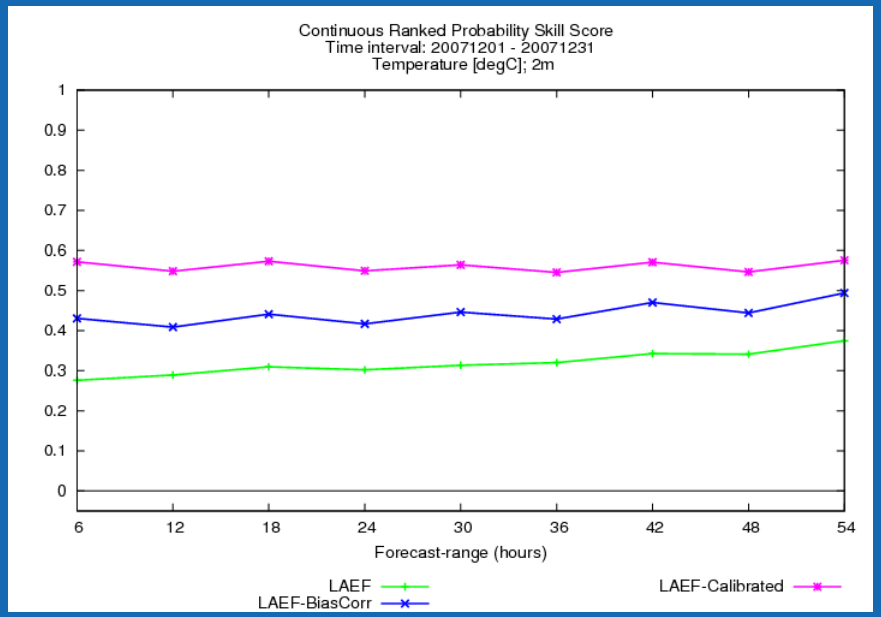


Brier Skill Score for 2m temperature anomaly <-2°C; raw LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). The calibrated ensemble performs much better, 30% - 40% of the improvement is achieved by the bias-correction.

Calibration results: Verification (CRPS; CRPSS)

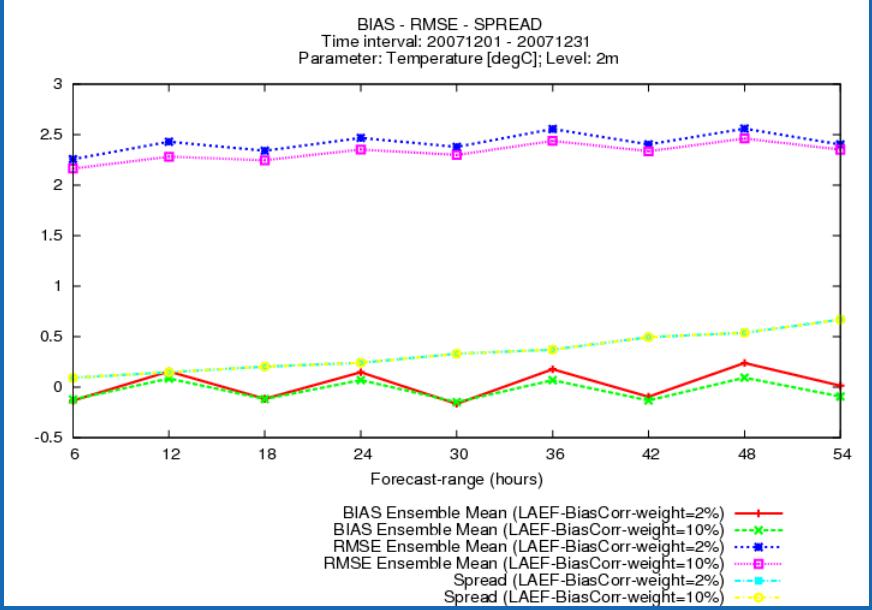


CRPS for T2M as a function of lead time. LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). About 50% of the calibration improvement is achieved by bias correction, the CRPS is decreased from ~2,3K to ~1,5K !!

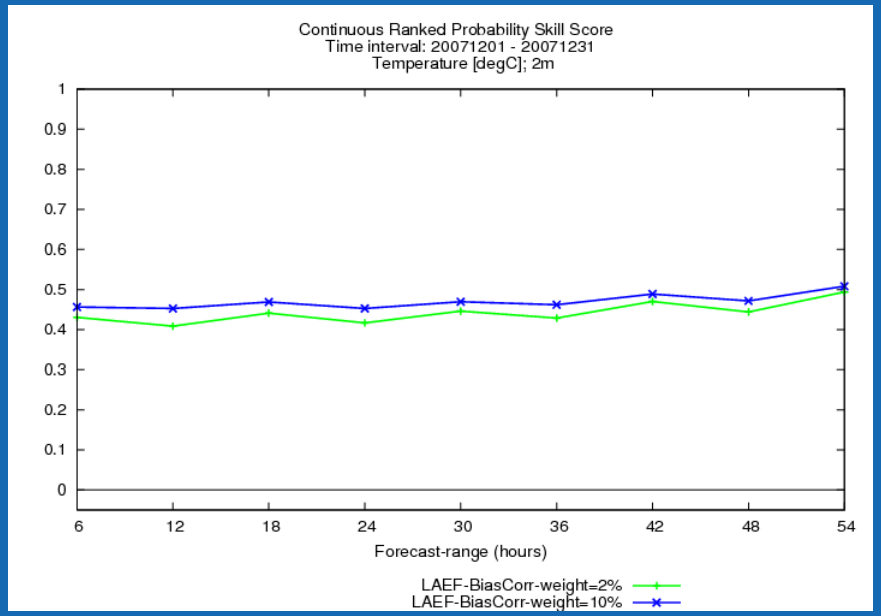


CRPSS (with deterministic Aladin-Austria as a reference) for T2M as a function of lead time. LAEF (green), bias-corrected LAEF (blue) and calibrated LAEF (purple). Again, about 50% of the total calibration improvement is obtained by bias correction, the CRPSS is approx. doubled from ~0,3 to almost 0,6 !

Impact of weighting on bias correction

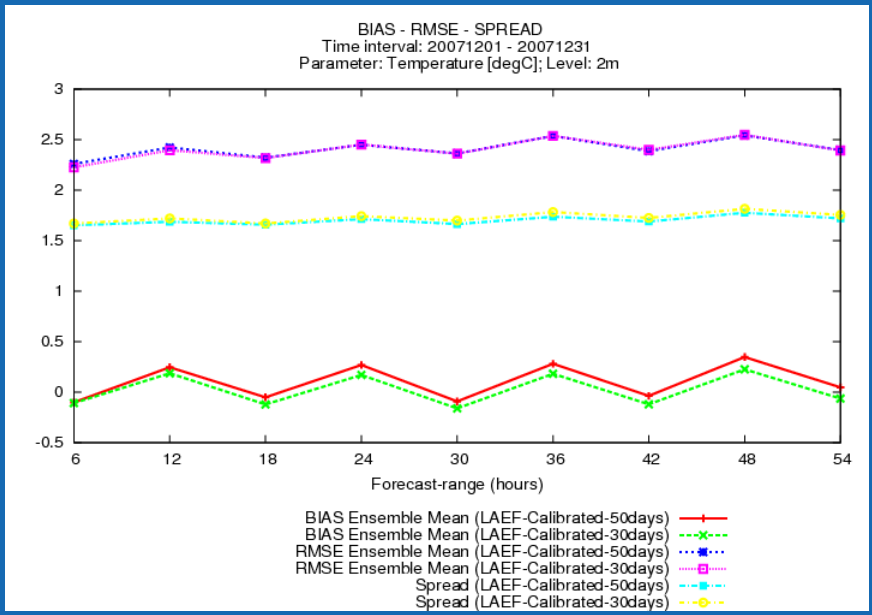


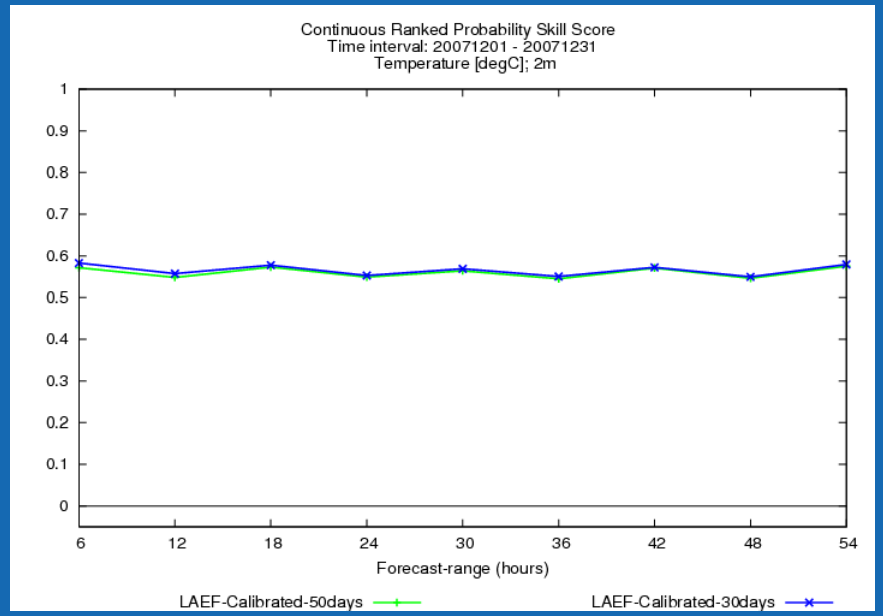
Bias, RMSE of Ensemble Mean and Ensemble Spread as a function of lead time for 2m temperature of bias-corrected EPS with w=2% and bias-corrected EPS with w=10%. Bias correction with higher weighting leads to a slight reduction of RMSE, the mean Bias rather remains.



CRPSS (with deterministic Aladin-Austria as a reference) for T2M as a function of lead time. Bias-corrected LAEF with w=2% (green) and bias-corrected LAEF with w=10% (blue). Higher weighting seems to improve skill in terms of CRPSS slightly.

Impact of training size on calibration

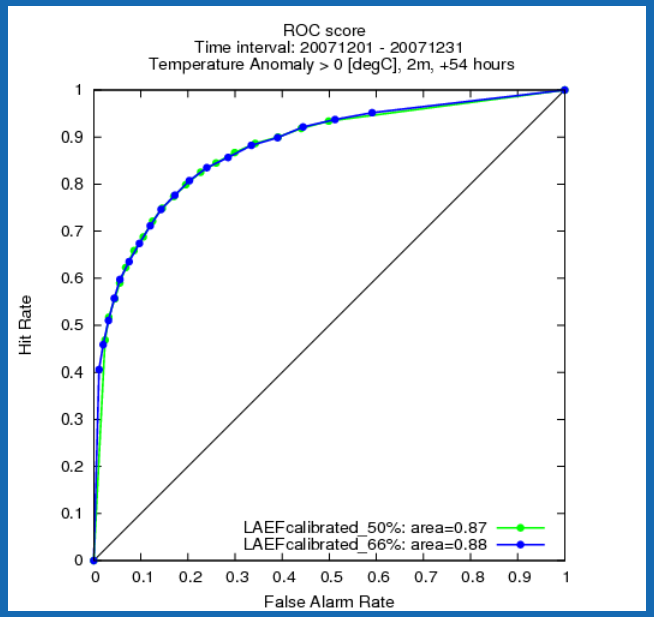




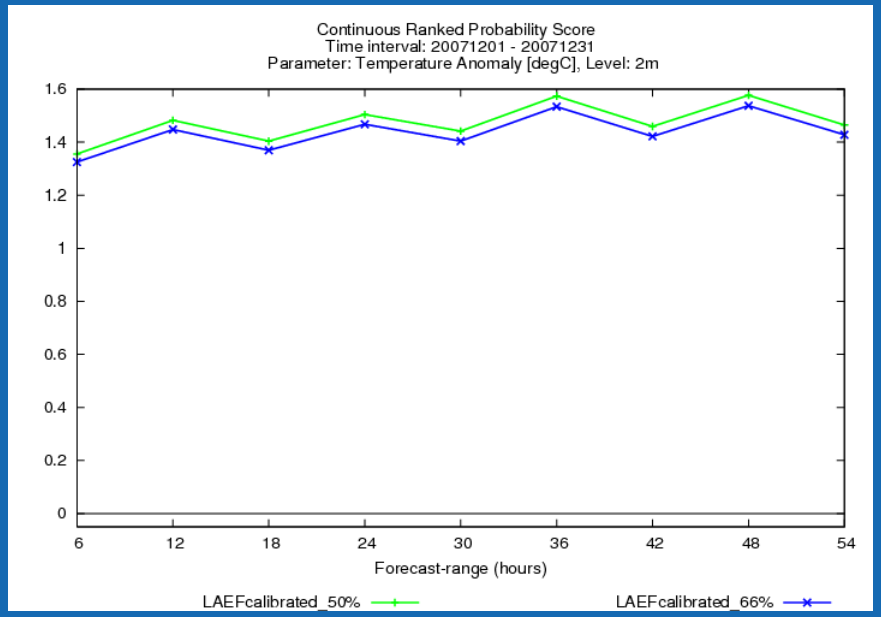
Bias, RMSE of Ensemble Mean and Ensemble Spread as a function of lead time for 2m temperature of calibrated EPS with 50 training days and calibrated EPS with 30 training days. Differences are marginal, possibly due to stable synoptical situation during December?

CRPSS (with deterministic Aladin-Austria as a reference) for T2M as a function of lead time. Calibrated LAEF with 50 training days (green) and calibrated LAEF with 30 training days (blue). Again, no significant impact on scores from using larger sample.

Impact of spread re-scaling on EPS performance

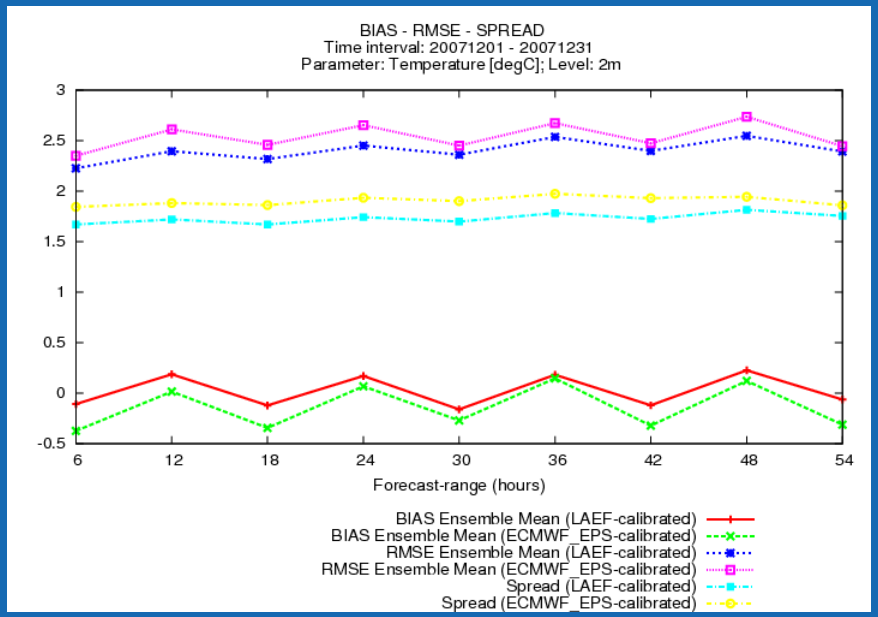


ROC curve and area under the ROC curve for T2M. Calibrated LAEF rescaled with f_{resc}=0.5 (green) and calibrated LAEF with f_{resc}=2/3 (blue). Both behave very similar, slightly better in case of f_{resc}=2/3.

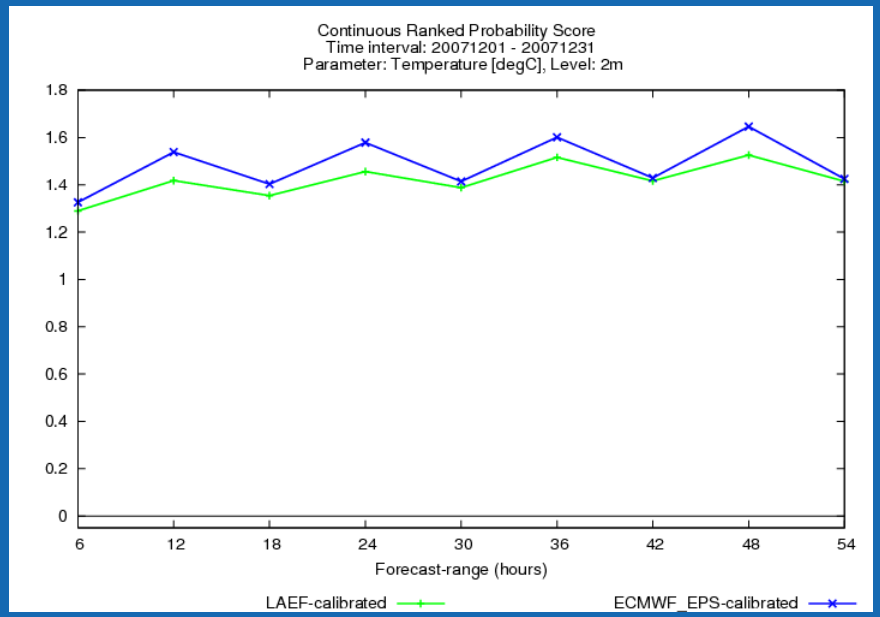


CRPS for T2M as a function of lead time. Calibrated LAEF rescaled with f_{resc}=0.5 (green) and calibrated LAEF with f_{resc}=2/3 (blue). Both scores are very similar, but slightly reduced CRPS using f_{resc}=2/3

ECMWF calibration vs. LAEF calibration



Bias, RMSE of Ensemble Mean and Ensemble Spread as a function of lead time for 2m temperature of calibrated LAEF and calibrated ECMWF-EPS. For ECMWF, a slight bias remains, both spread and RMSE are higher.



CRPS for T2M as a function of lead time. Calibrated LAEF (green), calibrated ECMWF-EPS (blue). LAEF performs slightly better, although differences are not very overwhelming.

Summary

- The state-of-the-art of LAM-EPS systems makes calibration inevitable!
- Methods that are successfully applied to global EPS systems are suitable for LAM-EPS (at least NGR method for Gaussian distributed variables).
- Calibration of 2m temperature using NGR leads to substantial improvement of probabilistic forecast.
- About 50% of the total improvement can be attributed to bias correction.
- Rank histograms of the calibrated LAEF are much flatter than of raw LAEF, but still remain slightly underdispersive.
- Bias correction: Higher weights on recent errors seem to be more appropriate to short range forecasting.
- The impact of using 50 days training data in contrast to 30 days is negligible for 2m temperature, at least for the chosen month December 2007 (we have to check if this is also the case during other seasons!)
- Calibrated probability forecasts from LAEF still perform slightly better than calibrated ECWMF-EPS forecasts (➔ statistical downscaling does not render dynamical downscaling unnecessary)

Open questions, outlook and on-going activities

1. Calibration for other variables (e.g. windspeed, mean sea level pressure)
2. Operational implementation of calibration for 2m temperature on high resolution (using INCA analysis)
3. Other methods for non-Gaussian distributed variables (e.g. logistic regression or analog technique for precipitation).
4. Still many open questions, e.g. how to deal with model changes? (no hindcast data available!)

THANKS FOR ATTENTION!