Calibration of ALADIN EPS precipitation forecasts

Lovro Kalin

Meteorological and hydrological service, Croatia

Acknowledgments

- results of efforts made on Aladin stay in April 2008 in ZAMG (Vienna)
- Work supervised by Yong Wang and Alexander Kann
- Blazenka Matjacic (DHMZ)

Outline

- Ensemble forecasts – calibration
- Calibration techniques: Logistic regression
- Aladin EPS and data
- Results
- Conclusion and future work

Ensemble forecasts – calibration

- why calibrate?
- ensemble members will inherit the deficiencies from the deterministic forecast
- if the deterministic model is biased, all individual ensemble members will probably exhibit the same bias, etc.

Calibration techniques

- various calibration techniques proposed:
- bias correction, Bayesian model averaging, analog method, logistic regression etc...
- for the 2m-temperature, nonhomogeneous Gaussian regression has been applied, with dramatic improvement (A. Kann, 2008)

Precipitation calibration

- Calibration of precipitation, due to it's complex temporal and spatial distribution, seems to be more complicated
- based on efforts mostly made by Tom Hamill from NOAA Earth System Research Laboratory, we applied a logistic regression to a small set of Aladin EPS forecasts
- Hamill, T. M., R. Hagedorn, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Wea. Rev.*

Logistic regression

- P - probability that precipitation O (predictand) will reach certain threshold T

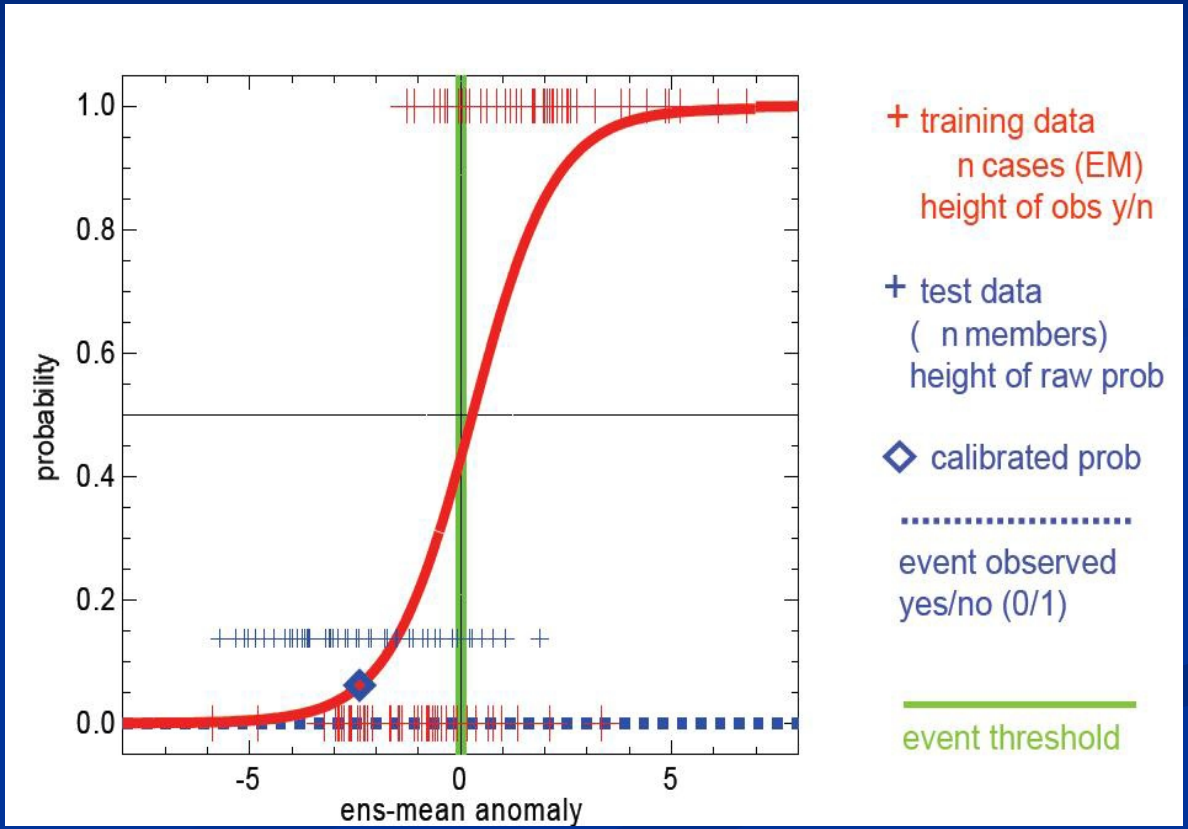
$$P(O > T) = 1.0 - \frac{1.0}{1.0 + \exp\left\{\beta_0 + \beta_1\left(\bar{x}^f\right) + \beta_2\left(\sigma^f\right)\right\}}.$$

- predictors:

  x – ensemble mean

  σ - spread of ensemble (RMSE)

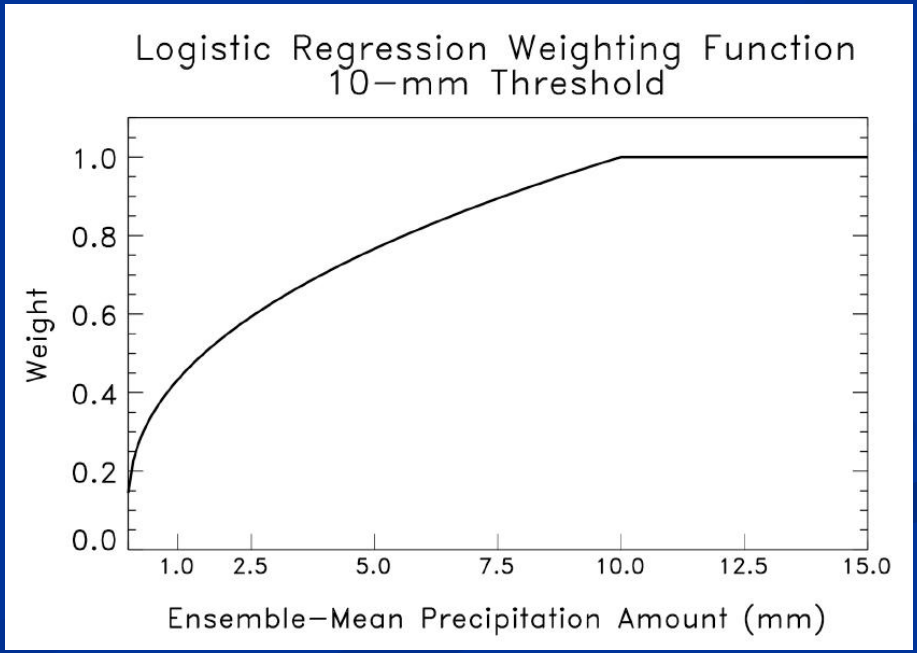
- β0, β1 and β2 are coefficients of regression, that have to be obtained



Logistic regression (cont...)

- a weighting function has also been proposed, in order to give more significance to cases with larger precipitation forecasts
- choice of the function is quite arbitrary
- implications of its implementation still have to be investigated...

$$w = \begin{cases} 1.0 & \text{if} \quad \overline{x}^f + 0.01 > T \\ 0.1 + 0.9 \times \exp\left(-1.0 \times \left|\log_{10}\left(\overline{x}^f + 0.01\right) - \log_{10}(T)\right|\right) & \text{if} \quad \overline{x}^f + 0.01 \le T \end{cases}$$



Logistic regression (cont...)

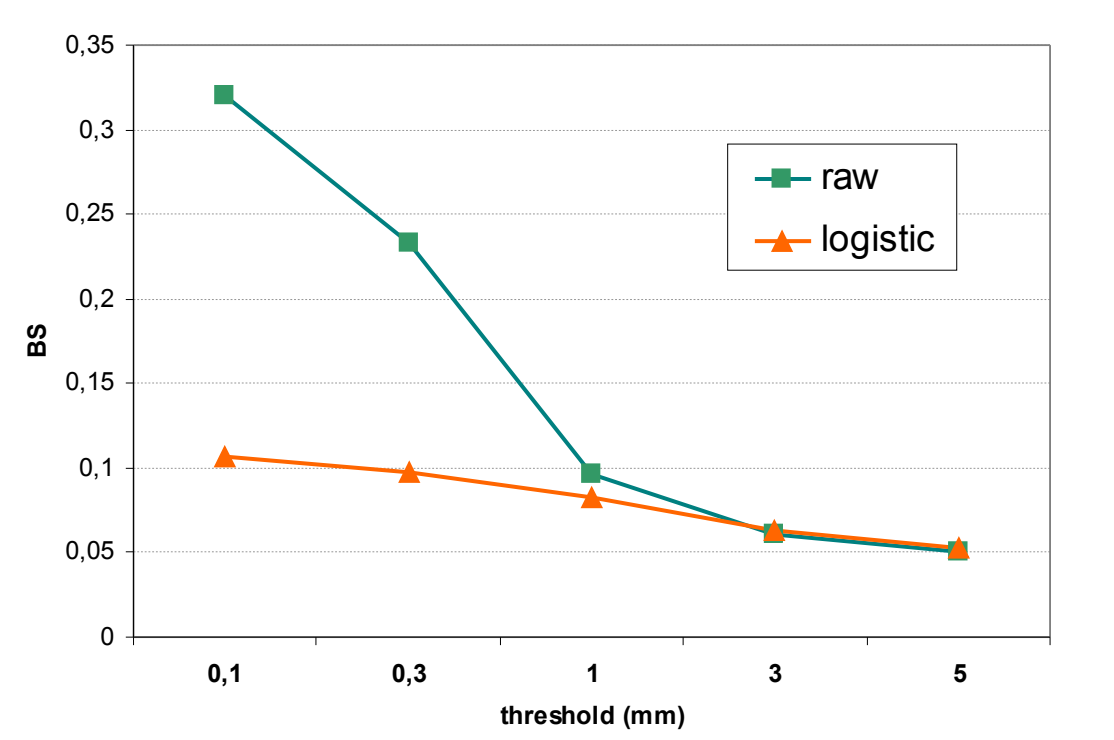
- such a fitting does not provide the full probability density function, but only the calibrated probability that a certain threshold will be exceeded
- fitting has to be done for each desired threshold

Aladin EPS and data

- at ZAMG, a set of Aladin EPS forecast data is stored, starting from June 2007
- a small seasonal sub-sample has been considered, covering summer of 2007 (June – September)
- 6-hour precipitation forecasts, different lead times
- ~ 100 events

Results

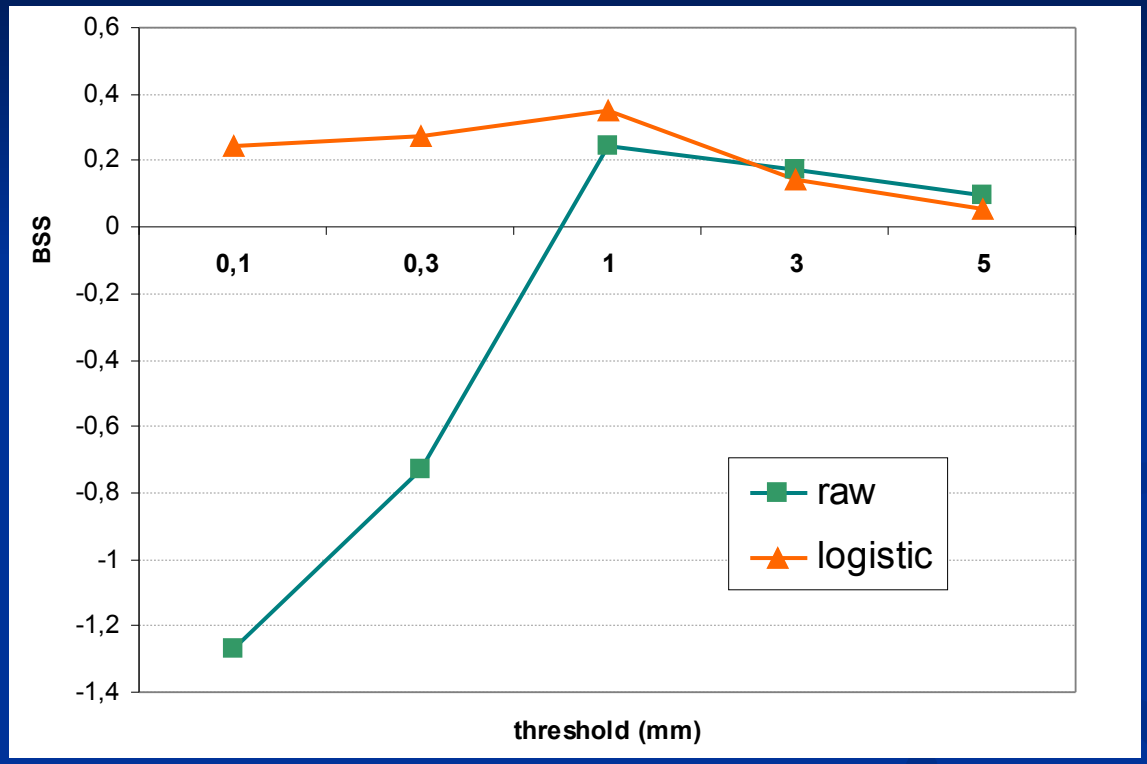
Brier score for 6-hour precipitation (12 - 18 UTC) for Zagreb, Jun - Sep 2007



- Improvement significant only for smaller thresholds

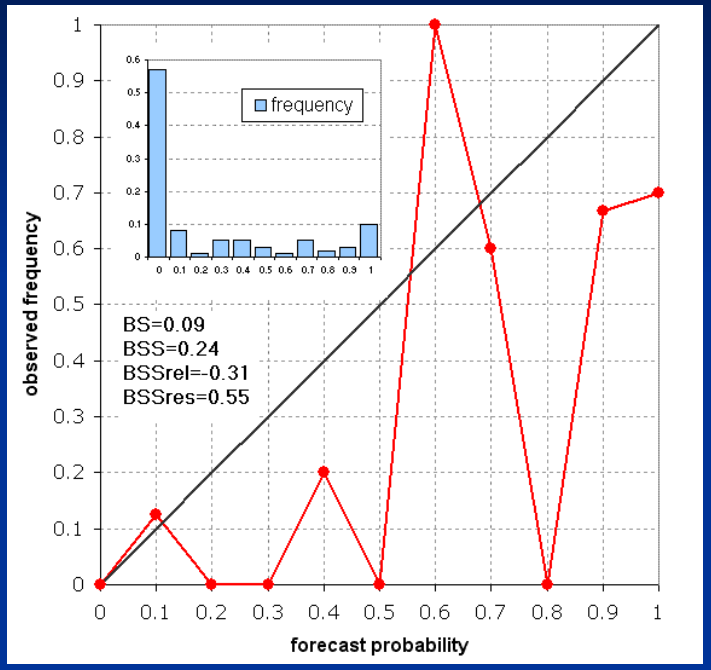
Results

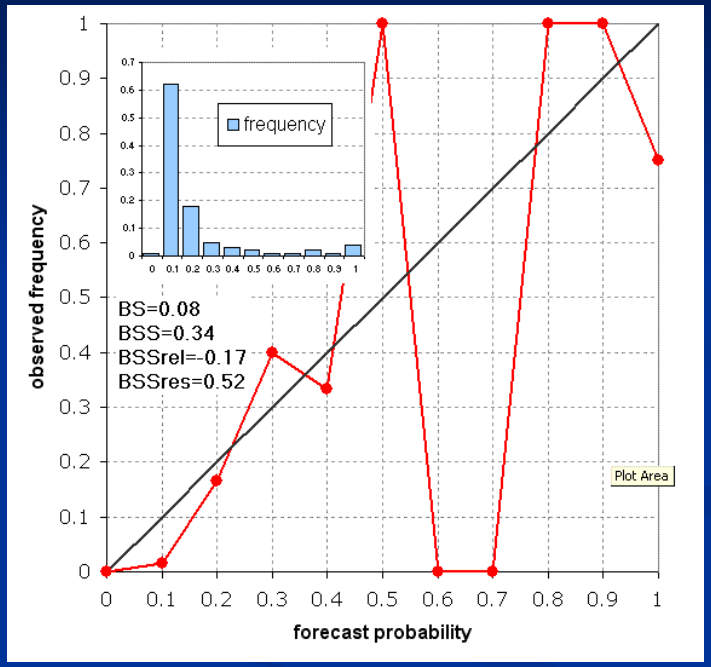
Brier skill score for 6-hour precipitation (12 - 18 UTC) for Zagreb, Jun - Sep 2007_



- skill improved only for smaller thresholds

Reliability plots...

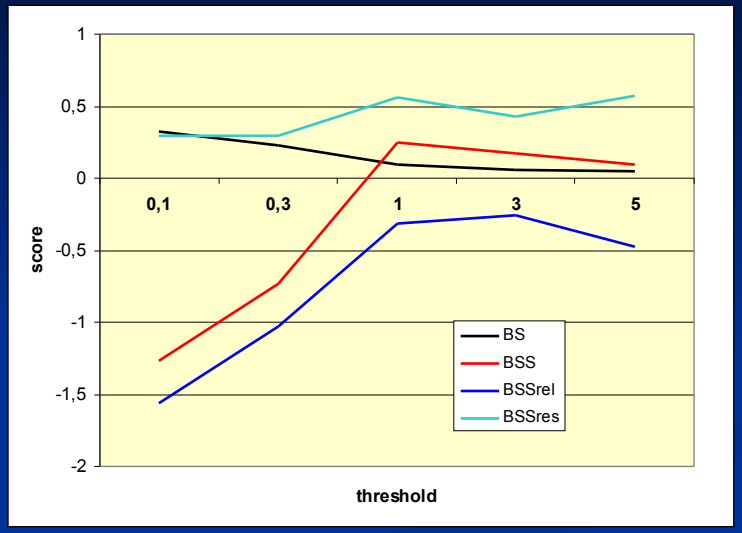


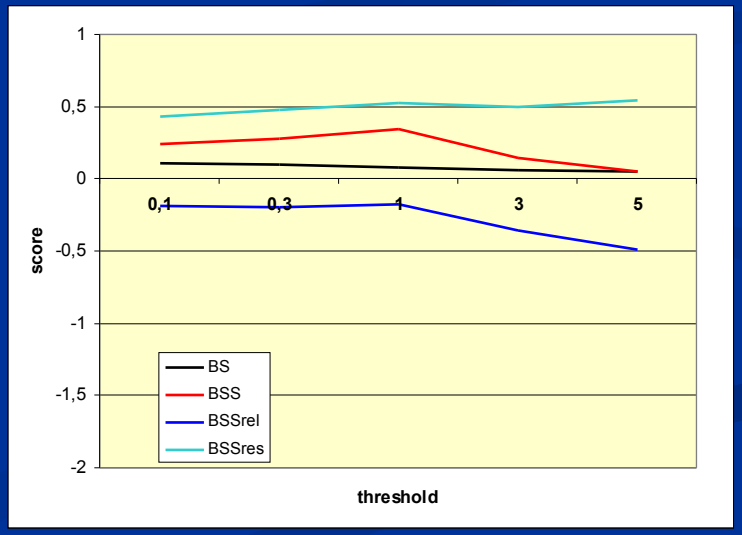


- calibration reduces sharpness of the forecasts (distribution less U-shaped than the raw)
- forecasts up to 40% (most frequent) are much better calibrated (red line closer to the diagonal)

Decomposition of Brier score...

- improvement of skill (red line) is more influenced by improvement of the reliability term (blue line) rather than improvement of the resolution term (light blue line)
- calibrated forecasts are not much sharper, but they are more accurate



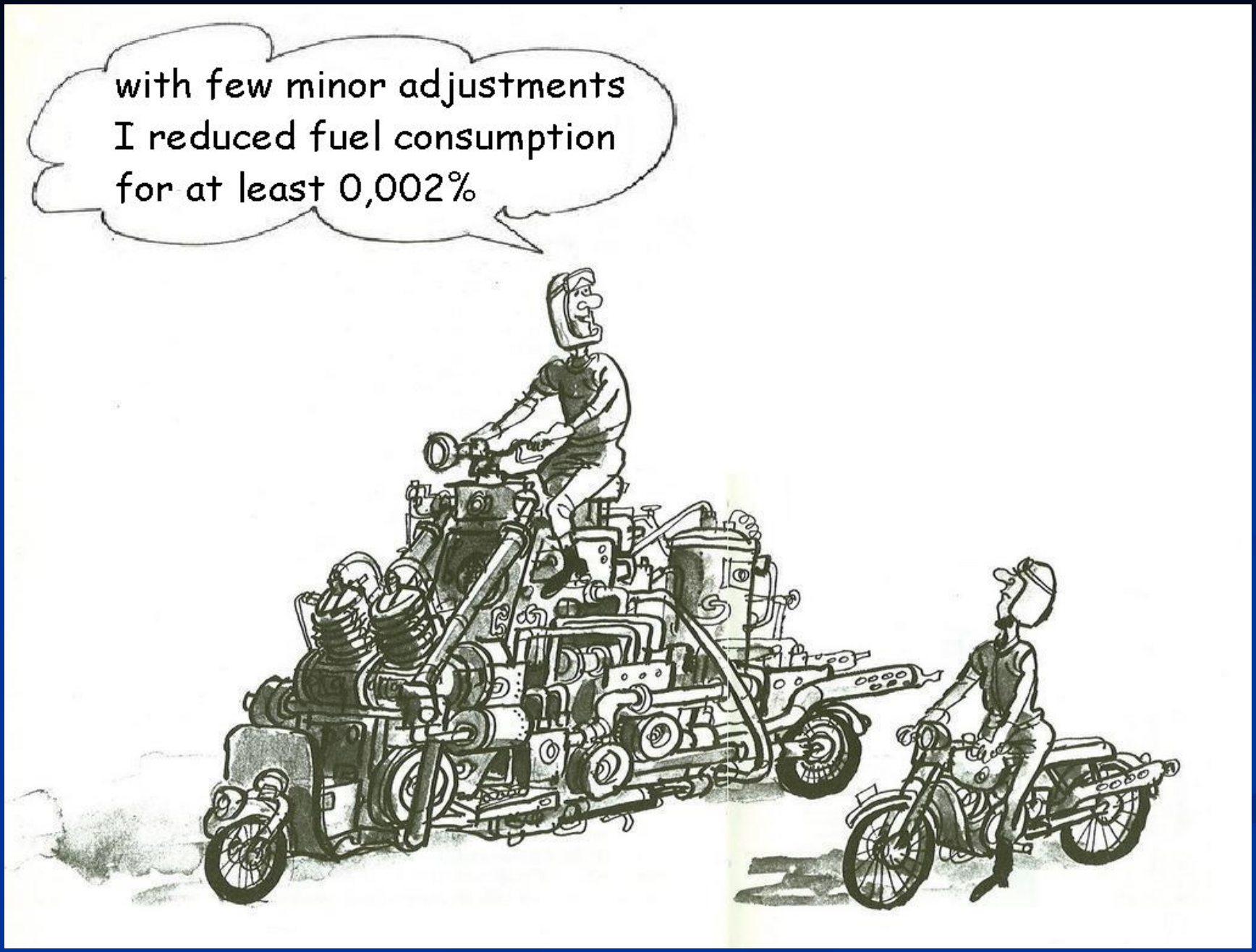


Conclusion...

- improvement is significant only for smaller thresholds (less than 1mm). According to Tom Hamill's paper, this is mostly due to relatively small sample
- Therefore, future work should be focused mostly on the impact of the sample size to calibration results (including more seasons, clustering different stations etc.)
- We assume that the increase of the sample could give benefit for the bigger thresholds also

Conclusion…

- significant daily variation
- no significant impact of forecast range (between D+1 and D+2)
- Impact of the weighting function is still not very clear
- other techniques (analog, BMA…) could also be applied and compared to logistic regression


with few minor adjustments
I reduced fuel consumption
for at least 0,002%