Post-Processing of Ensemble Forecasts

Tim Stockdale / Renate Hagedorn

European Centre for Medium-range Weather Forecasts

t.stockdale@ecmwf.int

Outline

- Motivation
- Methods
- Training data sets
- Results

This lecture is focussed on application to medium-range forecasts, but the theory and methods are general.

It is only an introductory lecture: some students may already be working with more advanced methods than those described

Motivation

- Raw, uncalibrated ensemble forecasts contain forecast bias and dispersion errors
- The goal of calibration is to correct for such deficiencies, i.e. to construct predictions with statistical properties similar to the observations
- A number of statistical methods exist for post-processing ensembles
- Calibration needs a record of prediction-observation pairs
  - In the (distant) past, these might come from e.g. the previous 2 months of operational forecasts
  - Nowadays, make use of large re-forecast sets covering many previous years, to allow a much more accurate calibration
  - "Observations" might be weather station data, or gridded global analyses
- Calibration of point forecasts is particularly successful at locations with long historical data records
- Calibration is often a form of downscaling

Calibration methods

- Bias correction
- Multiple implementation of deterministic MOS
- Ensemble dressing
- Bayesian model averaging
- Non-homogenous Gaussian regression
- Logistic regression
- Analogue method

Bias correction

- As a simple first order calibration a bias correction can be applied:

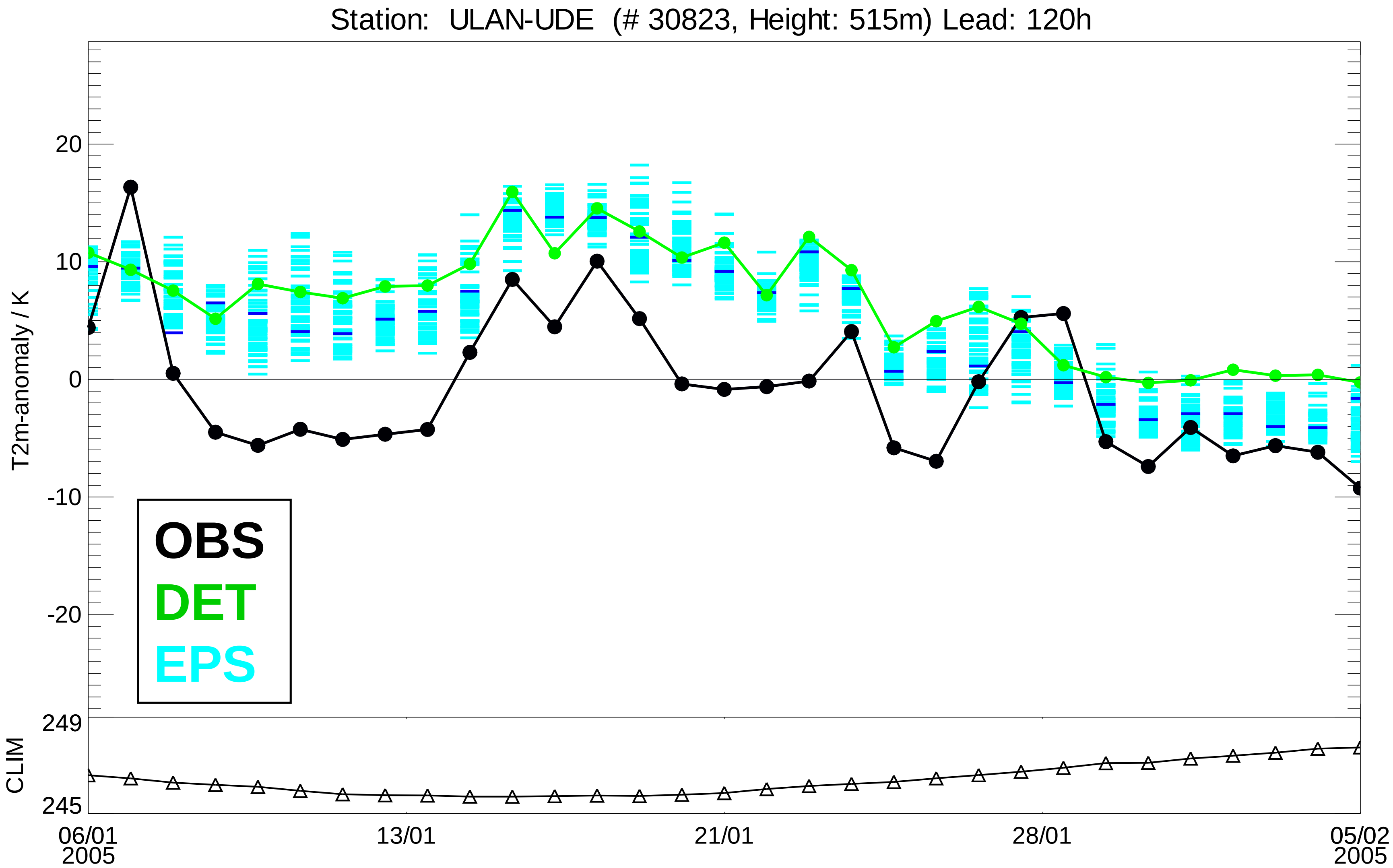
$$c = -\frac{1}{N}\sum_{i=1}^{N}\bar{e}_i + \frac{1}{N}\sum_{i=1}^{N} o_i$$

with: $\bar{\boldsymbol{e}}_{\boldsymbol{i}}$ = ensemble mean of the ith forecast
$\boldsymbol{o}_{\boldsymbol{i}}$ = value of ith observation
N = number of observation-forecast pairs

- This correction is added to each ensemble member, i.e. spread is not affected

- Particularly useful/successful at locations with features not resolved by model and causing significant bias

Bias correction

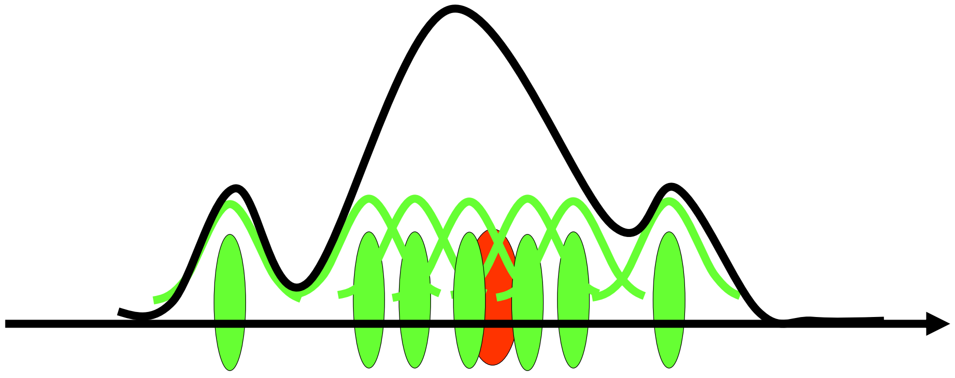


Multiple implementation of deterministic MOS

- A possible approach for calibrating ensemble predictions is to simply correct each individual ensemble member according to its deterministic model output statistic (MOS)
- **BUT**: this approach is conceptually inappropriate since for longer lead-times the MOS tends to correct towards climatology
  - all ensemble members tend towards climatology with longer lead-times
  - decreased spread with longer lead-times
  - in contradiction to increasing uncertainty with increasing lead-times
- (Discontinued) experimental product at http://www.nws.noaa.gov/mdl/synop/enstxt.php

Ensemble dressing

- Define a probability distribution around each ensemble member ("dressing")



- A number of methods exist to find appropriate dressing kernel ("best-member" dressing, "error" dressing, "second moment constraint" dressing, etc.)
- Average the resulting n_{ens} distributions to obtain final pdf

Ensemble Dressing

- (Gaussian) ensemble dressing calculates the forecast probability for the quantiles q as:

$$P(v \leq q) = \frac{1}{n_{ens}} \sum_{i=1}^{n_{ens}} \Phi\left[\frac{q - \tilde{x}_i}{\sigma_D}\right]$$

with: $\boldsymbol{\Phi}$ = CDF of standard Gaussian distribution
$\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$ = bias-corrected ensemble-member

- Key parameter is the standard deviation of the Gaussian dressing kernel
- One simple approach: "best member" dressing, take standard deviation from r.m.s. difference of (obs-best member) from training set.

Ensemble Dressing

- Common approach: second-moment constraint dressing

$$\sigma_D{}^2 = \sigma^2{}_{\bar{x}-y} - \left(1 + \frac{1}{n_{ens}}\right)\overline{\sigma}_{ens}{}^2$$

error variance of the ensemble-mean FC

average of the ensemble variances over the training data

•BUT: this can give negative or unstable variances, if model is already near to or over-dispersive.

•Ensemble dressing to generate a pdf is only suitable for *under-dispersive* forecasts.

Bayesian Model Averaging

- BMA closely linked to ensemble dressing
- Differences:
  - dressing kernels do not need to be the same for all ensemble members
  - different estimation method for kernels
- Useful for giving different ensemble members (models) different weights:

$$P(v \leq q) = w_1 \Phi\left[\frac{q - \tilde{x}_1}{\sigma_1}\right] + w_e \sum_{j=2}^{n_{ens}} \Phi\left[\frac{q - \tilde{x}_j}{\sigma_e}\right]$$

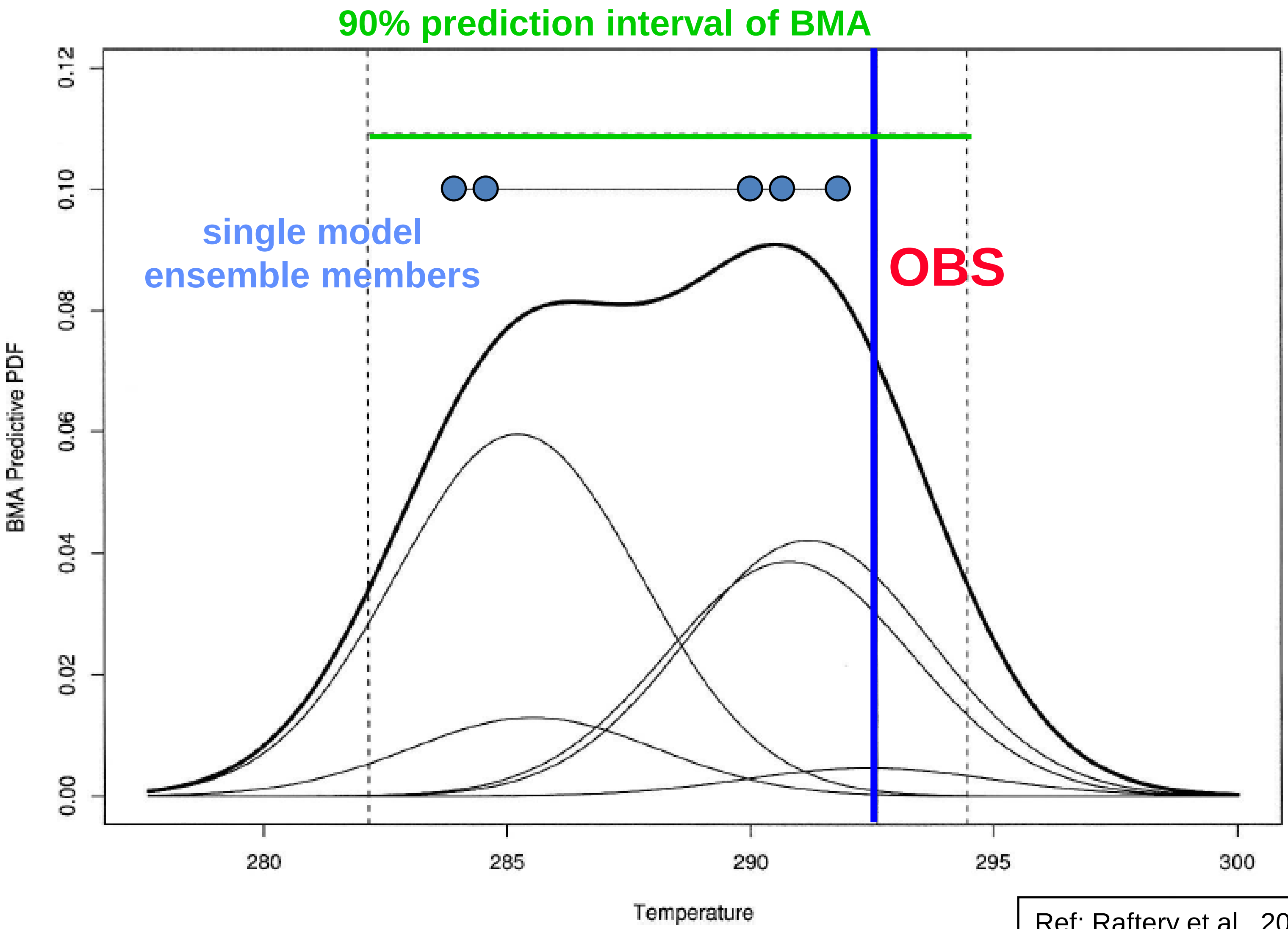
with: $w_1 + w_e(n_{ens} - 1) = 1$

- Estimation of weights and kernels simultaneously via maximum likelihood, i.e. maximizing the log-likelihood function:

$$\ln(\Lambda) = -\sum_{i=1}^{N} \ln\left[w_1 g_1(v_i | \tilde{x}_{1,i}, \sigma_1^{\,2}) + w_e \sum_{j=2}^{n_{ens}} g_e(v_i | \tilde{x}_{j,i}, \sigma_e^{\,2})\right]$$

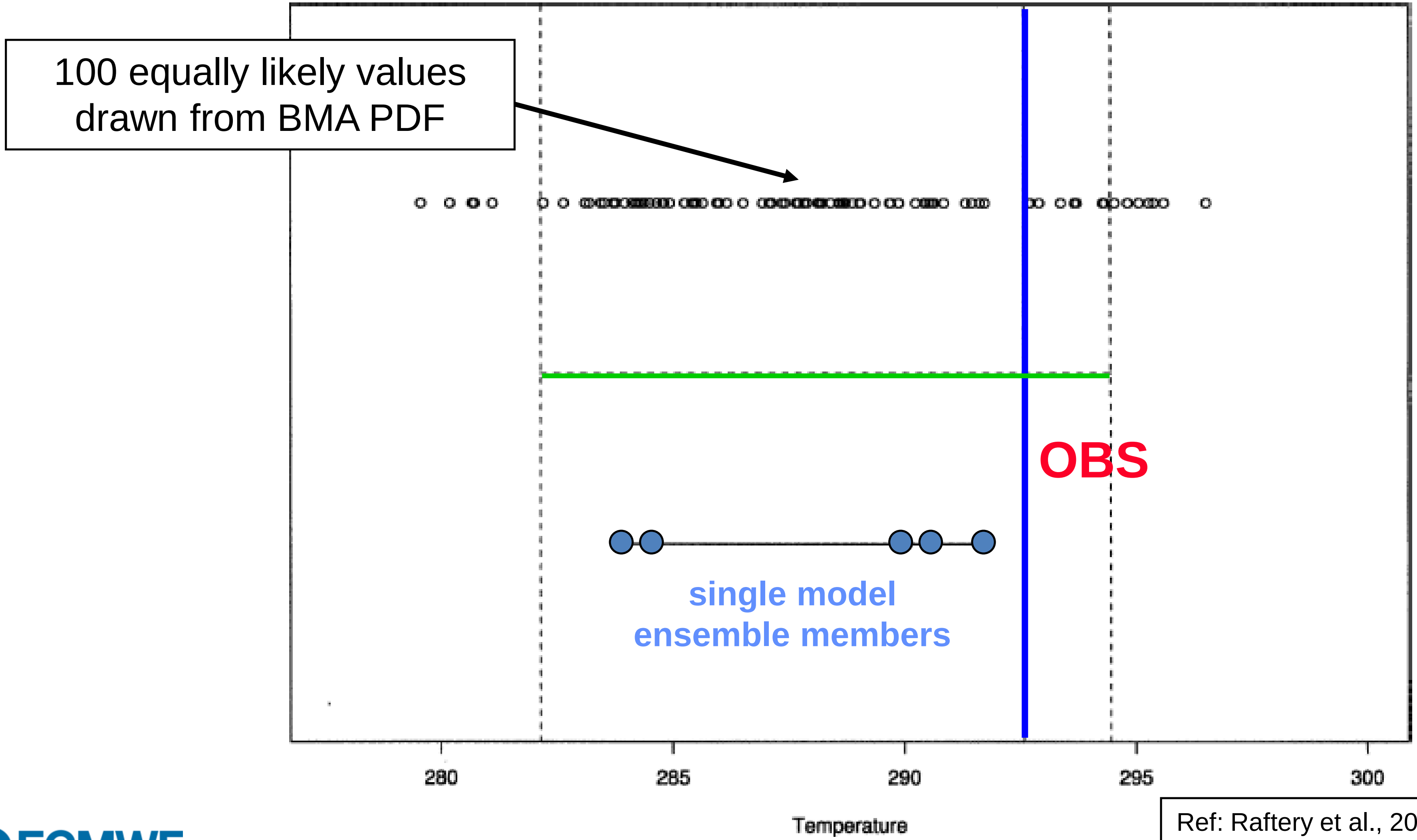
g_1, g_e = Gaussian PDF's

BMA: example



Ref: Raftery et al., 2005, MWR

BMA: recovered ensemble members



Ref: Raftery et al., 2005, MWR

Non-homogenous Gaussian Regression

- In order to account for existing spread-skill relationships we model the variance of the error term as a function of the ensemble spread s_{ens}:

$$P(v \le q) = \Phi\left[\frac{q-(a+b\bar{x}_{ens})}{\sqrt{c+d{s_{ens}}^2}}\right]$$

- The parameters a,b,c,d are fit iteratively by minimizing the CRPS of the training data set

- Interpretation of parameters:
  - bias & general performance of ens-mean are reflected in a and b
  - large spread-skill relationship: $c \approx 0.0,\ d \approx 1.0$
  - small spread-skill relationship: $d \approx 0.0$

- Calibration provides mean and spread of Gaussian distribution

(called non-homogenous since variances of regression errors not the same for all values of the predictor, i.e. non-homogenous)

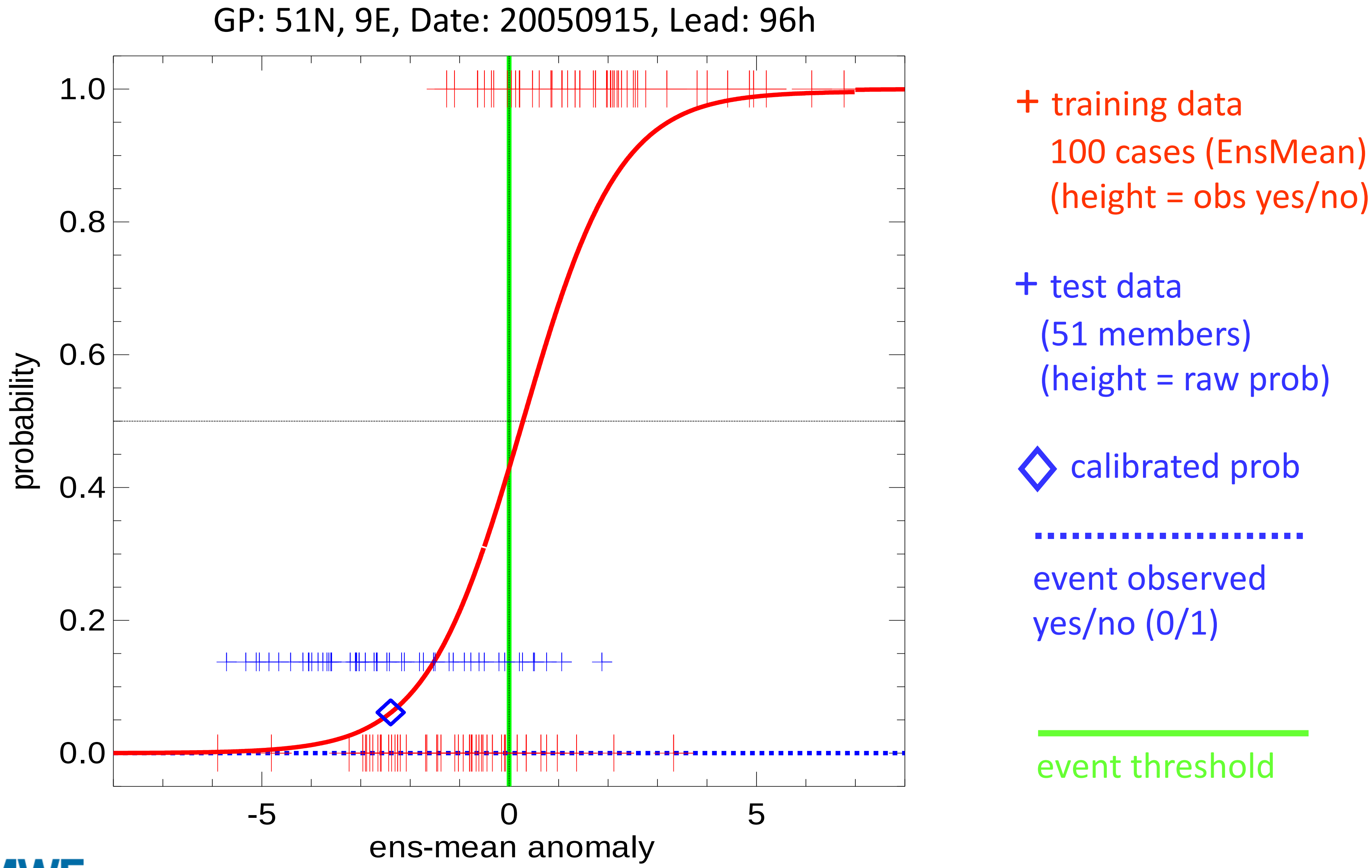
Logistic regression

- Logistic regression is a statistical regression model for Bernoulli-distributed dependent variables

$$P(v \le q) = \frac{\exp(\beta_0 + \beta_1 \bar{x}_{ens})}{1 + \exp(\beta_0 + \beta_1 \bar{x}_{ens})}$$

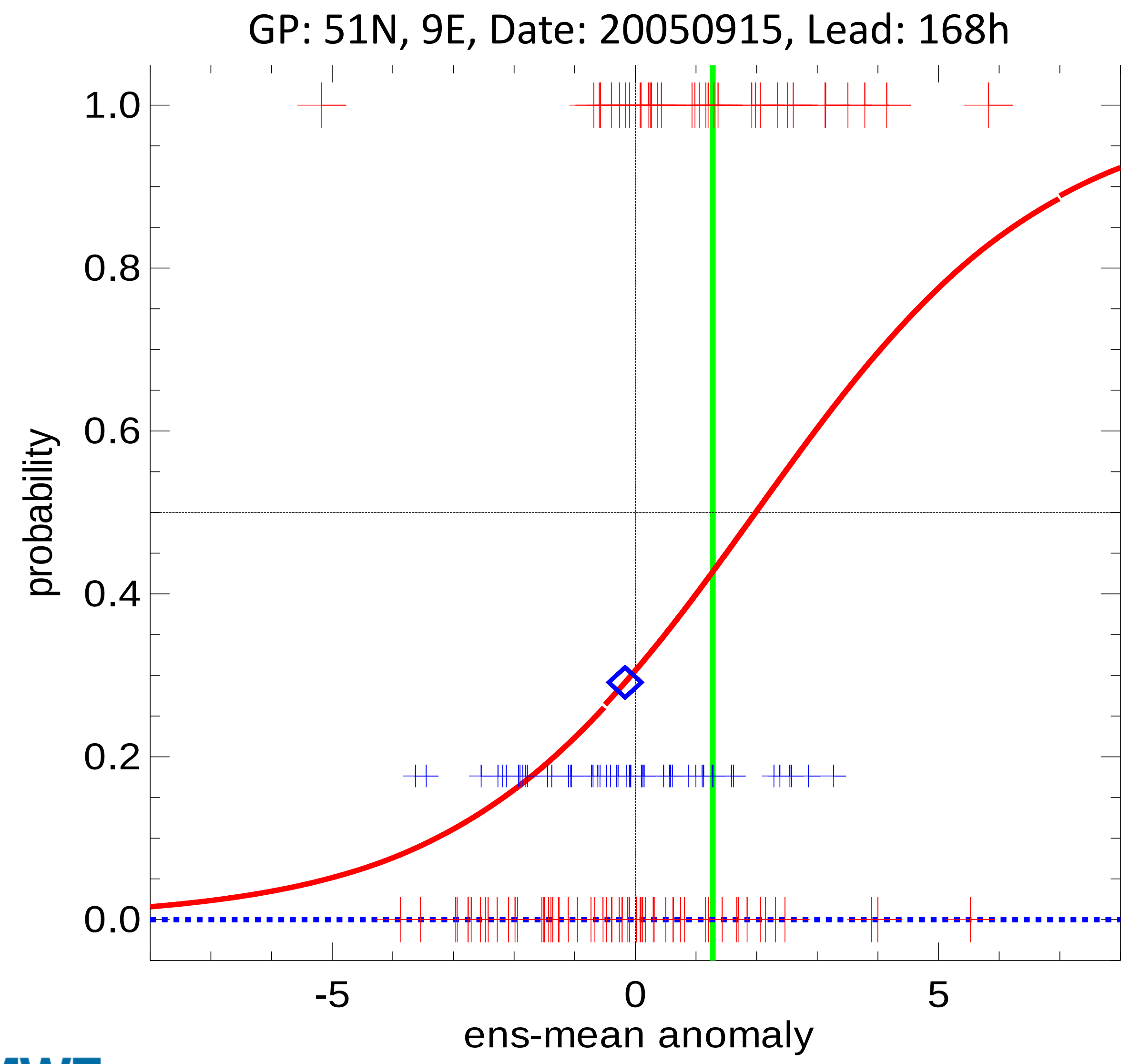
- P is bound by 0,1 and produces an s-shaped prediction curve
  - steepness of curve (β_1) increases with decreasing spread, leading to sharper forecasts (more frequent use of extreme probabilities)
  - parameter β_0 corrects for bias, i.e. shifts the s-shaped curve

How does logistic regression work?



Event did not happen in this case

Example: LR-Probability worse in this case



+ training data
100 cases (EM)
height of obs y/n

+ test data
(51 members)
(height = raw prob)

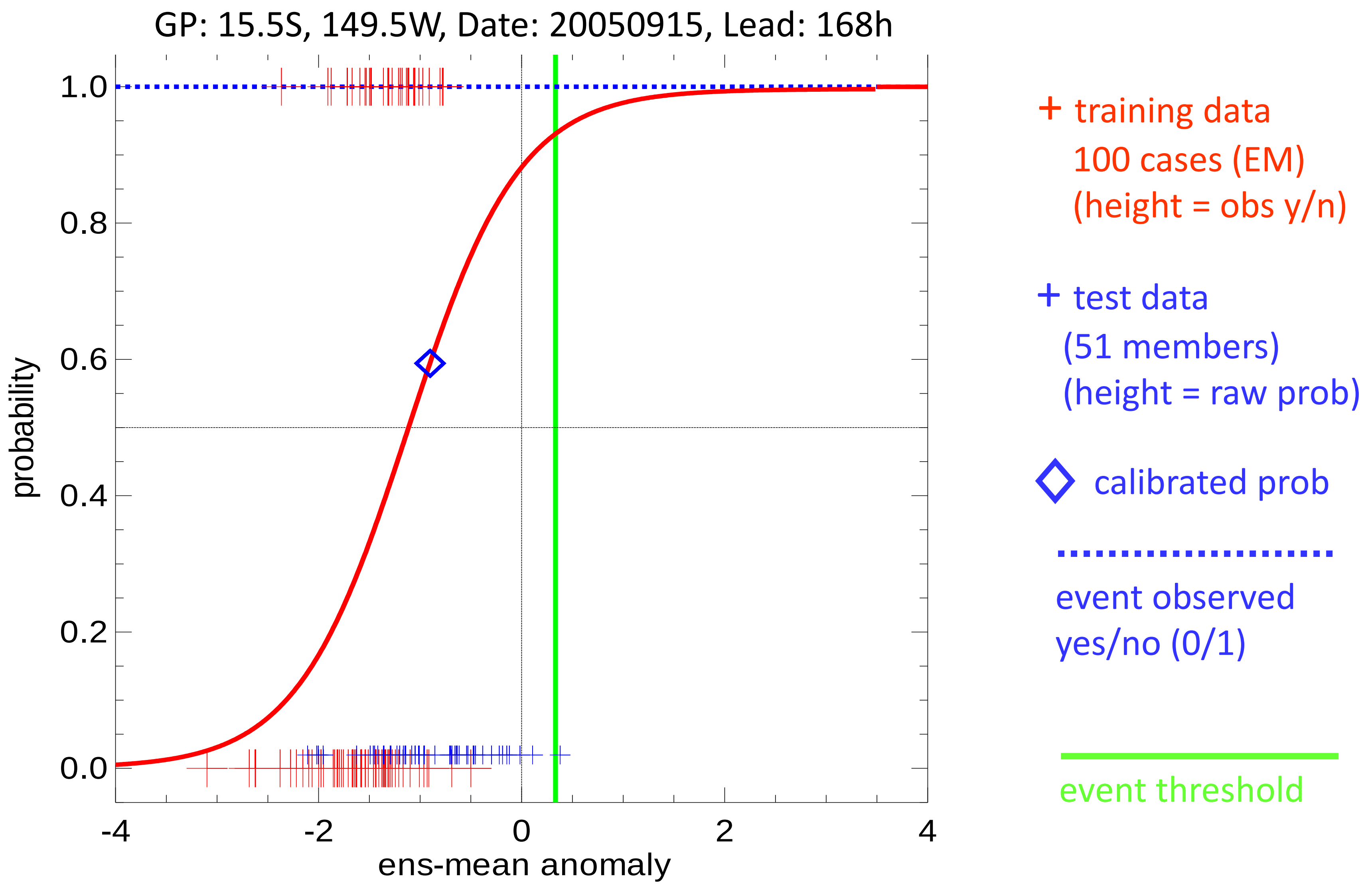
calibrated prob

event observed yes/no (0/1)

event threshold

Event did not happen

Example: LR-Probability (much) better!



Event **did** happen in this case

Analogue method

- Full analogue theory assumes a nearly infinite training sample
- Nonetheless, can be justified under simplifying assumptions:
  - Search only for local analogues
  - Match the ensemble-mean fields
  - Consider only one model forecast variable in selecting analogues
- General procedure:
  - Take the ensemble mean of the forecast to be calibrated and find the n_{ens} closest forecasts to this in the training dataset
  - Take the corresponding observations to these n_{ens} re-forecasts and form a new calibrated ensemble
  - Construct probability forecasts from this analogue ensemble

Analogue method

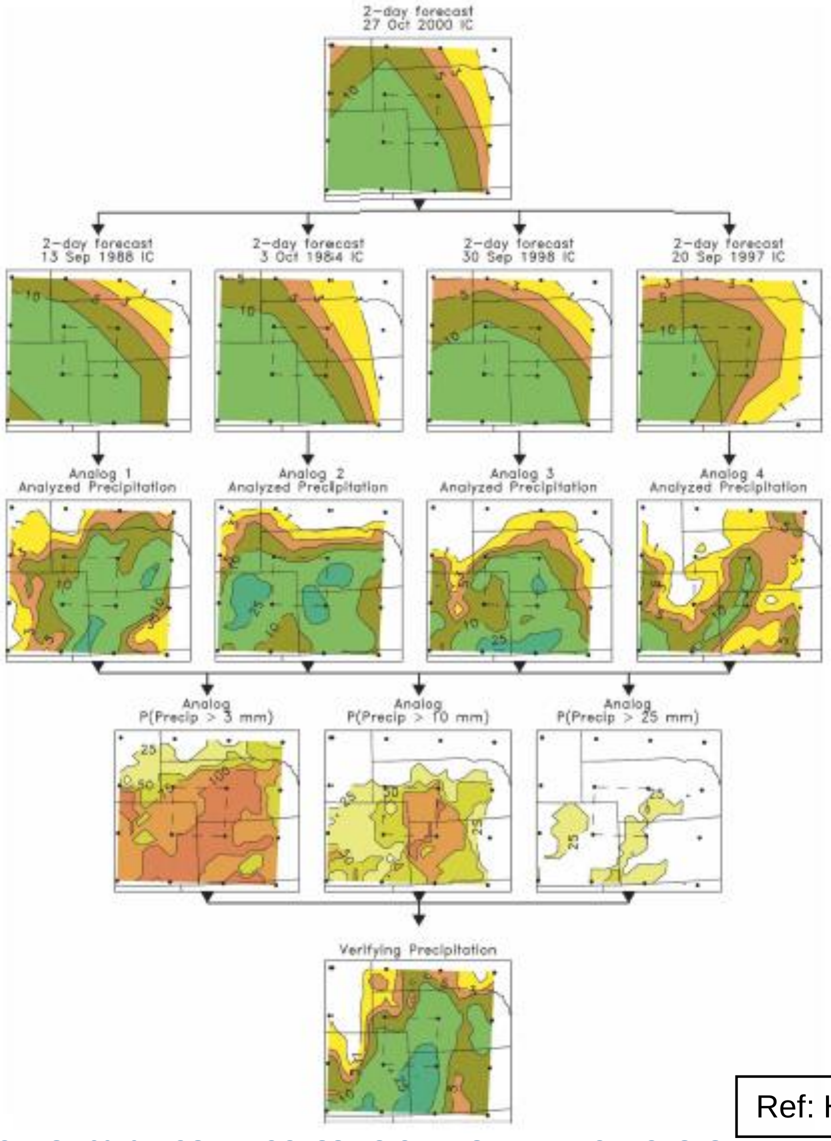
Forecast to be calibrated

Closest re-forecasts

Corresponding obs

Probabilities of analog-ens

Verifying observation



Ref: Hamill & Whitaker, 2006, MWR

Training datasets

- All calibration methods need a training dataset, containing a number of forecast-observation pairs from the past
  - The more training cases the better
  - The model version used to produce the training dataset should be as close as possible to the operational model version
- For research applications often only one dataset is used to develop and test the calibration method. In this case cross-validation has to be applied.
- For operational applications one can use:
  - Operational available forecasts from e.g. past 30-40 days
  - Data from a re-forecast dataset covering a larger number of past forecast dates / years

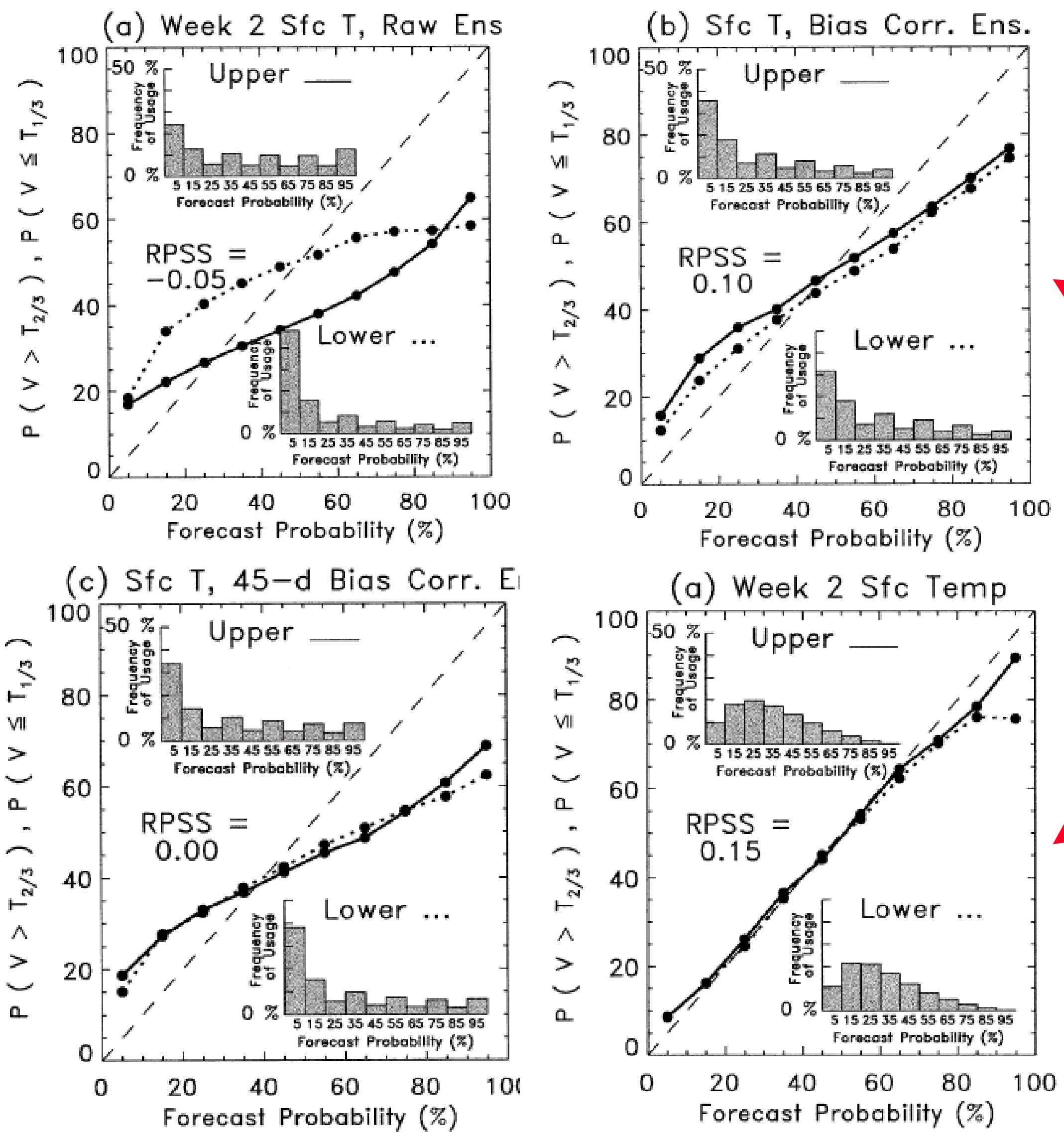
“Ideal” Reforecast Data Set

	2018																																		
	Feb		Mar																															Apr	
	27	28	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	01	02
1981																																			
1982																																			
1983																																			
1984																																			
1985																																			
. . . .																																			
2006																																			
2007																																			
2008																																			
2009																																			
2010																																			

Early motivating results from Hamill et al., 2004

Raw ensemble

Bias corrected with refc data



Achieved with “ideal” reforecast system!

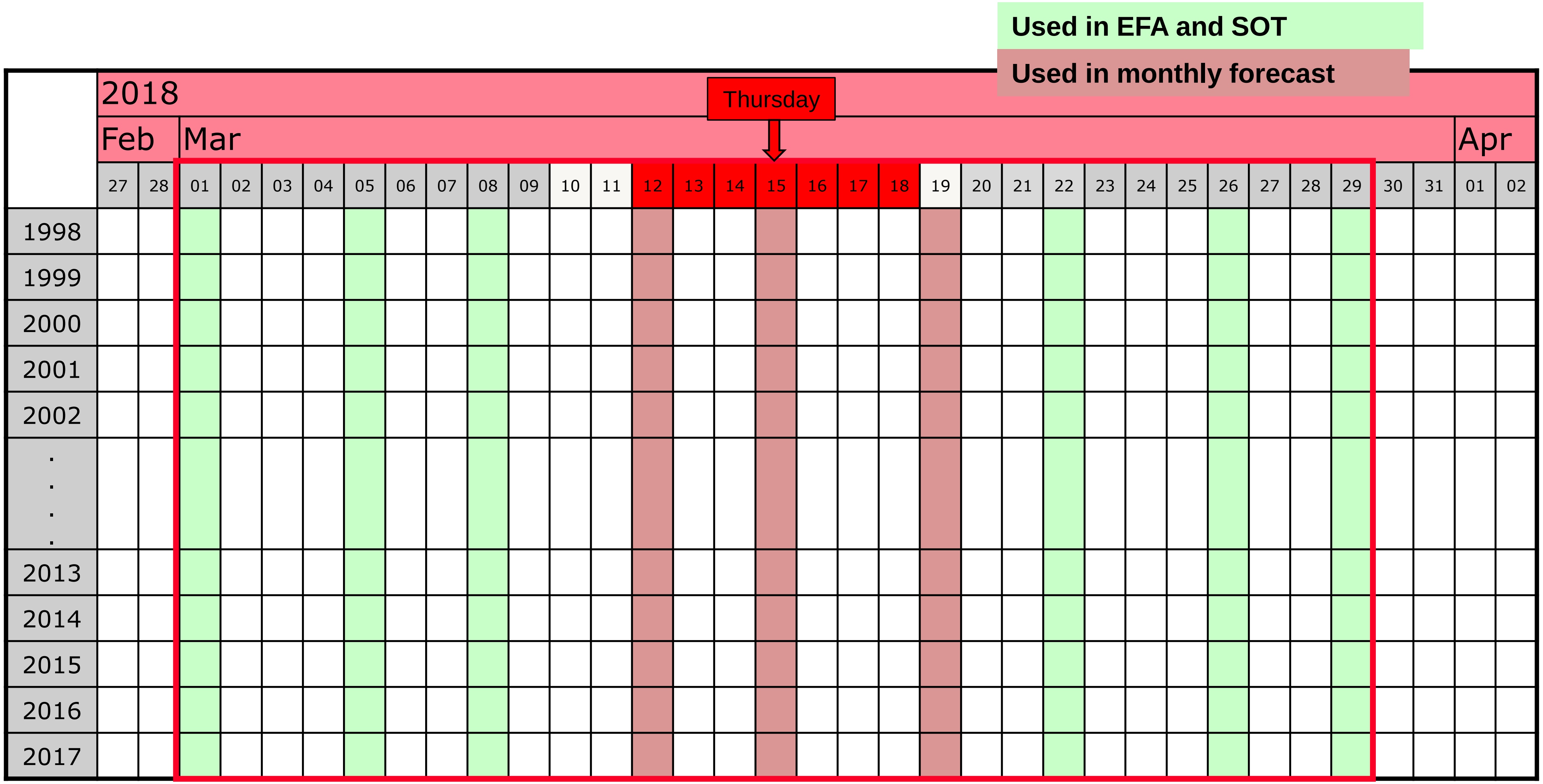
Bias corrected with 45-d data

LR-calibrated ensemble

The 45-day unified ENS ensemble system

- Unified ENS ensemble system enables the production of a unified reforecast data set, to be used by:
  - EFI model climate
  - 15 day ENS calibration
  - Monthly forecasts anomalies and verification
- Efficient use of resources (computational and operational)
- “Realistic” reforecast system has to be an optimal compromise between affordability and needs of all three applications
- Use 11 member ensemble, twice per week, for last 20 years

Unified ENS Reforecasts



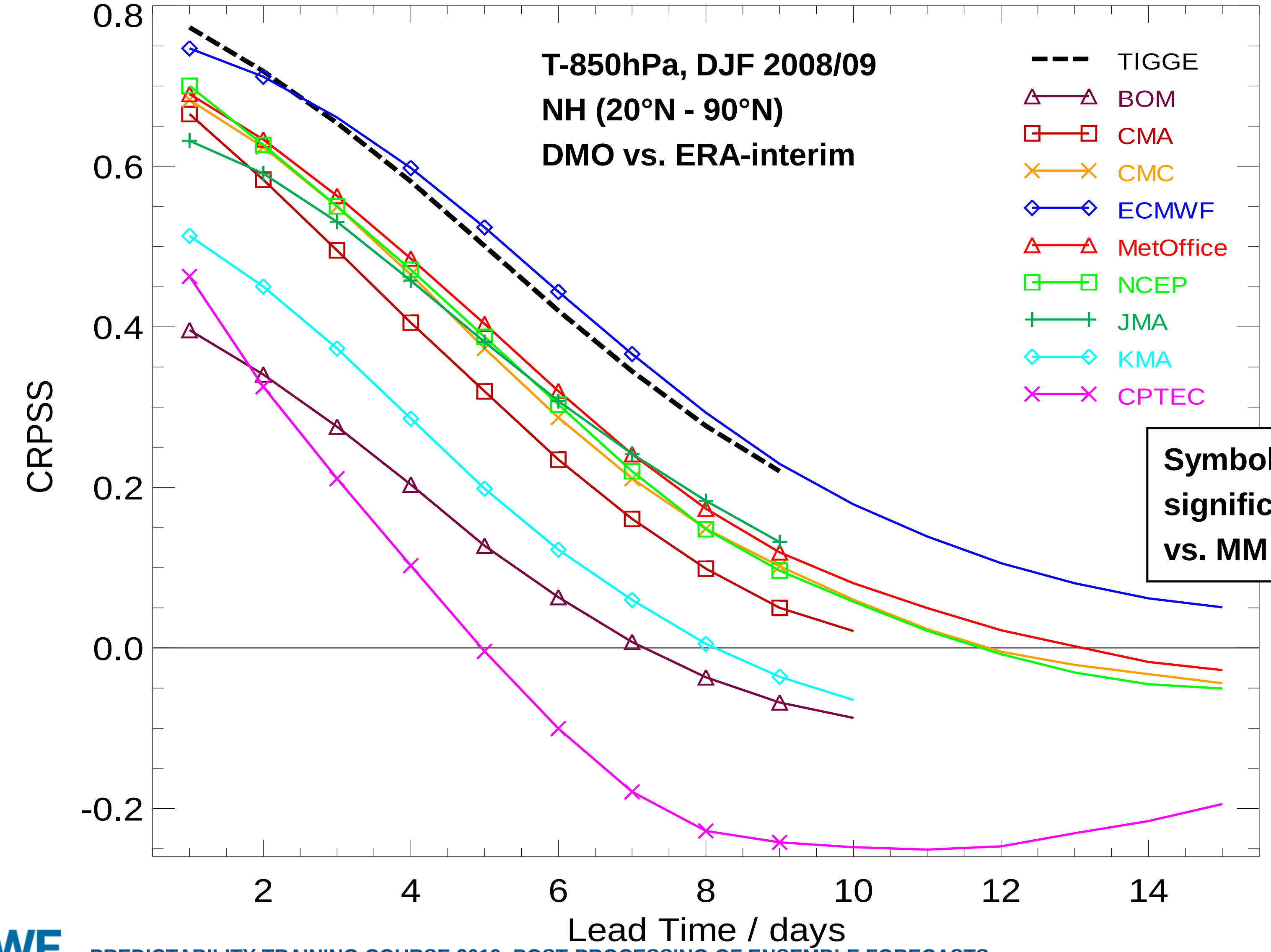
Testing the benefits of reforecast calibration

Reference: Hagedorn et al, 2012

- One goal of the TIGGE project is to investigate whether multi-model predictions are an improvement to single model forecasts
- The goal of using reforecasts to calibrate single model forecasts is to provide improved predictions
- Questions:
  - What are the relative benefits (costs) of both approaches?
  - What is the mechanism behind the improvements?
  - Which is the “better” approach?

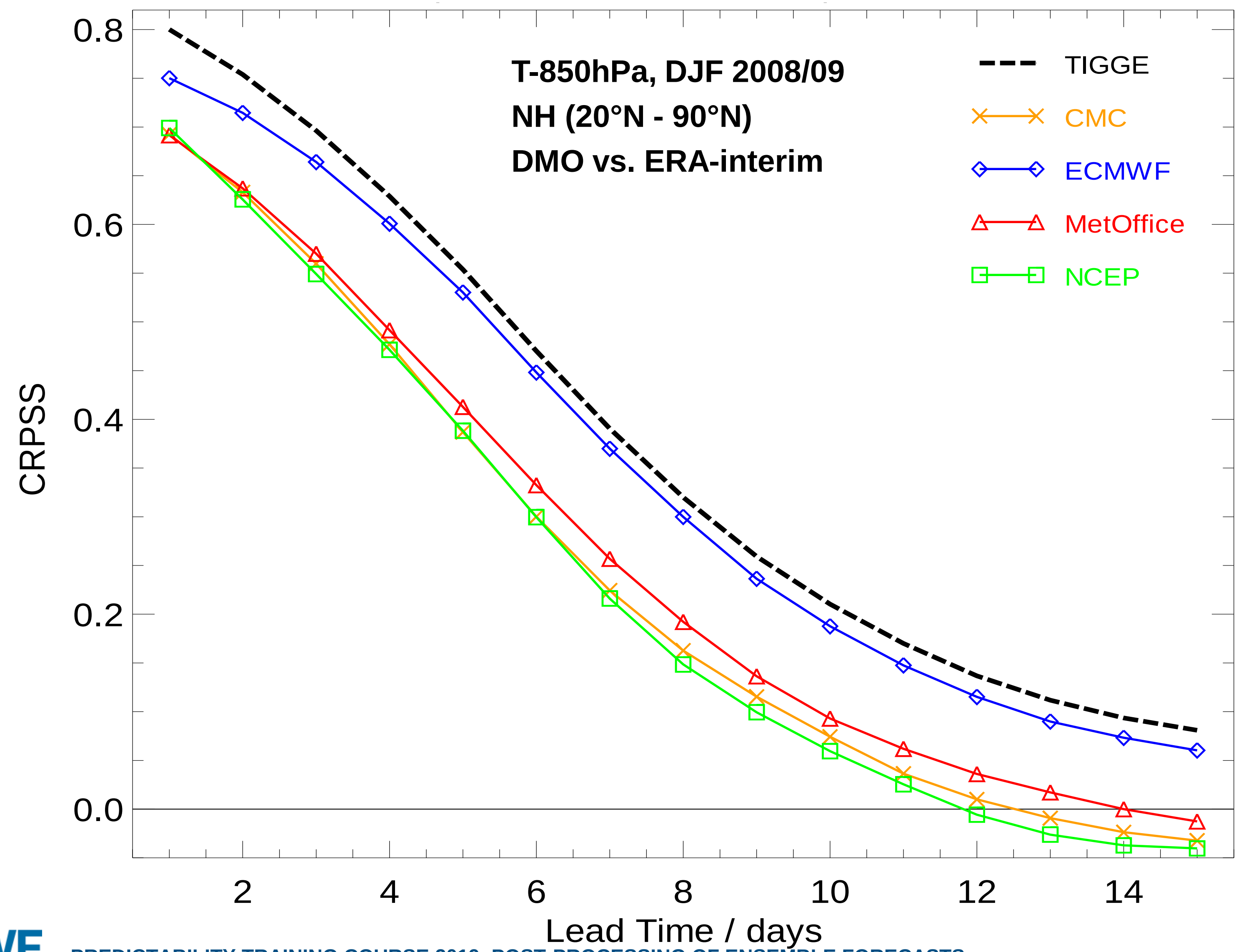
* TIGGE stands for: THORPEX Interactive Grand Global Ensemble

Comparing 9 TIGGE models & the MM

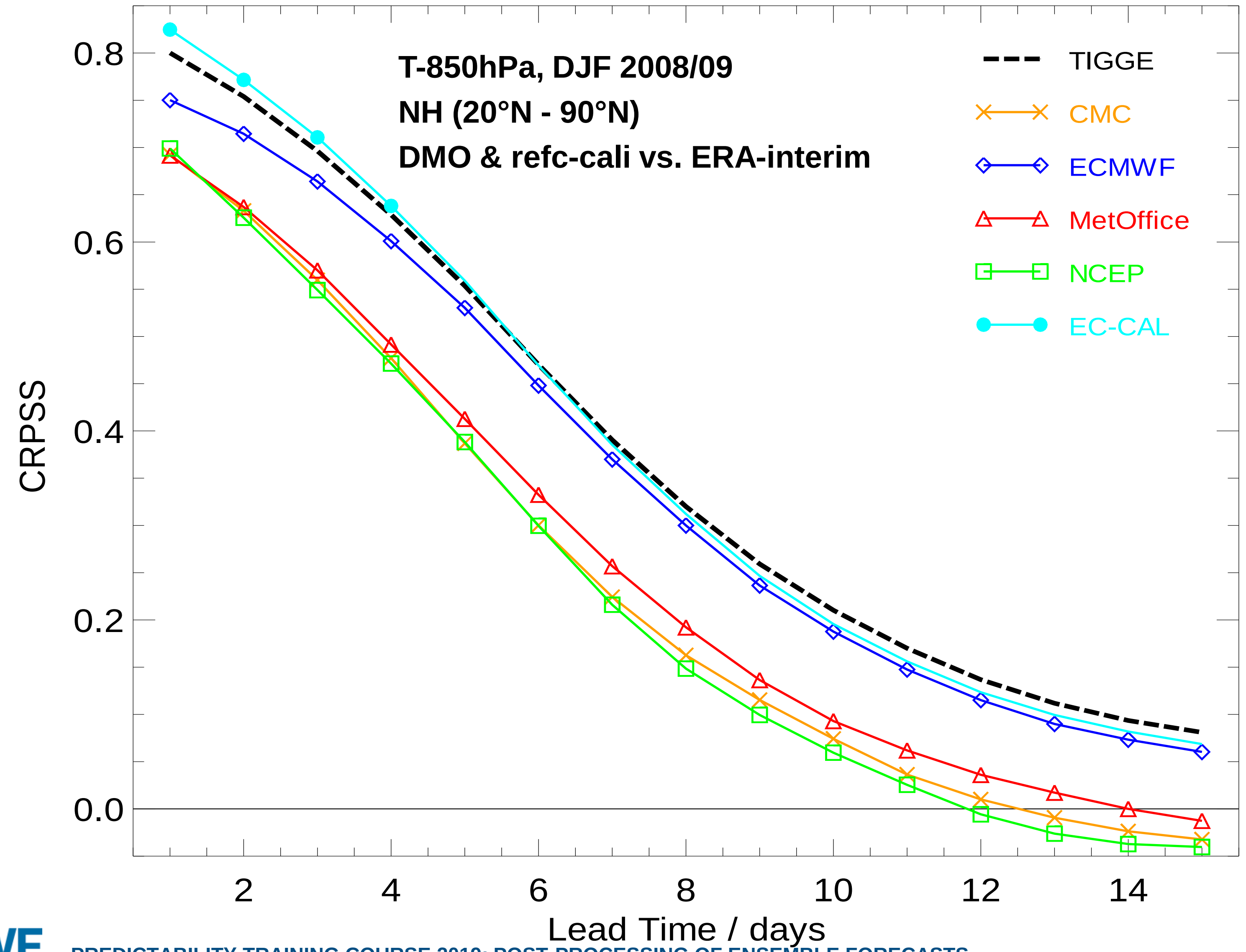


Symbols used for significance level vs. MM (1%)

Comparing 4 TIGGE models & the MM

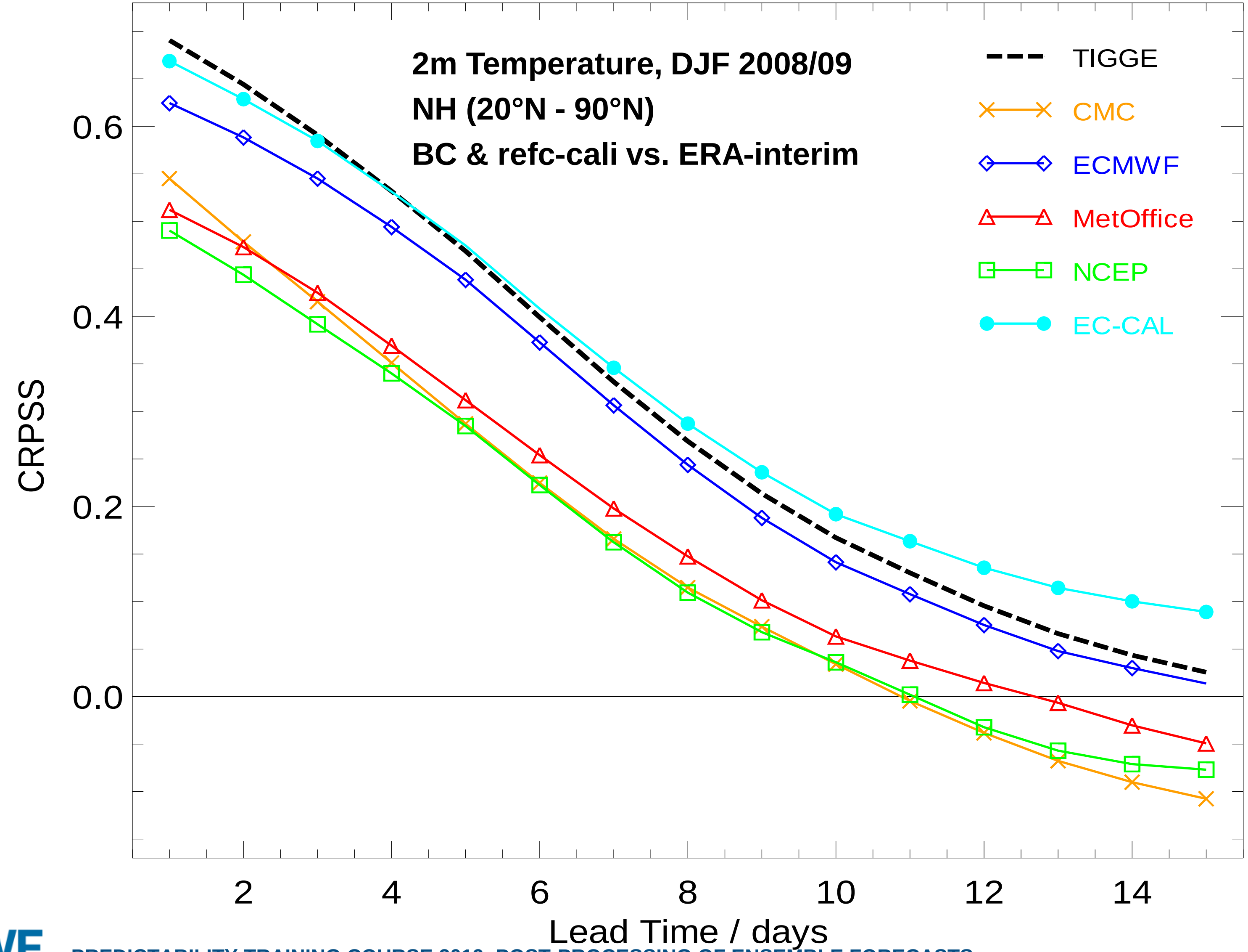


Comparing 4 TIGGE models, MM, EC-CAL

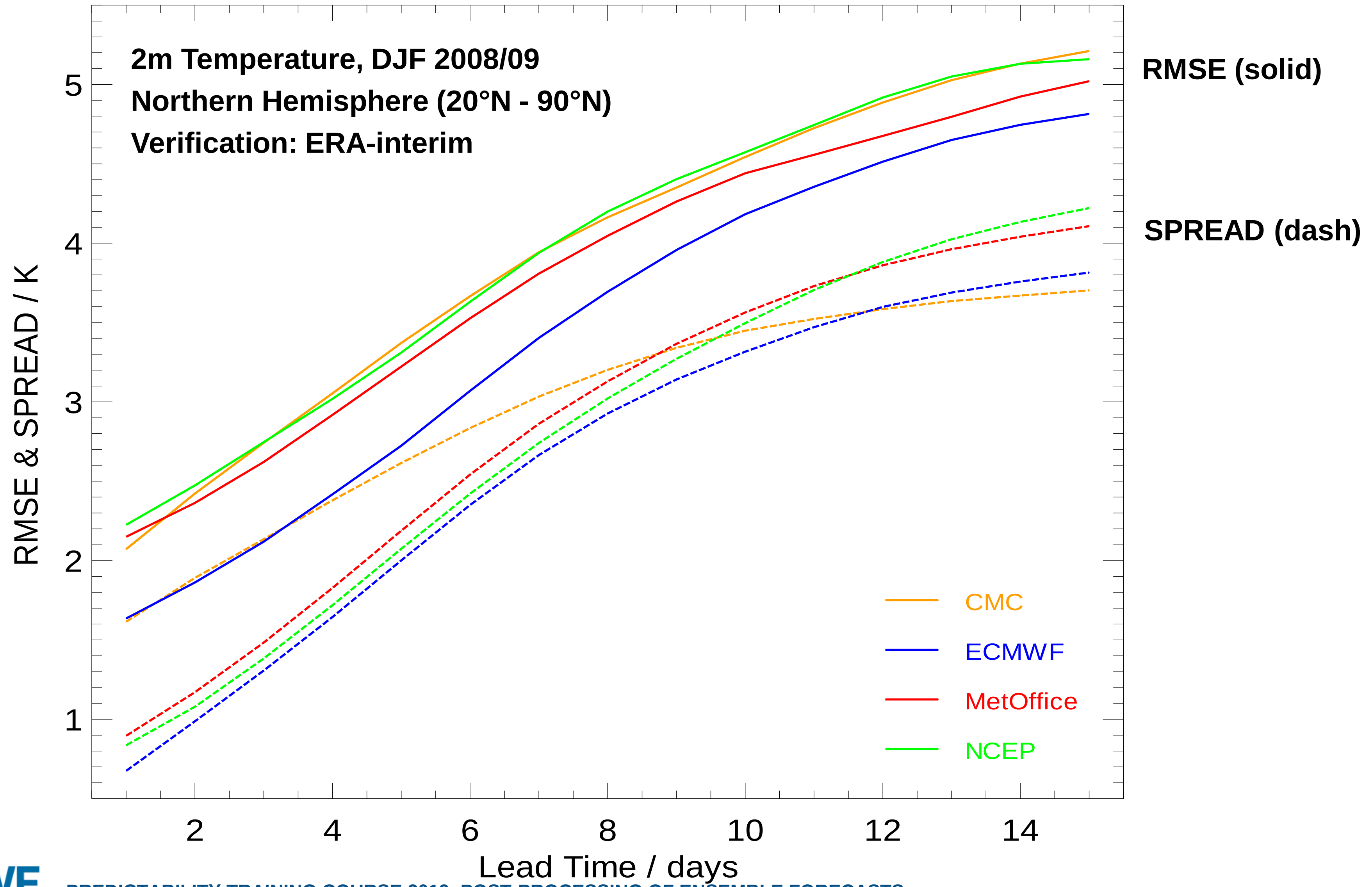


Note: *only* ECMWF is calibrated; other models do not have re-forecast datasets

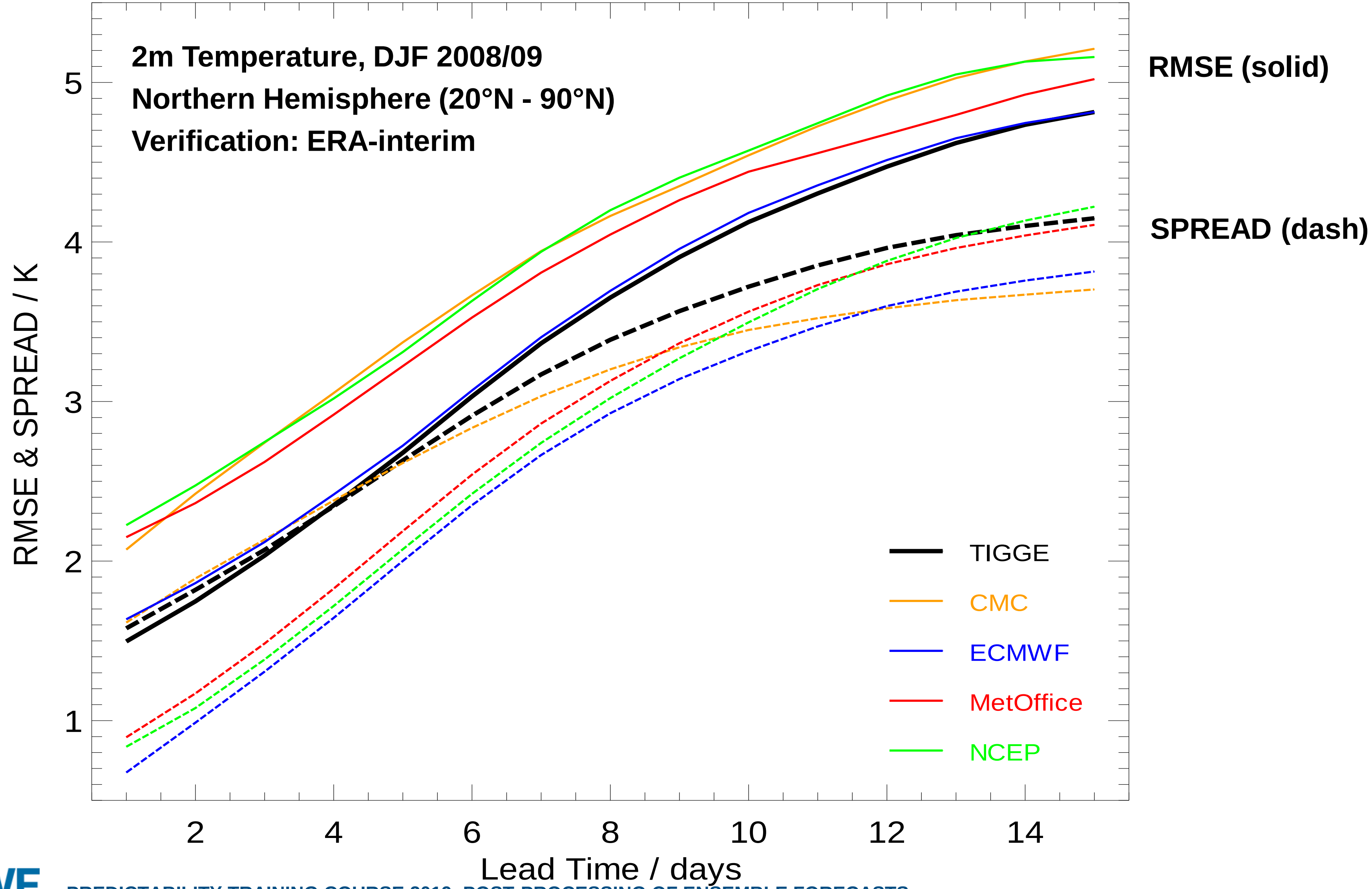
Comparing 4 TIGGE models, MM, EC-CAL



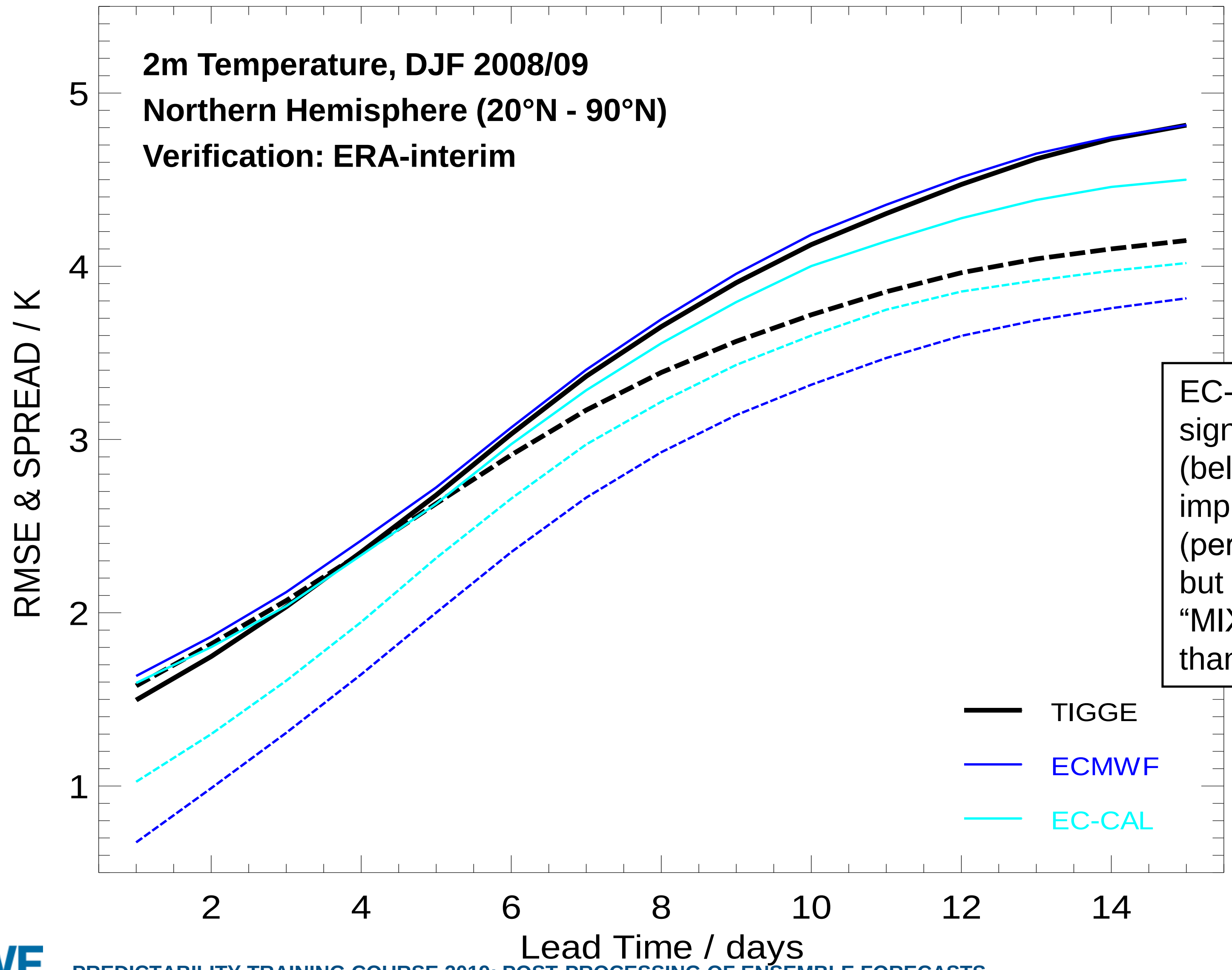
Mechanism behind improvements



Mechanism behind improvements



Mechanism behind improvements



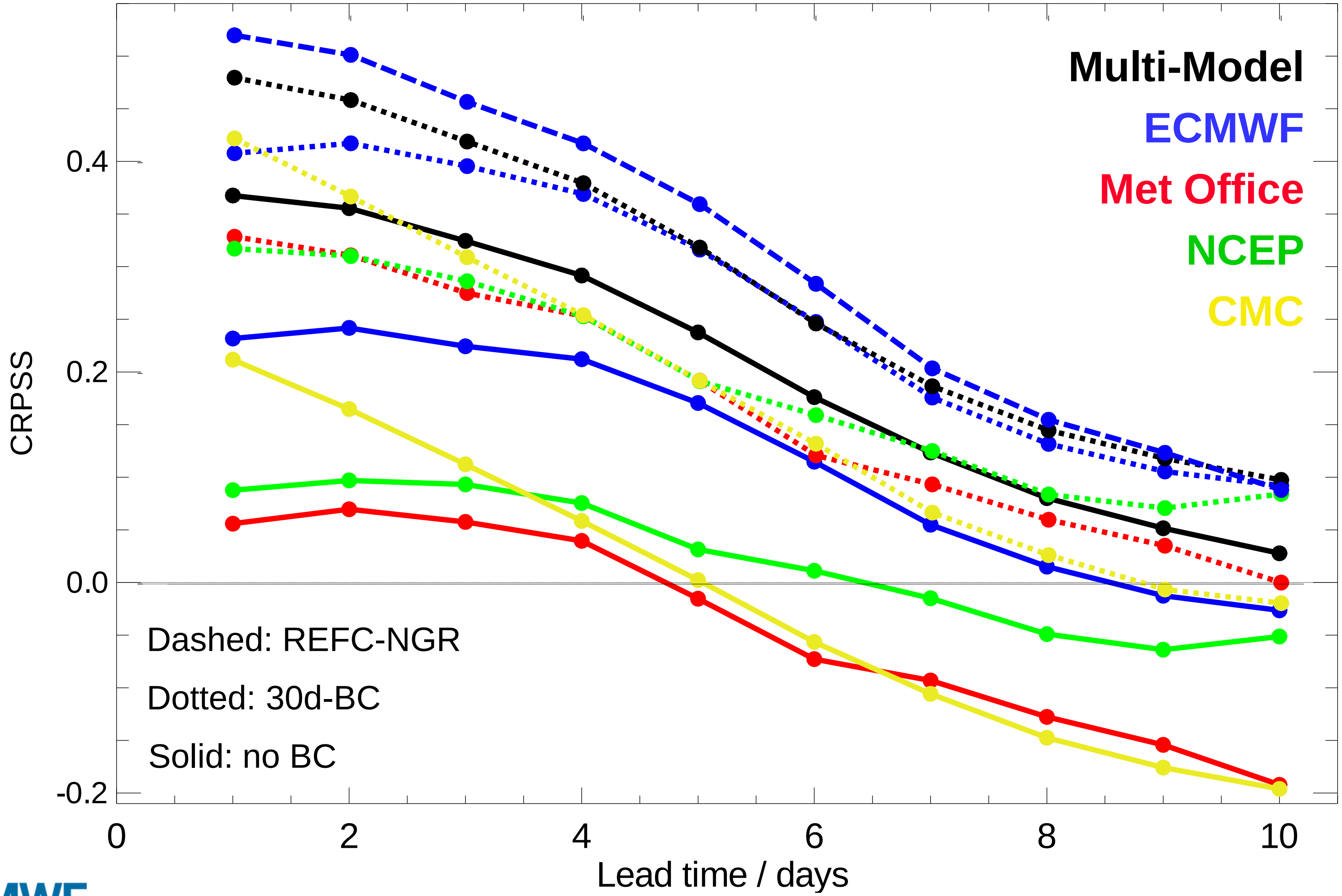
RMSE (solid)

SPREAD (dash)

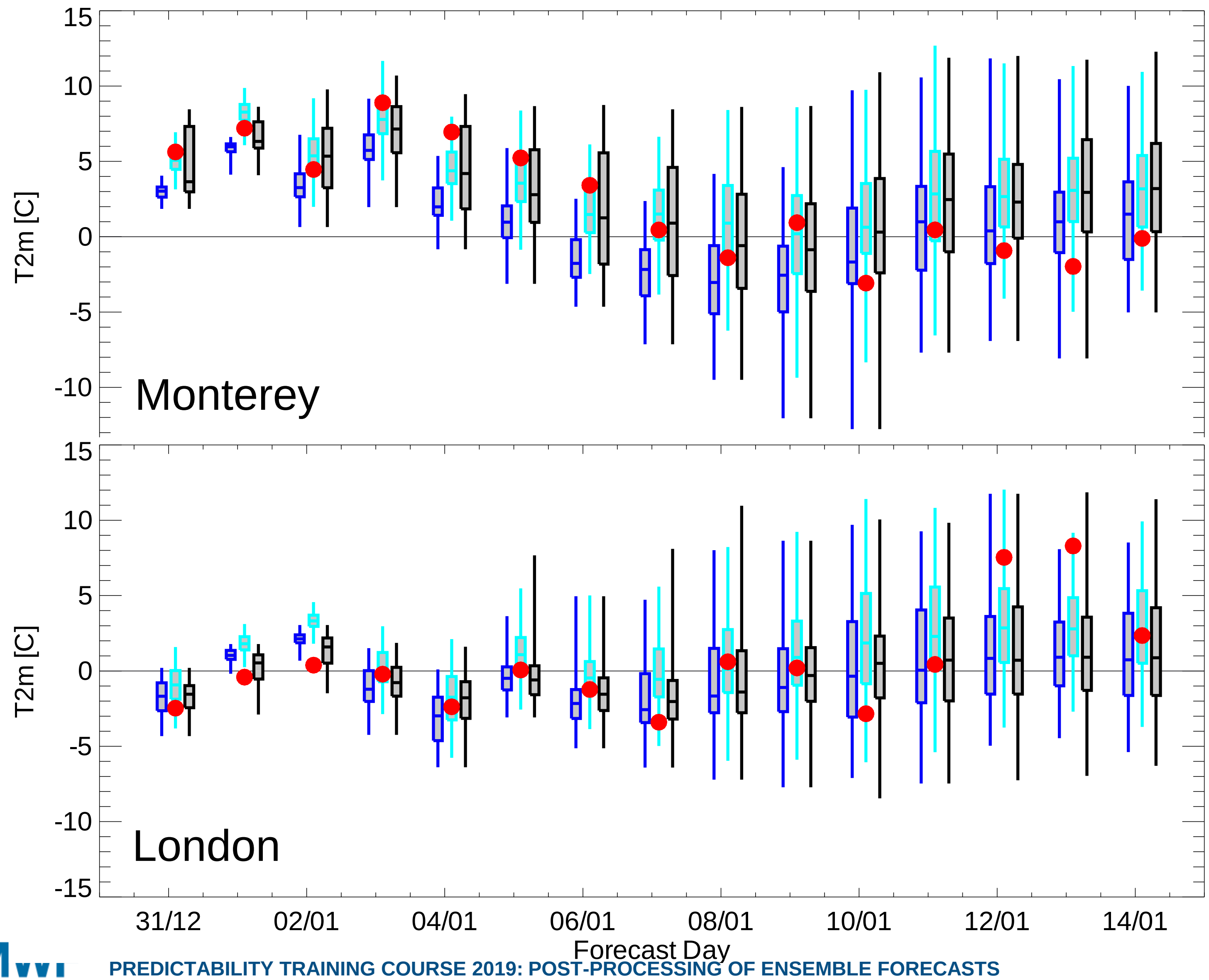
EC-CAL:
significant reduction of RMSE
(below MM-RMSE after day5)
improved SPR-ERR relation
(perfect for “pure” NGR,
but greater RMSE reduction of
“MIX” calibration more important
than better SPR-ERR)

What about station data?

T-2m, 250 European stations
DJF 2008/09



Impact of calibration & MM in EPSgrams



2m Temperature
FC: 30/12/2008

ECMWF
ECMWF-NGR
TIGGE
Analysis

An alternative view …

Reference: Hamill, 2012

- Examining precipitation forecasts over the US
- Four high skill models; compare ECMWF “re-forecast calibrated” with multi-model (no re-forecasts)
- Conclusions:
- “Raw multimodel PQPFs were generally more skilful than reforecast-calibrated ECMWF PQPFs for the light precipitation events but had about the same skill for the higher-precipitation events”
- “Multimodel ensembles were also postprocessed using logistic regression and the last 30 days of prior forecasts and analyses; Postprocessed multimodel PQPFs did not provide as much improvement to the raw multimodel PQPF as the reforecast-based processing did to the ECMWF forecast.”
- “The evidence presented here suggests that all operational centers, even ECMWF, would benefit from the open, real-time sharing of precipitation forecast data and the use of reforecasts.”

Summary on MM vs. calibration

- What are the relative benefits/costs of both approaches?
  - Both multi-model and a reforecast calibration approach can improve predictions, in particular for (biased and under-dispersive) near-surface parameters
- What is the mechanism behind the improvements?
  - Re-forecast calibration is effective at correcting local mis-representations in the model, and ensuring forecast uncertainty is well estimated
  - Multi-model approach can reduce forecast error as well as increasing spread; it tends to improve reliability but not necessarily in an optimal way
- Which is the “better” approach?
  - On balance, reforecast calibration seems to be the easier option for a reliable provision of medium-range forecasts in an operational environment
  - Both approaches can be useful in achieving the ultimate goal of an optimized, well tuned forecast system

Overall summary

- The goal of calibration is to correct for known forecasting system deficiencies
- A number of statistical methods exist to post-process ensembles
- Each method has its own strengths and weaknesses
  - Analogue methods seem to be useful when large training dataset available
  - Logistic regression can be helpful for extreme events not seen so far in training dataset
  - NGR method useful when strong spread-skill relationship exists, but relatively expensive in computational time
- Greatest improvements can be achieved on local station level
- Bias correction constitutes a large contribution for all calibration methods
- ECMWF re-forecasts are a very valuable training dataset for calibration

References and further reading

- Gneiting, T. et al, 2005: Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, **133**, 1098-1118.
- Hagedorn, R, T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: 2-meter temperature. *Monthly Weather Review*, **136**, 2608-2619.
- Hagedorn, R., Buizza, R., Hamill, T. M., Leutbecher, M. and Palmer, T. N., 2012: Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Q.J.R. Meteorol. Soc.* doi: 10.1002/qj.1895
- Hamill, T.M., 2012: Verification of TIGGE Multi-model and ECMWF Reforecast-Calibrated Probabilistic Precipitation Forecasts over the Contiguous US. *Monthly Weather Review*, doi: 10.1175/MWR-D-11-00220.1
- Hamill, T.M. et al., 2004: Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts. *Monthly Weather Review*, **132**, 1434-1447.
- Hamill, T.M. and J.S. Whitaker, 2006: Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application. *Monthly Weather Review*, **134**, 3209-3229.
- Raftery, A.E. et al., 2005: Using Bayesian Model Averaging to Calibrate Forecast Ensembles. *Monthly Weather Review*, **133**, 1155-1174.
- Wilks, D. S., 2006: Comparison of Ensemble-MOS Methods in the Lorenz ’96 Setting. *Meteorological Applications*, **13**, 243-256.