Post-Processing of Ensemble Forecasts

Tim Stockdale
(incorporating material adapted from Renate Hagedorn)

European Centre for Medium-range Weather Forecasts
t.stockdale@ecmwf.int

ECMWF

Outline

- Motivation
- Calibration methods
- Training data sets
- Multi-model forecasts

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

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

Motivation

- Raw, uncalibrated ensemble forecasts contain forecast **bias** and errors in **spread**
- 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
- Ensemble dressing
- Bayesian model averaging
- Non-homogenous Gaussian regression
- Logistic regression
- Analogue method

Bias correction

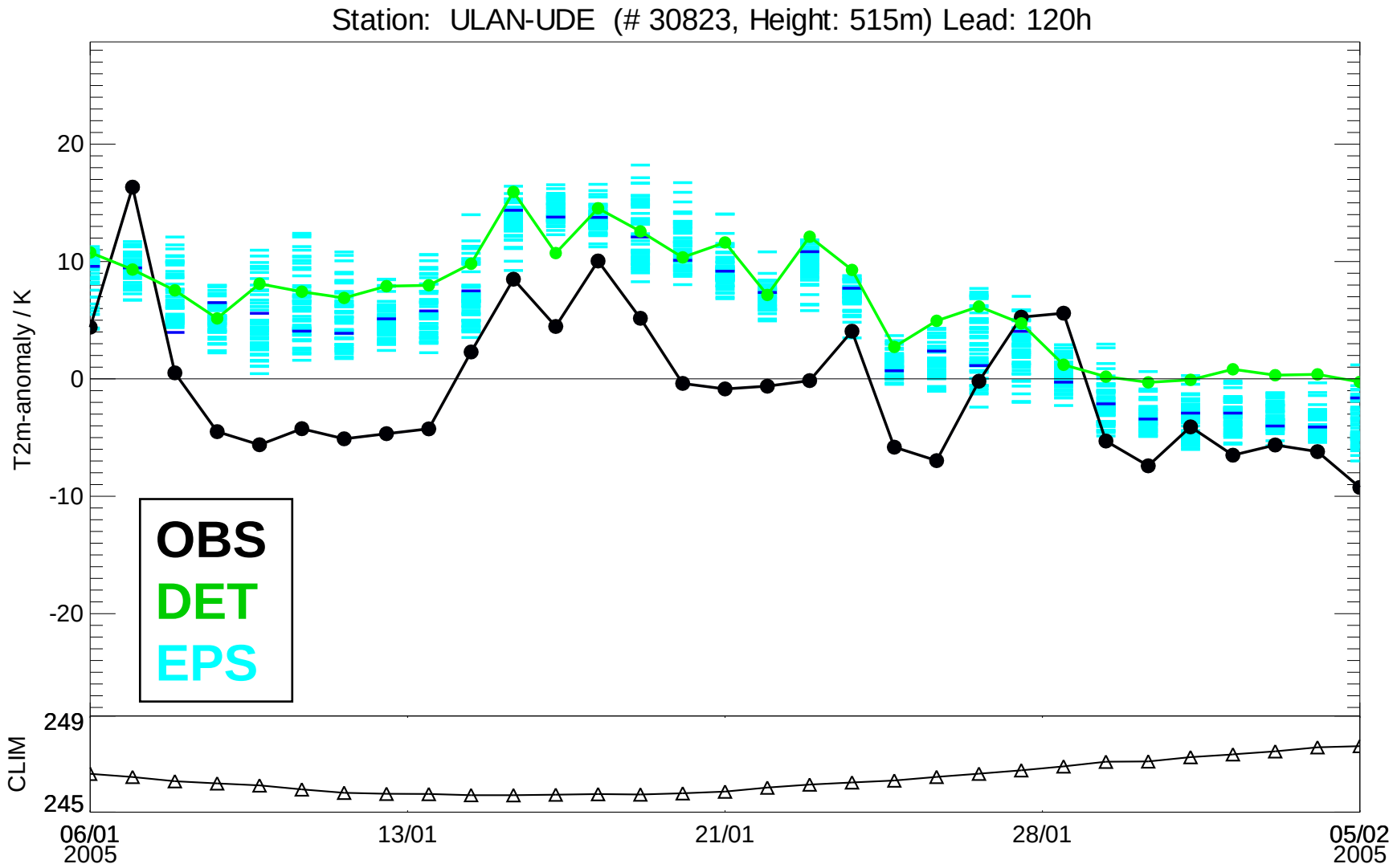
- As a simple first-order calibration, a bias correction can be applied:

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

with: $\boldsymbol{e_i}$ = ensemble mean of the ith forecast
$\boldsymbol{o_i}$ = value of ith observation
N = number of observation-forecast pairs

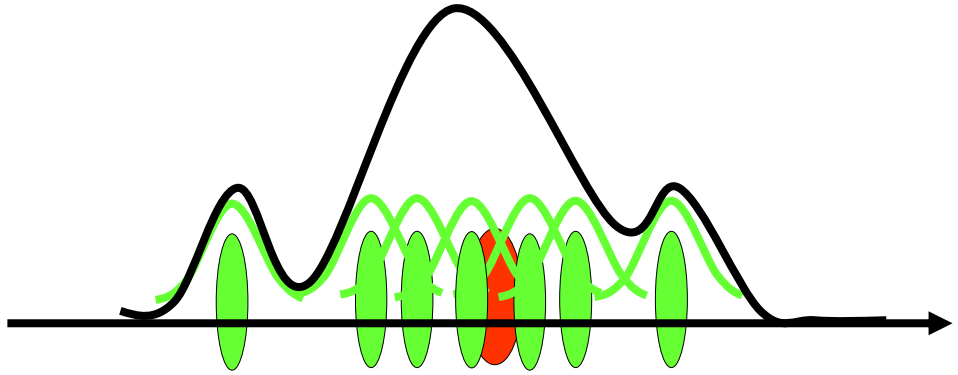
- This correction is added to each ensemble member, i.e. spread is not affected
- Particularly useful/successful at locations with features not resolved by model and causing significant bias

Bias correction



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
- Consider bias correcting the ensemble first

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: Φ = CDF of standard Gaussian distribution
x_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

- A more 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 is 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:

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

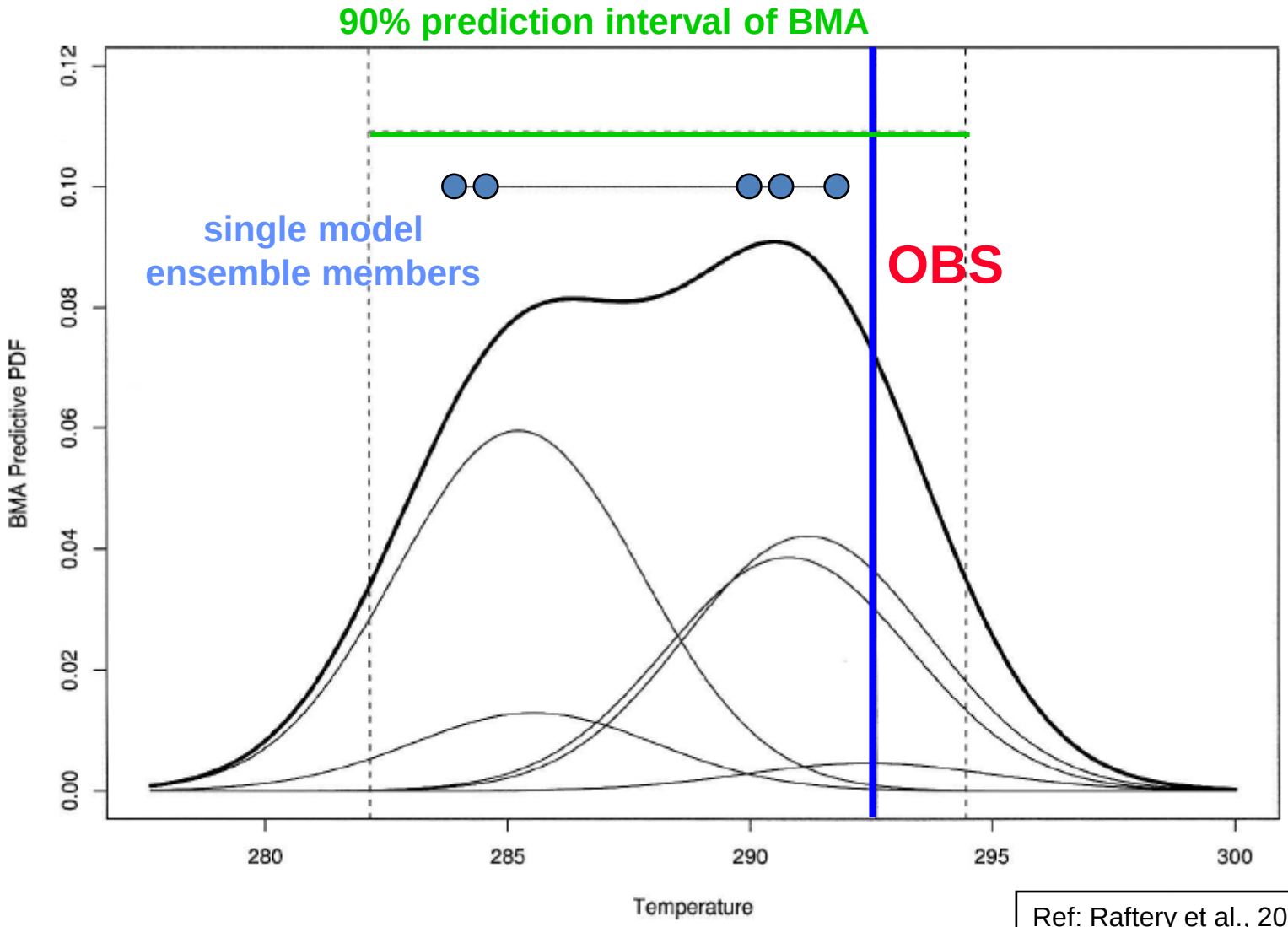
with: $w_1 + w_e (n_{ens} - 1) = 1$

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

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

g_1, g_e = Gaussian PDF's

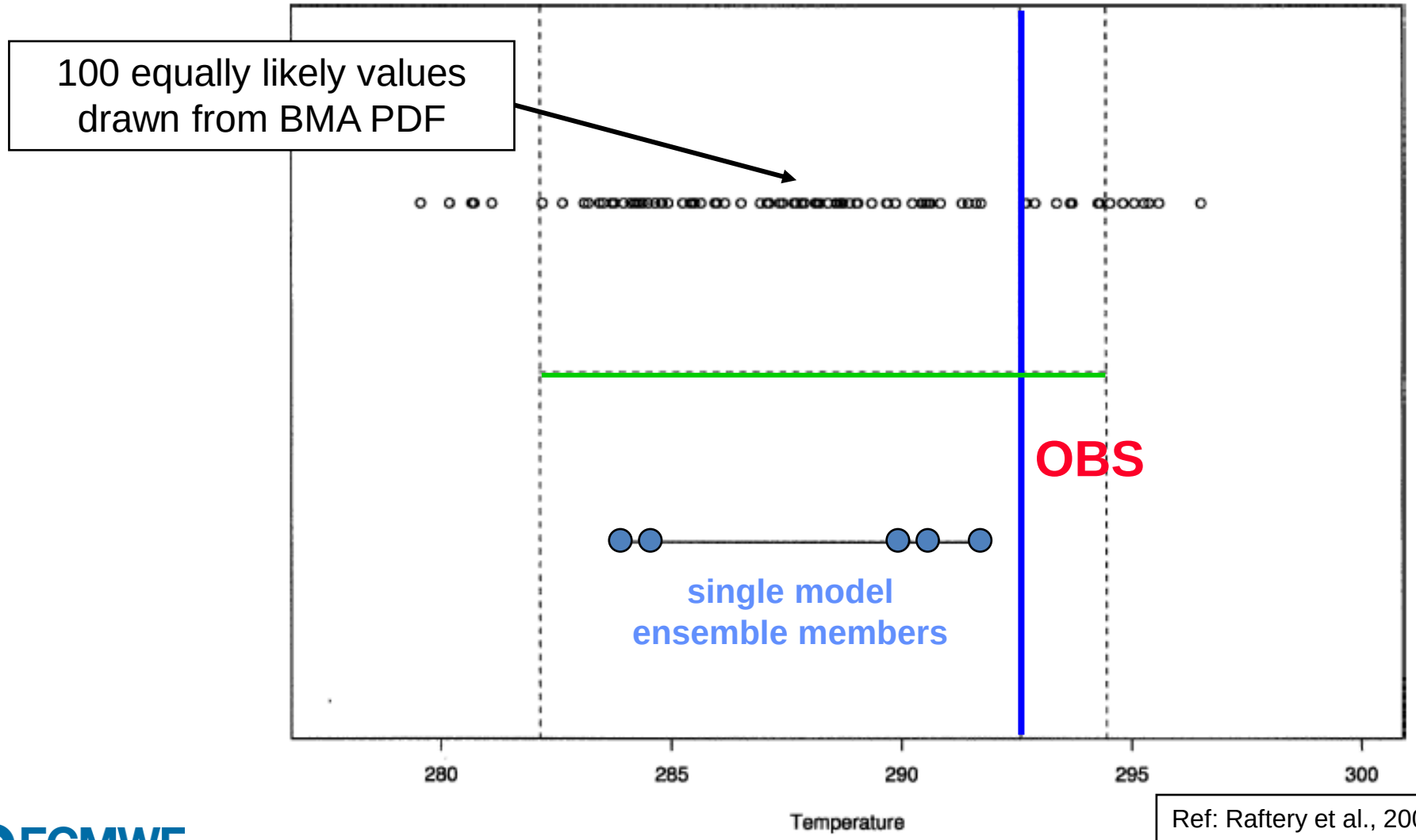
BMA: example



Ref: Raftery et al., 2005, MWR

BMA: recovered ensemble members

Can use this technique for any pdf, of course



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}:

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

- The parameters a,b,c,d are fit iteratively by minimizing the CRPS of the training data set
- Interpretation of parameters:
  - bias & scaling/general performance of ensemble-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 – different from BMA)

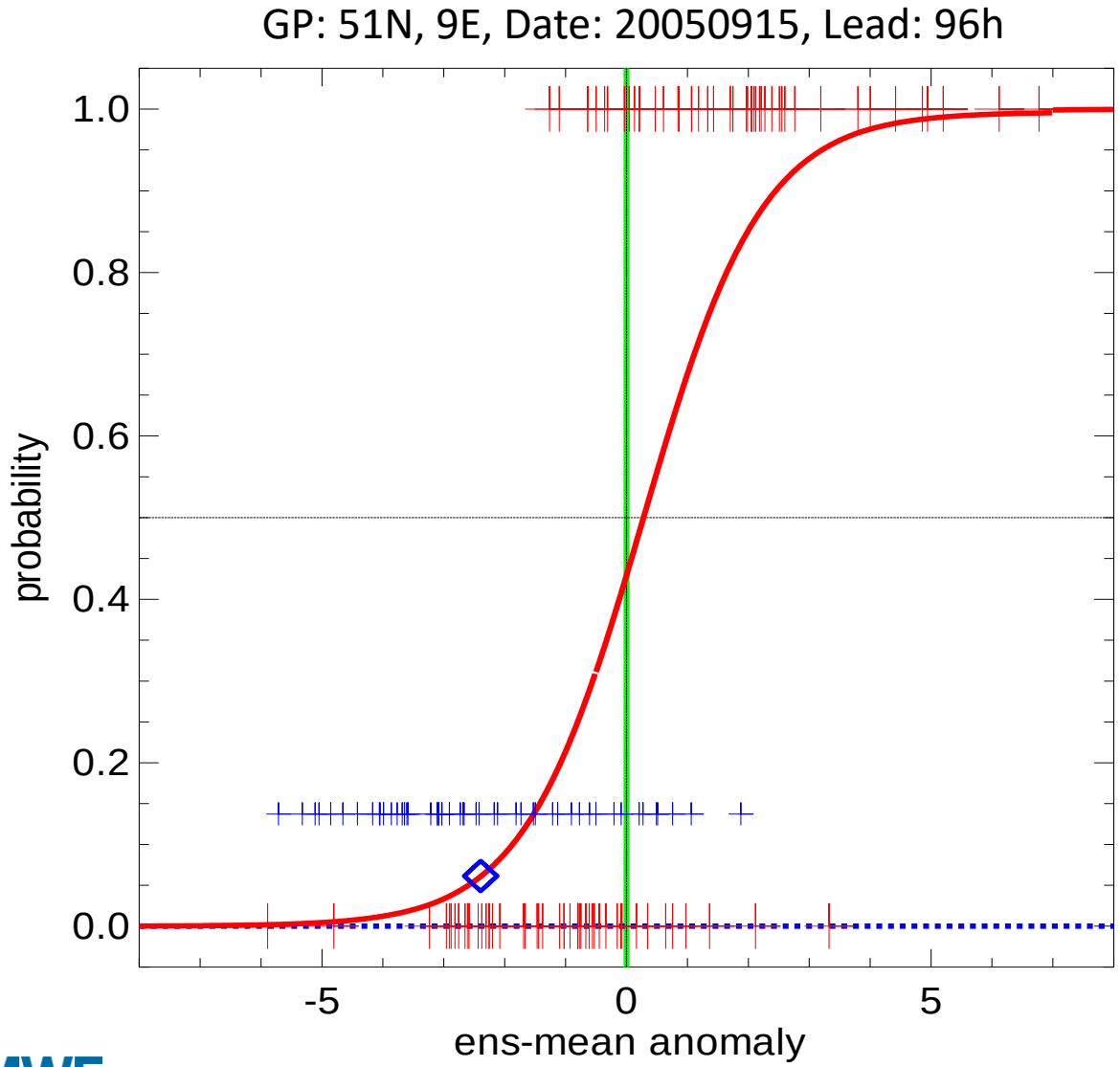
Logistic regression for **event** probabilities

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

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

- 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
  - Estimate β_0 and β_1 by fitting data to maximize likelihood, in our case using ensemble means from a forecast set and corresponding event outcomes

How does logistic regression work?



+ training data
  100 cases (EnsMean)
  (height = obs yes/no for ens mean>0)

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

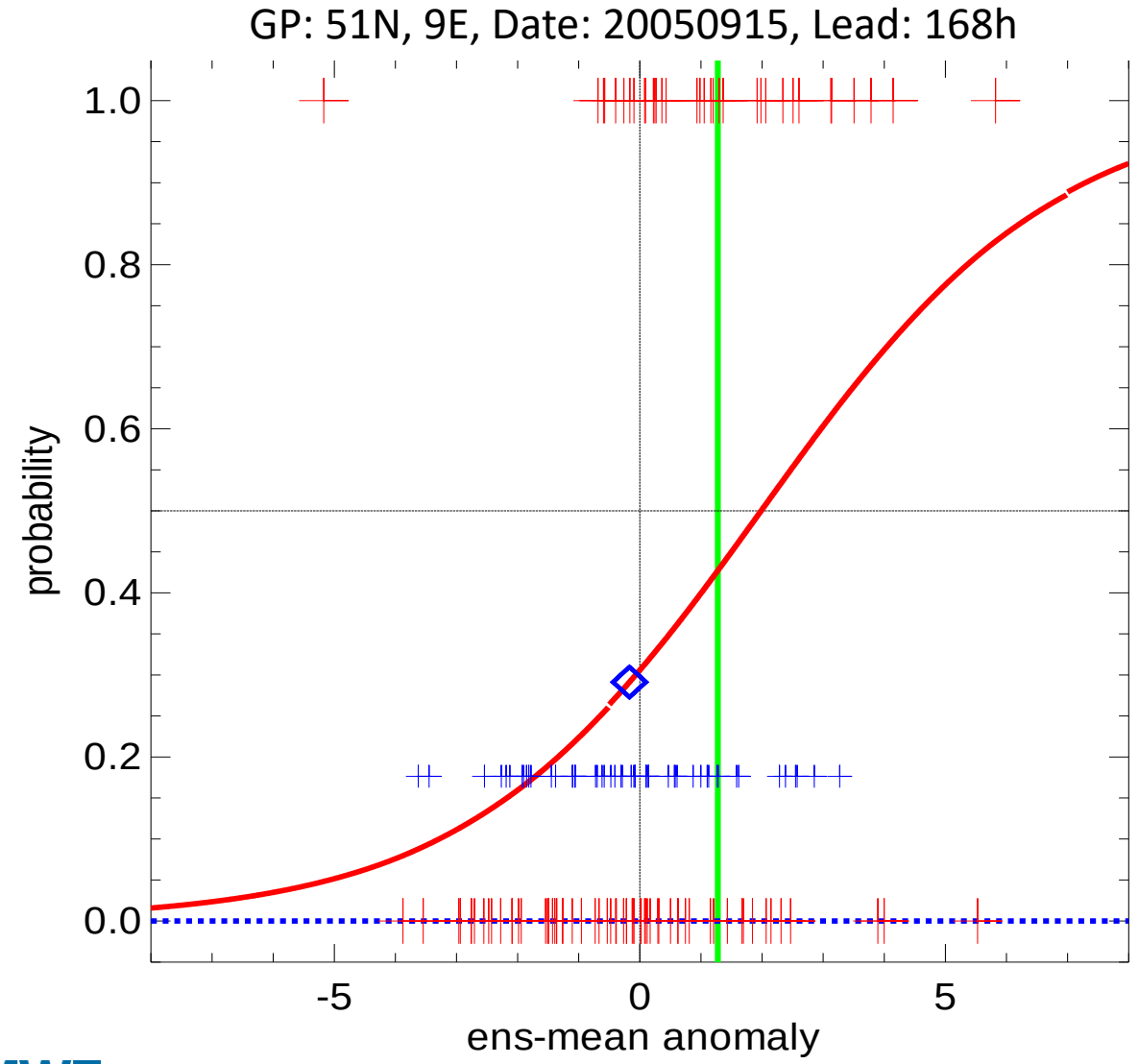
◇ calibrated prob

event observed
yes/no (0/1)

event threshold

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)

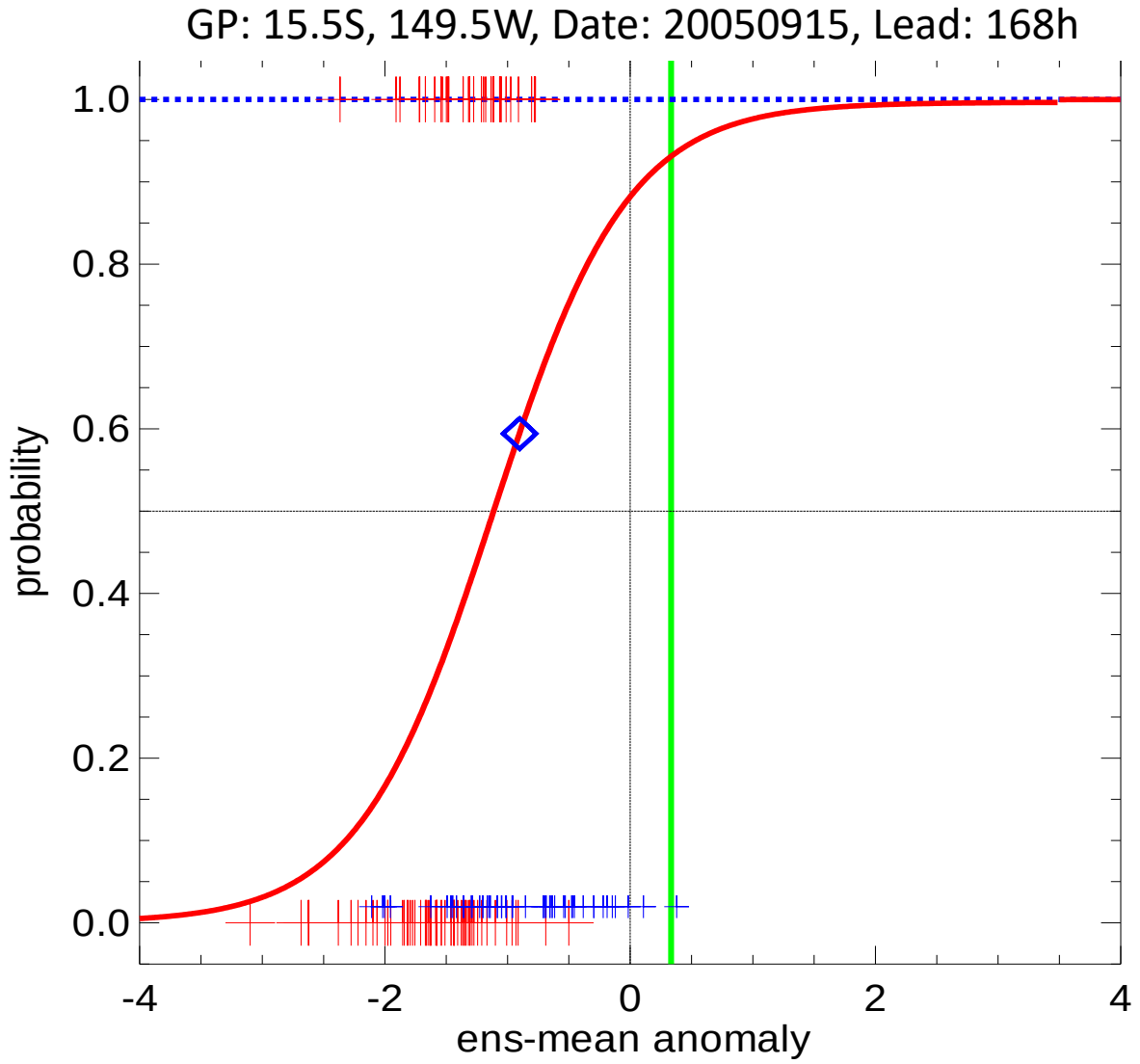
◇ calibrated prob

event observed
yes/no (0/1)

event threshold

Event did not happen

Example: LR-Probability (much) better!



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

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

◇ calibrated prob

event observed
yes/no (0/1)

event threshold

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

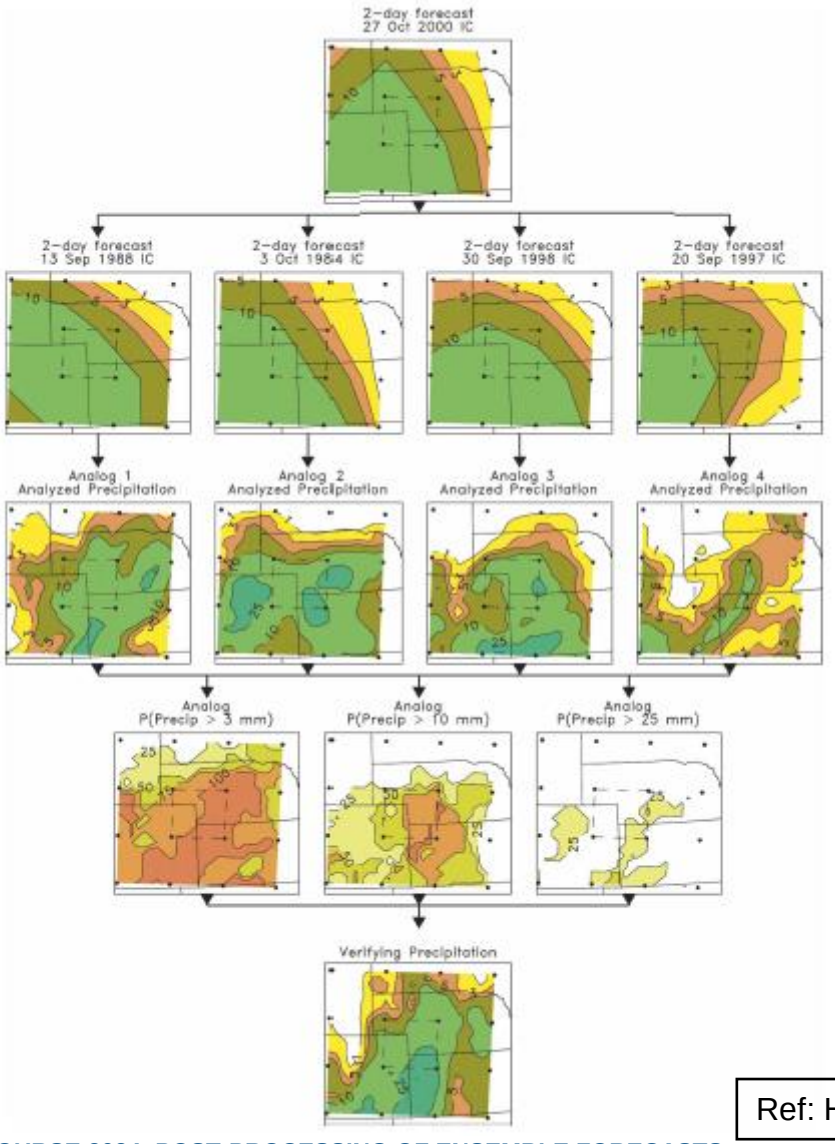
Forecast to be calibrated

Closest re-forecasts

Corresponding obs

Probabilities given by 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 forecast system used to produce the training dataset should be as close as possible to the operational forecast system (easy for the model, may be less easy for initial conditions)
- 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

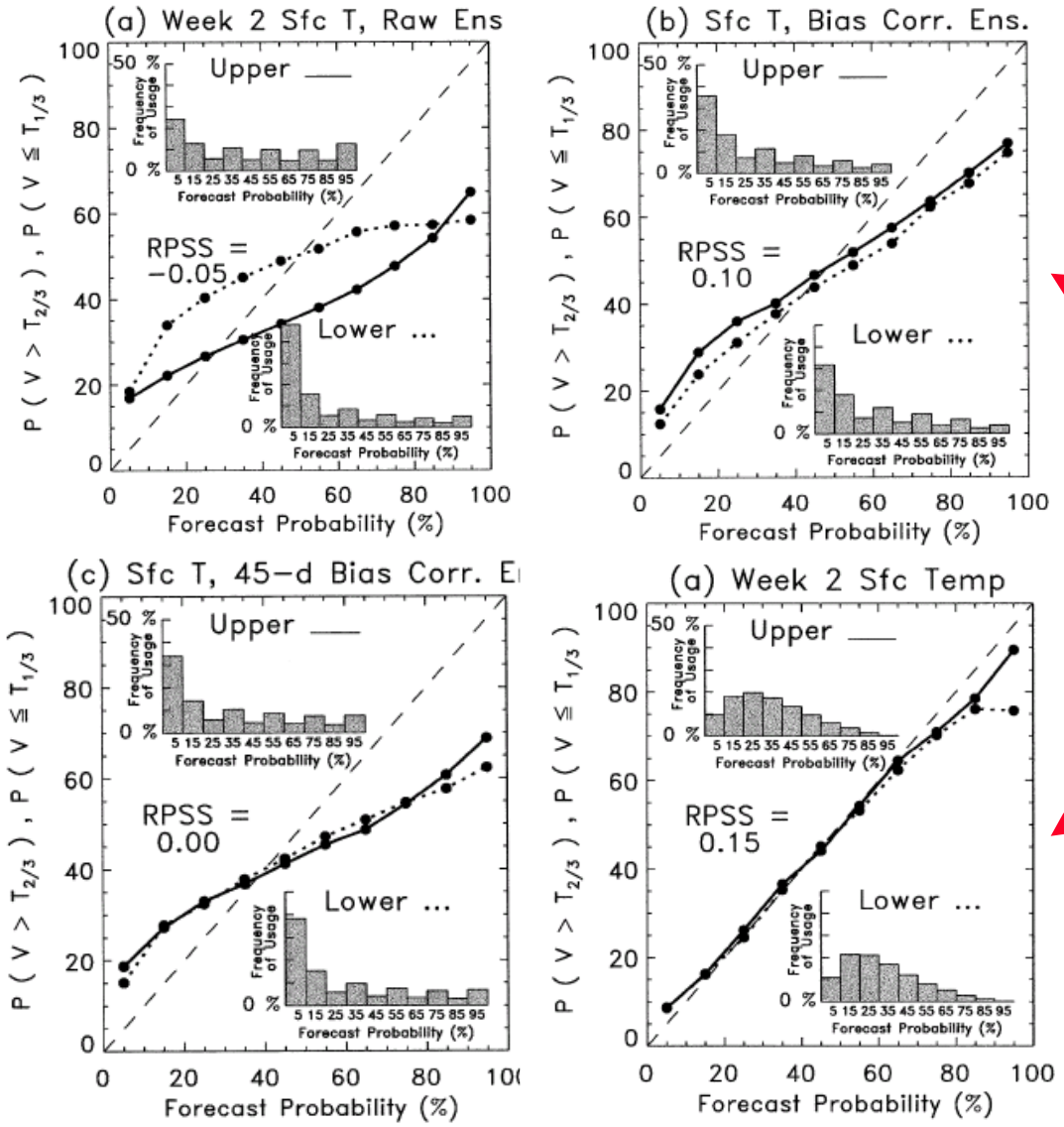
- Aside on calibration, prediction and model selection
  - ML prediction: distinguish between training data, validation data and test data
  - Calibration: calibration data and evaluation data

“Ideal” Reforecast Data Set

	2022																																
	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																																			
. . . .																																			
2017																																			
2018																																			
2019																																			
2020																																			
2021																																			

Early motivating results from Hamill et al., 2004

Raw ensemble



Bias corrected with refc data

Achieved with "ideal" reforecast system!

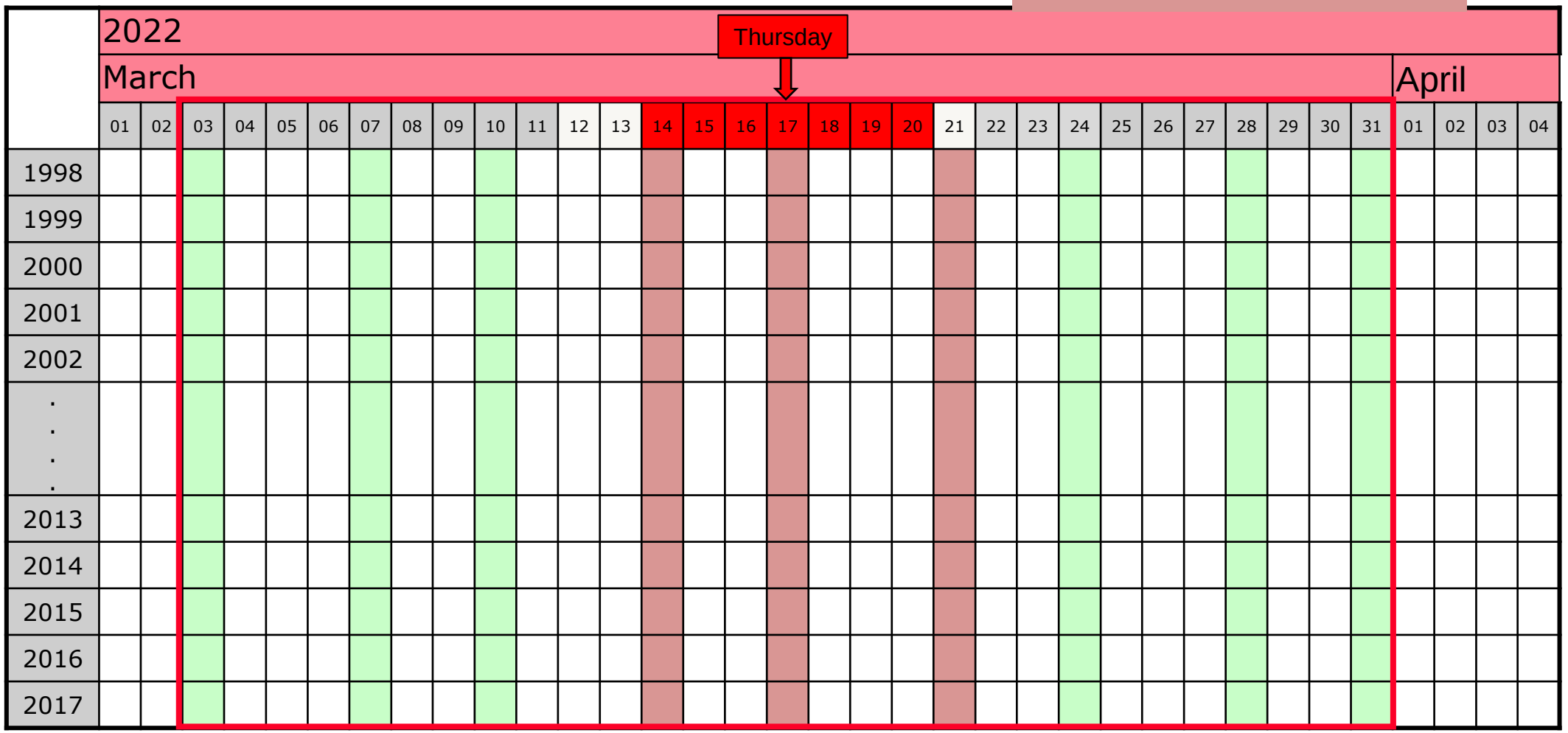
Bias corrected with 45-d data

LR-calibrated ensemble

Medium-range and Sub-seasonal reforecasts (at present)

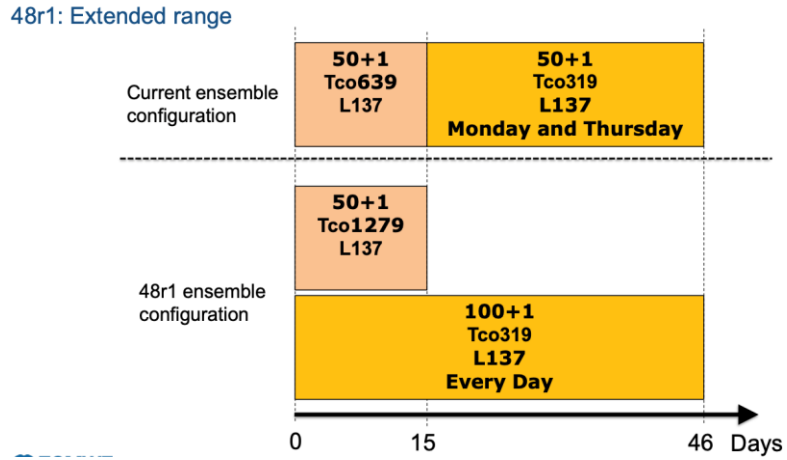
Used by medium-range in EFI and SOT

Used in sub-seasonal forecast



Changes to medium-range and sub-seasonal re-forecasts

- Separate ensemble systems now enable the production of two reforecast data sets, to be used by:
  - EFI model climate and 15-day medium-range calibration
  - Sub-seasonal forecast anomalies and verification
- Re-forecast configurations have to be an optimal compromise between affordability and needs of different applications
- At present, both re-forecasts are still 11 members, twice per week
- Will change in Cy49r1 to 11 members every 2 days (sub-seasonal) or every 4 days (medium-range)



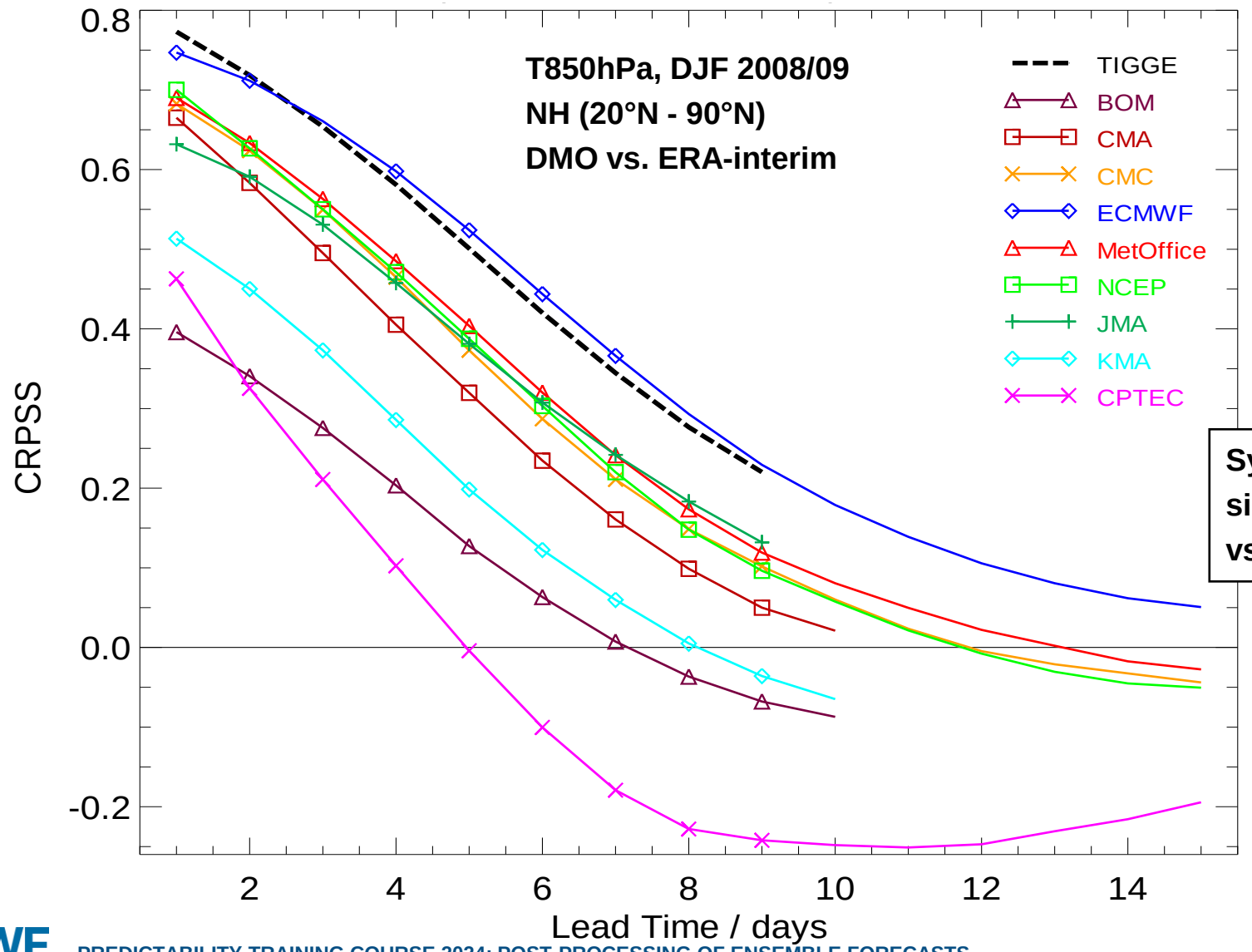
Reforecast calibration and multi-model forecasts

Reference: Hagedorn et al, 2012

- One goal of the TIGGE* project was to investigate whether multi-model predictions are an improvement compared to single model forecasts
- The goal of using reforecasts to calibrate single model forecasts is also to provide improved predictions
- Questions:
  - What are the relative benefits (and costs) of both approaches?
  - What are the mechanisms behind any improvements?
  - Can we say which approach is “better”?

* TIGGE stands for: The International Grand Global Ensemble (originally THORPEX interactive GGE). It provides a research database of medium-range ensemble forecasts from 10 forecasting centres, downloadable with a delay of 48h.

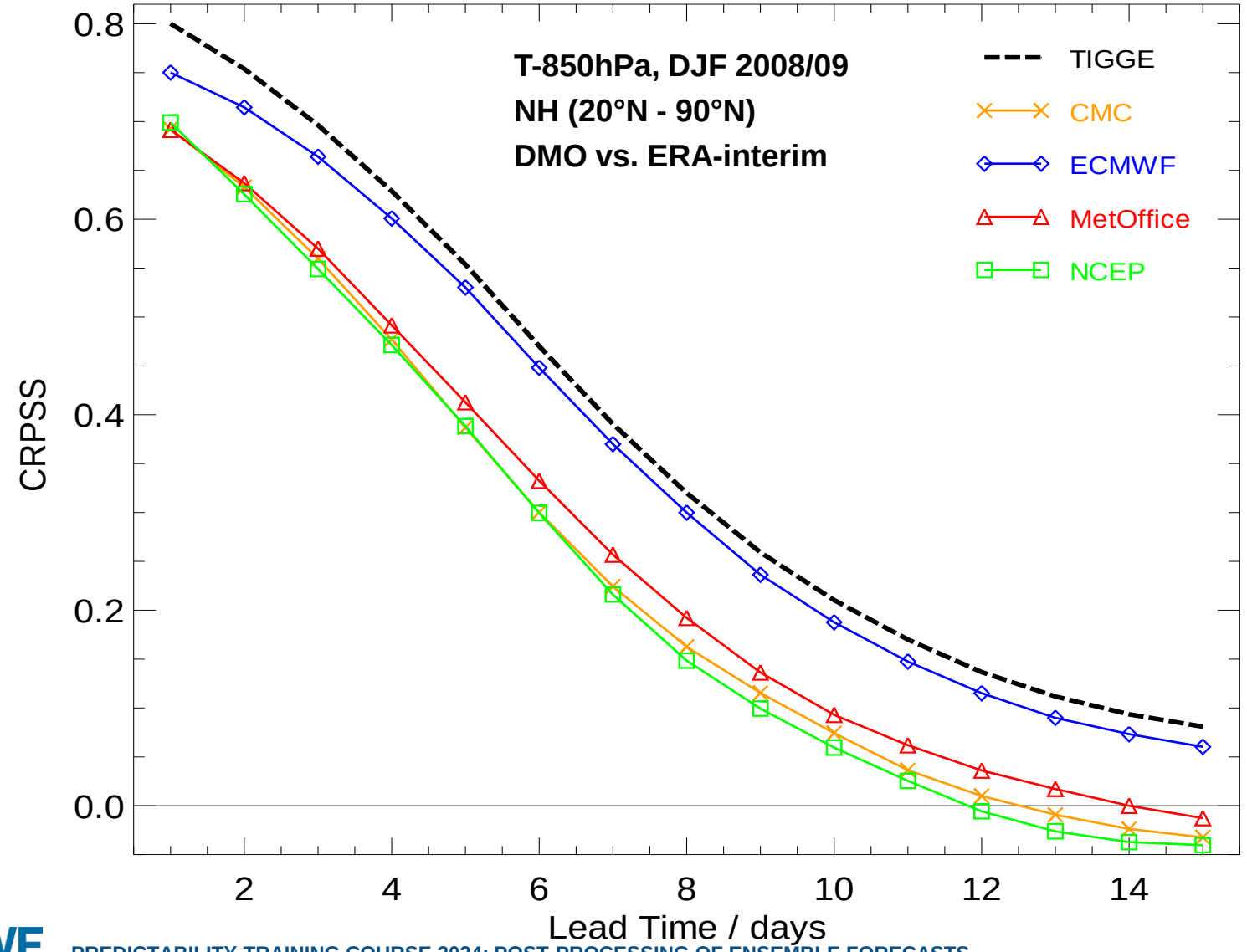
Comparing 9 TIGGE models & the resulting multi-model mean (MM)



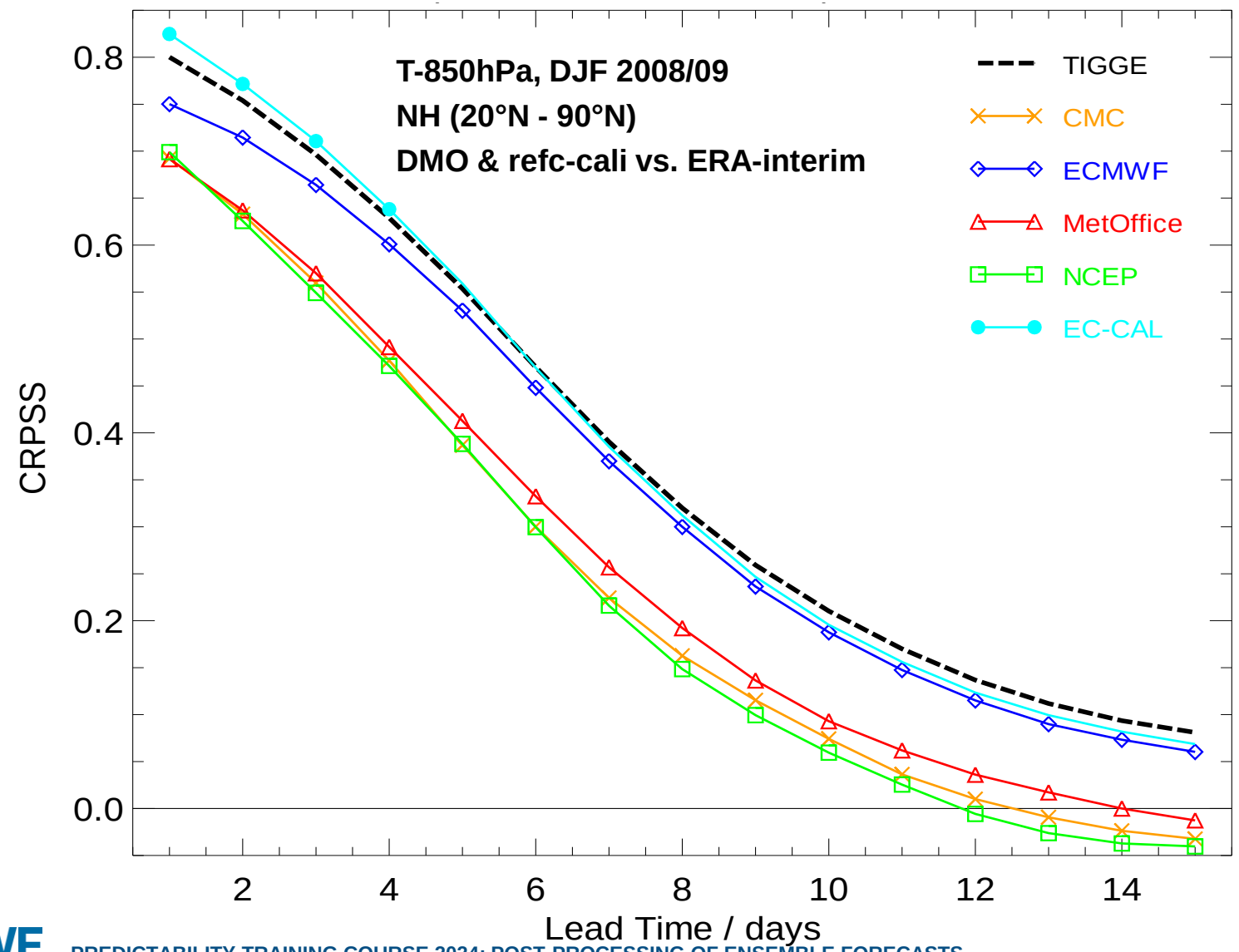
Symbols used for significance level vs. MM (1%)

DMO=Direct model output

Comparing 4 best TIGGE models & the MM

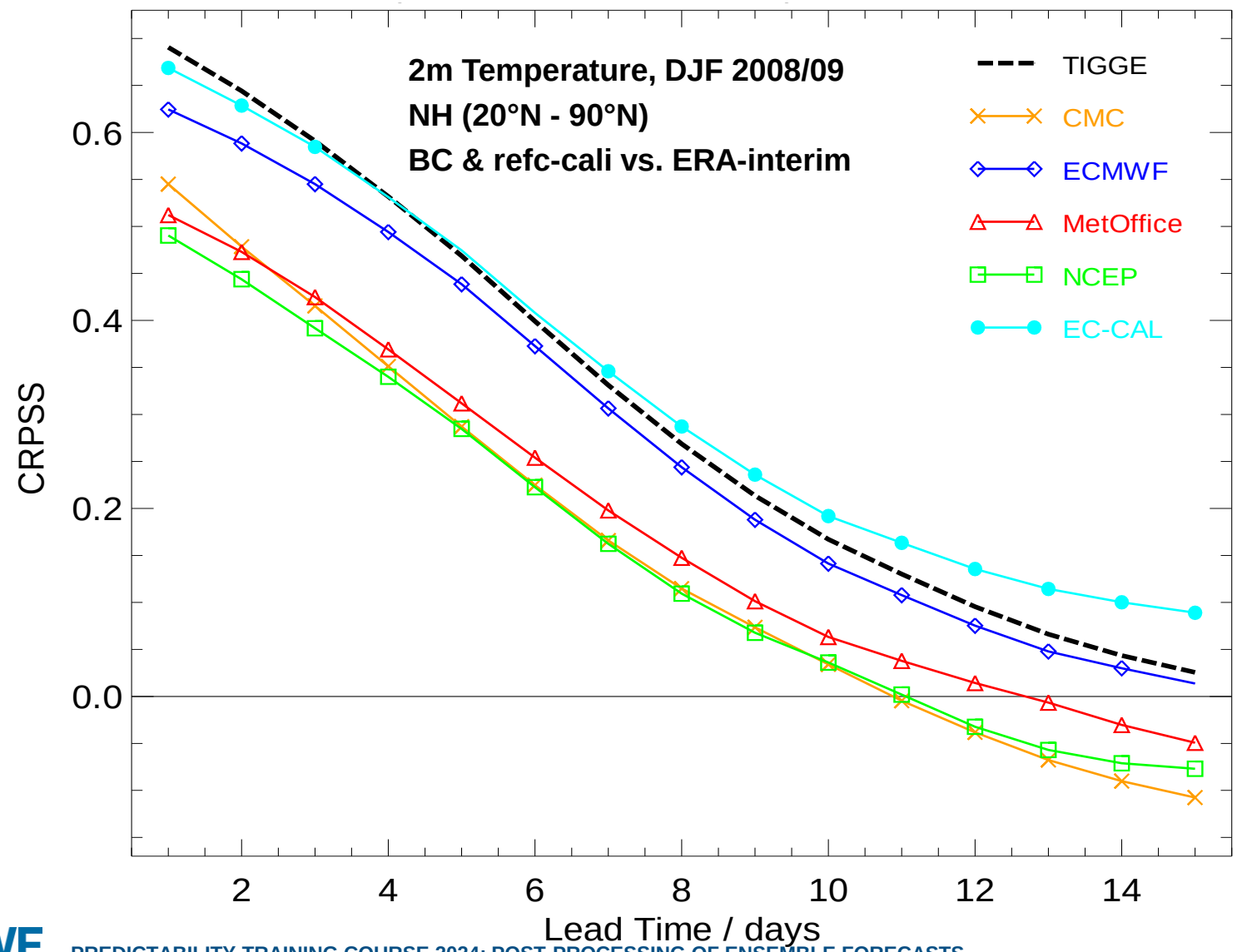


Comparing 4 TIGGE models, MM, EC-CAL

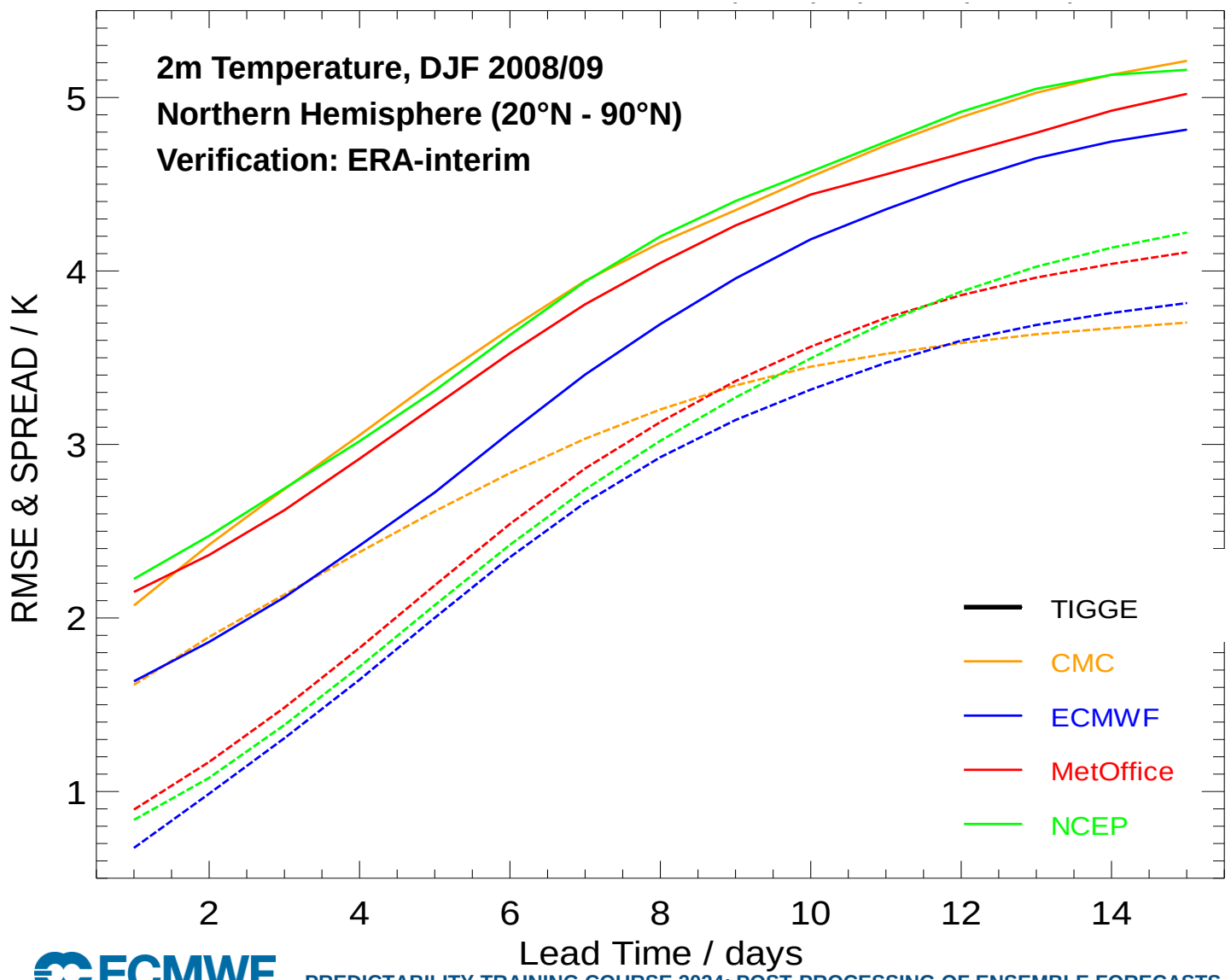


Note: *only* ECMWF is calibrated; other models do not have re-forecast datasets

Comparing 4 TIGGE models, MM, EC-CAL

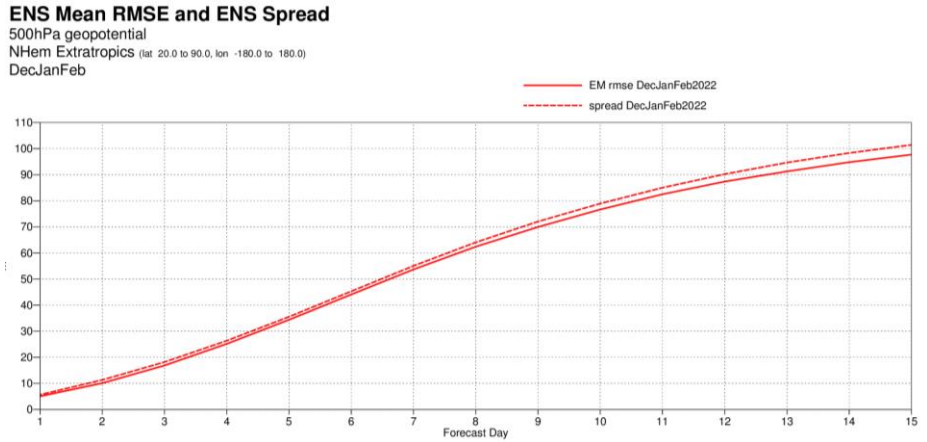


Mechanism behind improvements



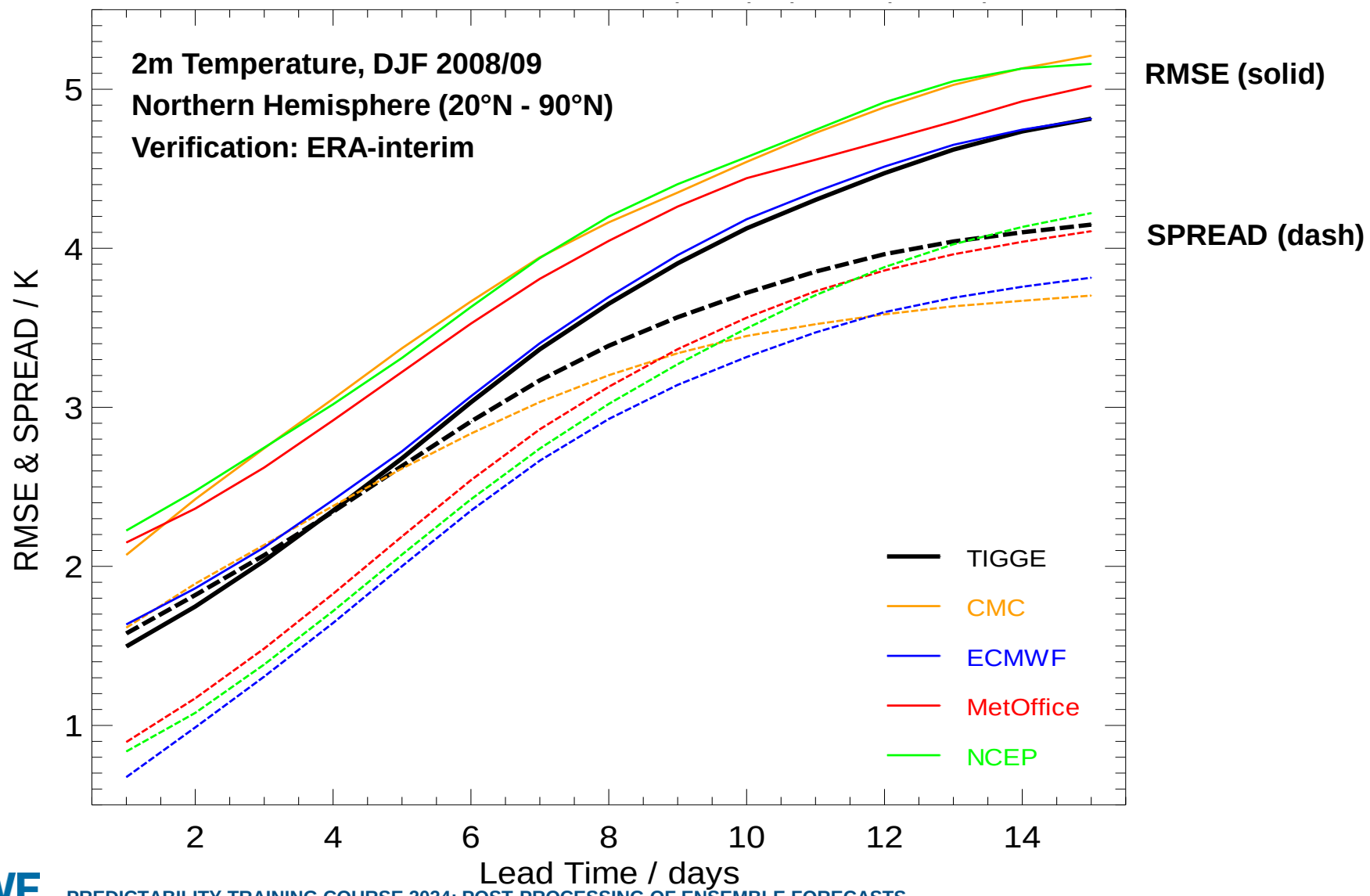
RMSE (solid)

SPREAD (dash)



Aside: spread/error is much better matched for upper-air scores, especially today.

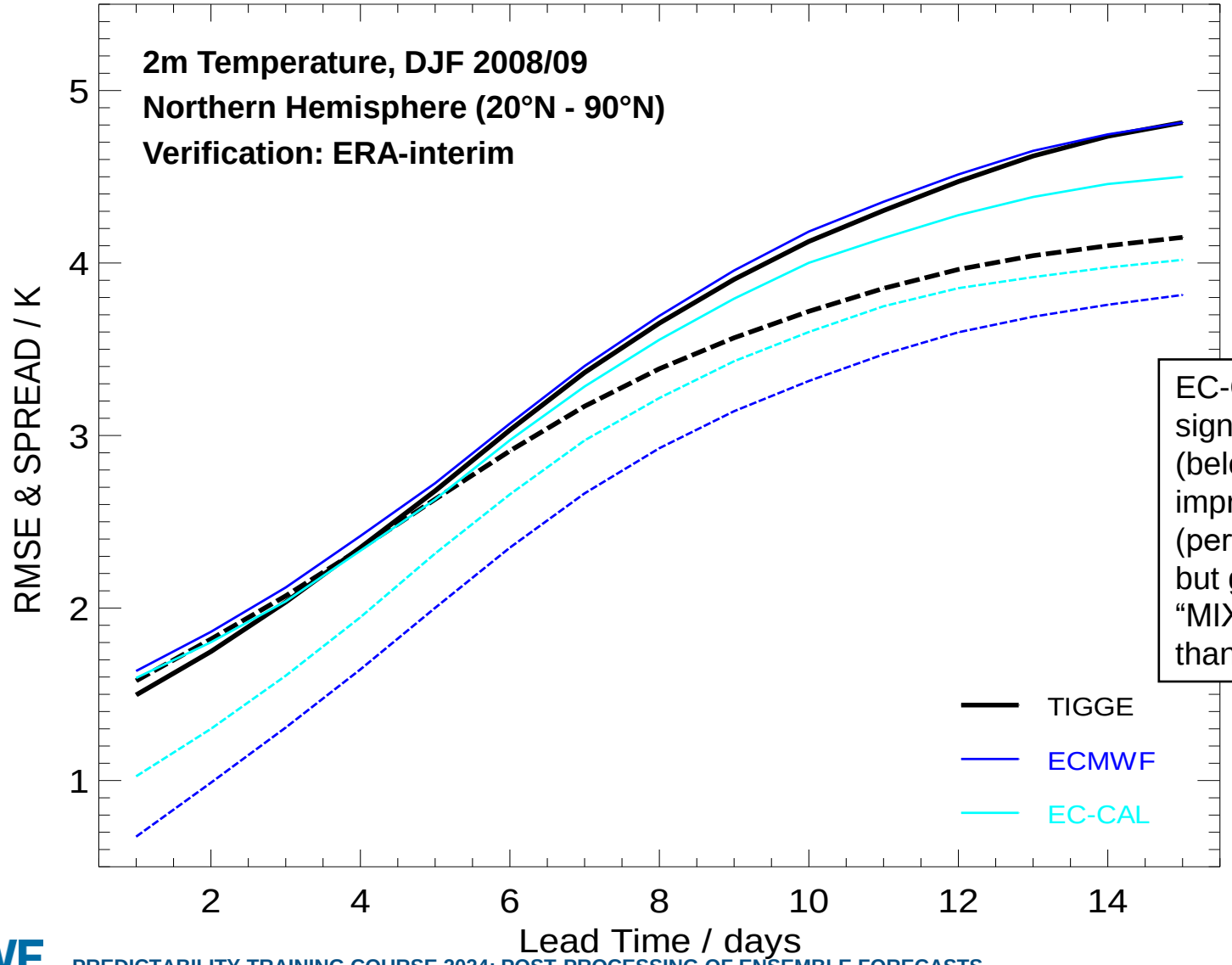
Mechanism behind improvements



RMSE (solid)

SPREAD (dash)

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)

An alternative view of TIGGE …

Reference: Hamill, 2012

- Examining precipitation forecasts over the US
- Four high skill models; compare ECMWF “re-forecast calibrated” with multi-model (no re-forecasts)

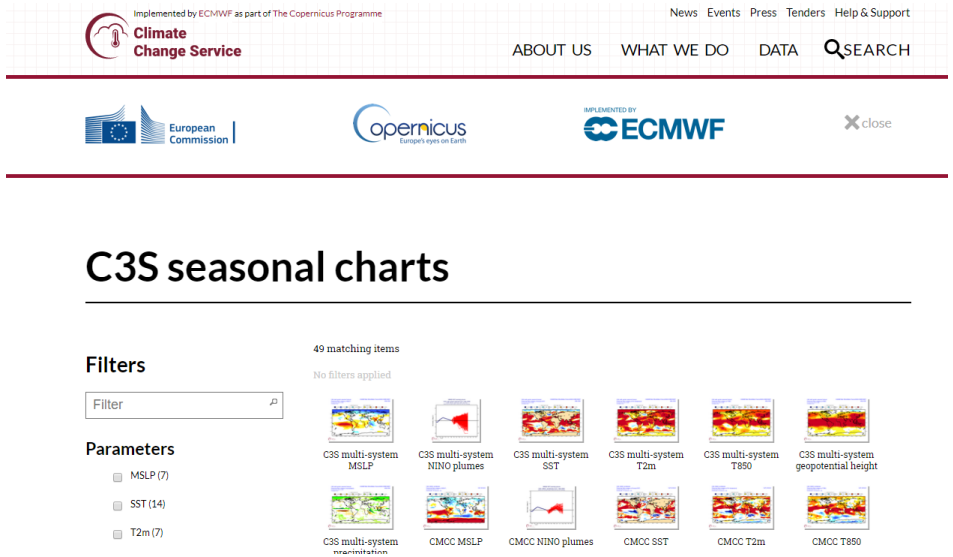
Conclusions:

- “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.”

*PQPF=probabilistic quantative precipitation forecasts

Multi-model seasonal forecasting

- Scientific exploration: DEMETER and ENSEMBLES
- First operational system: EUROSIP
- New incarnation: C3S from COPERNICUS

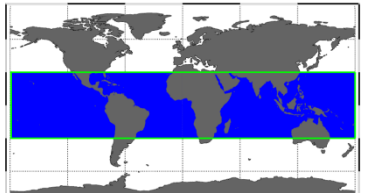


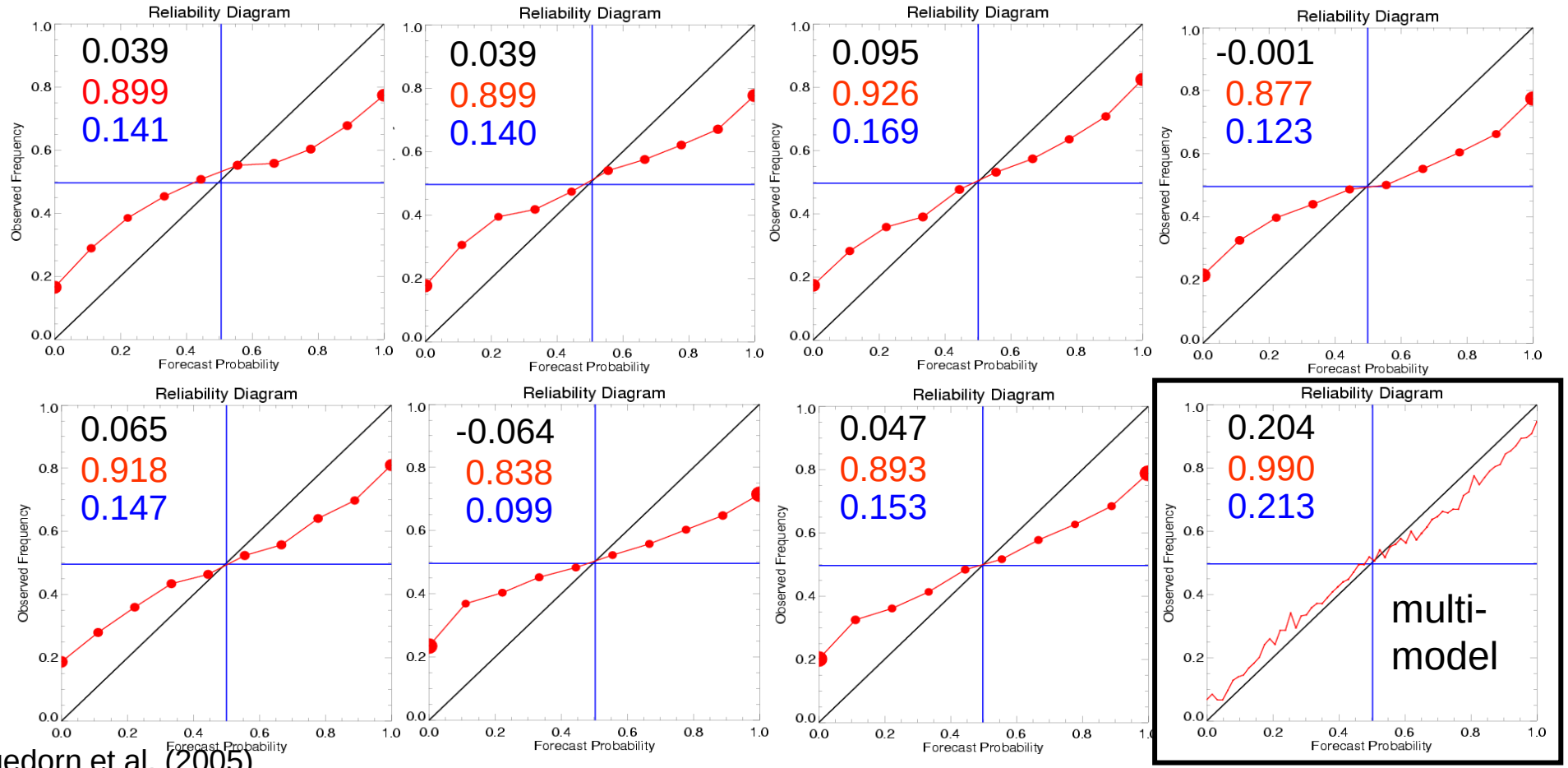
DEMETER: multi-model vs single-model

BSS
Rel-Sc
Res-Sc

Reliability diagrams (T2m > 0)

1-month lead, start date May, 1980 - 2001



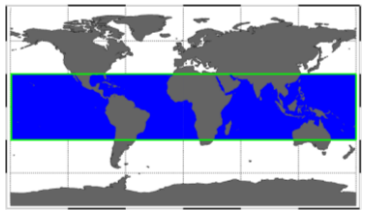


Hagedorn et al. (2005)

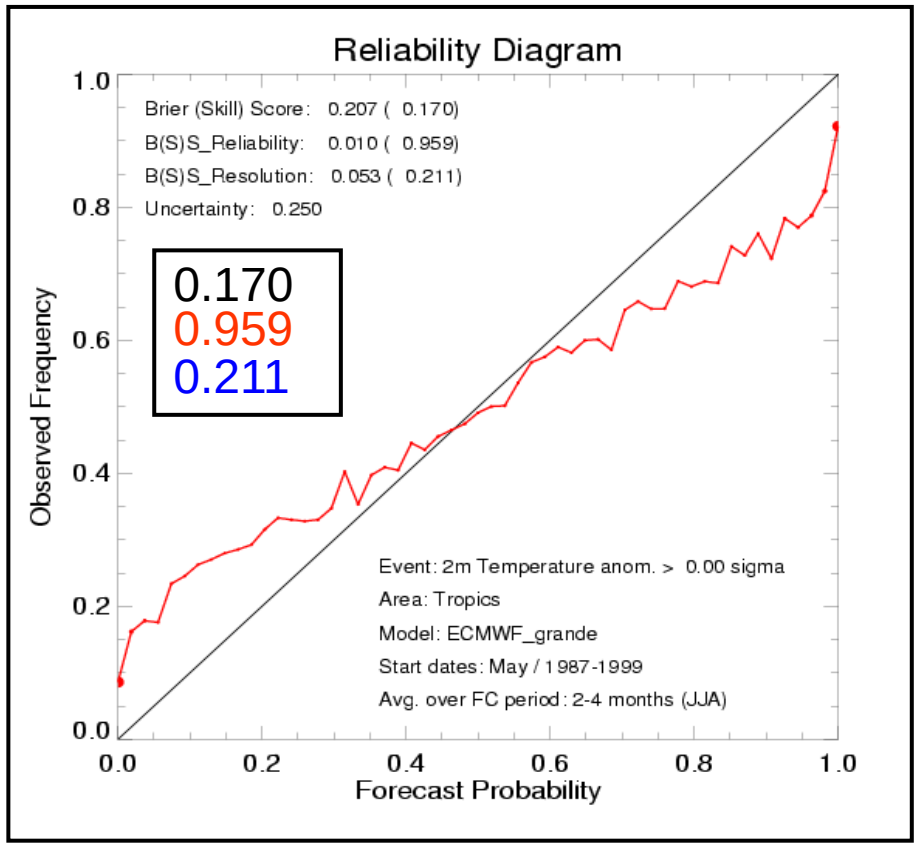
DEMETER: checking impact of ensemble size

BSS
Rel-Sc
Res-Sc

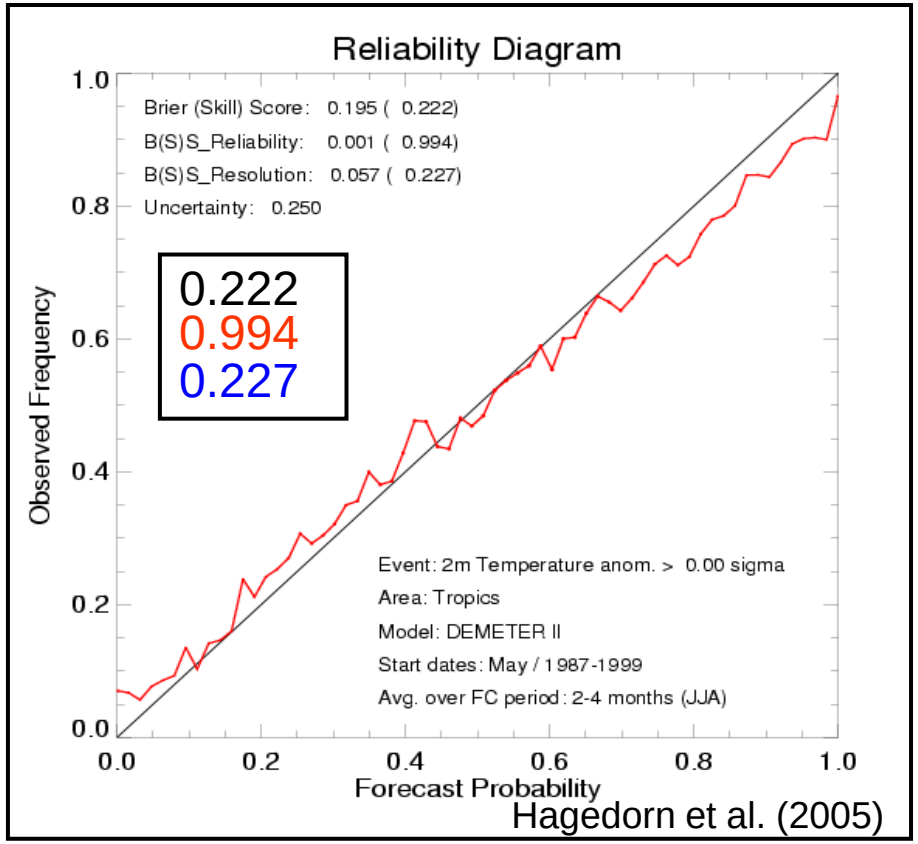
Reliability diagrams (T2m > 0)
1-month lead, start date May, 1987 - 1999



single-model [54 members]



multi-model [54 members]

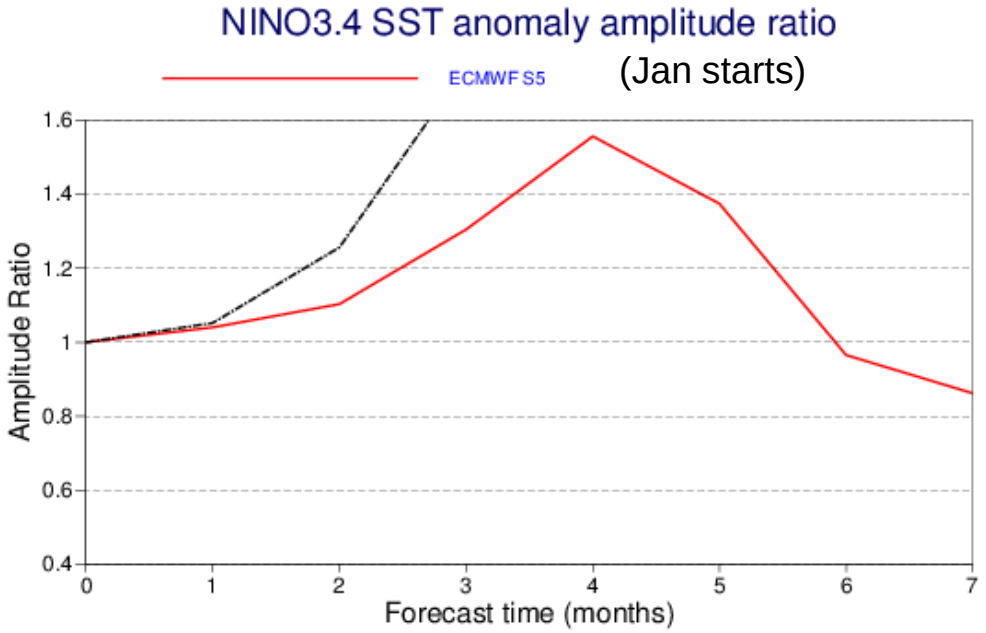


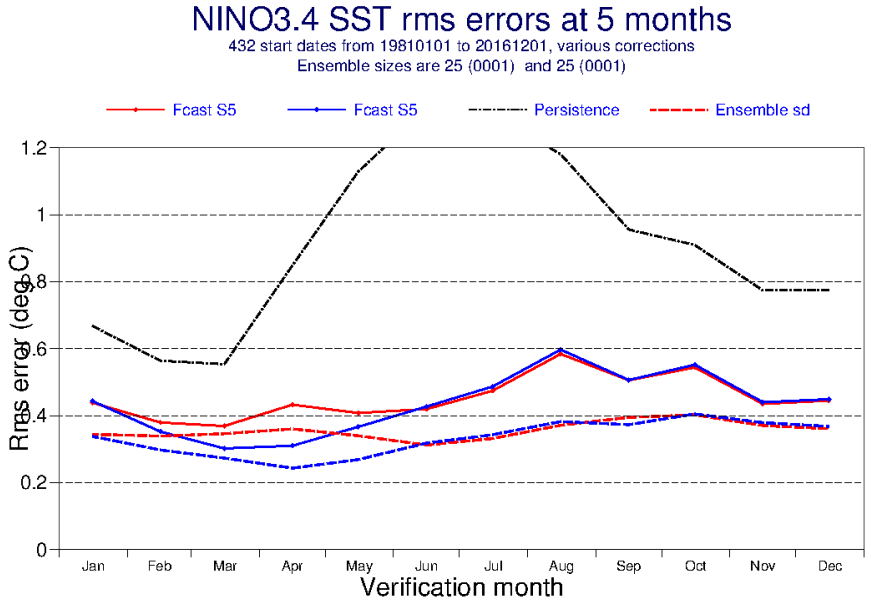
Hagedorn et al. (2005)

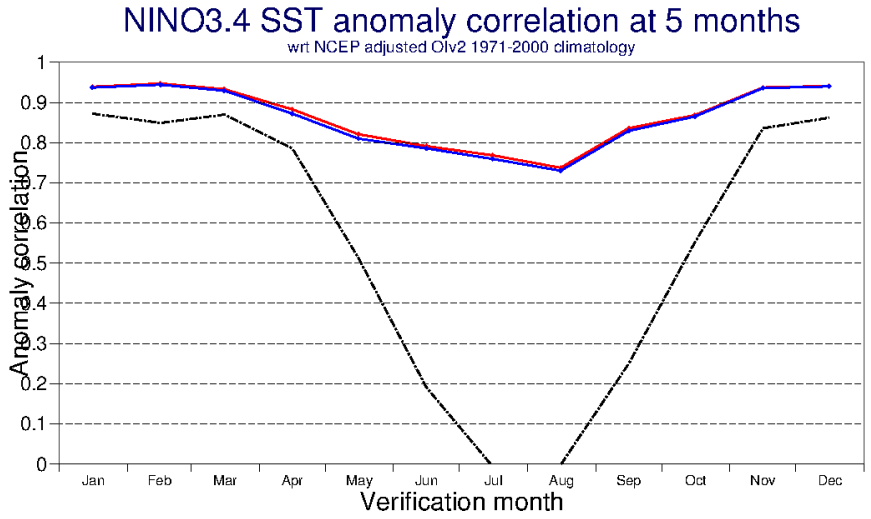
ENSO Variance adjustment

This very simple calibration scales the forecast climatological variance to match the observed climatological variance. The scaling is seasonally dependent. This calibration can substantially improve forecast products (and their verification scores). This calibration was used for our previous system but was turned off in SEAS5. It is used in C3S Nino plots.

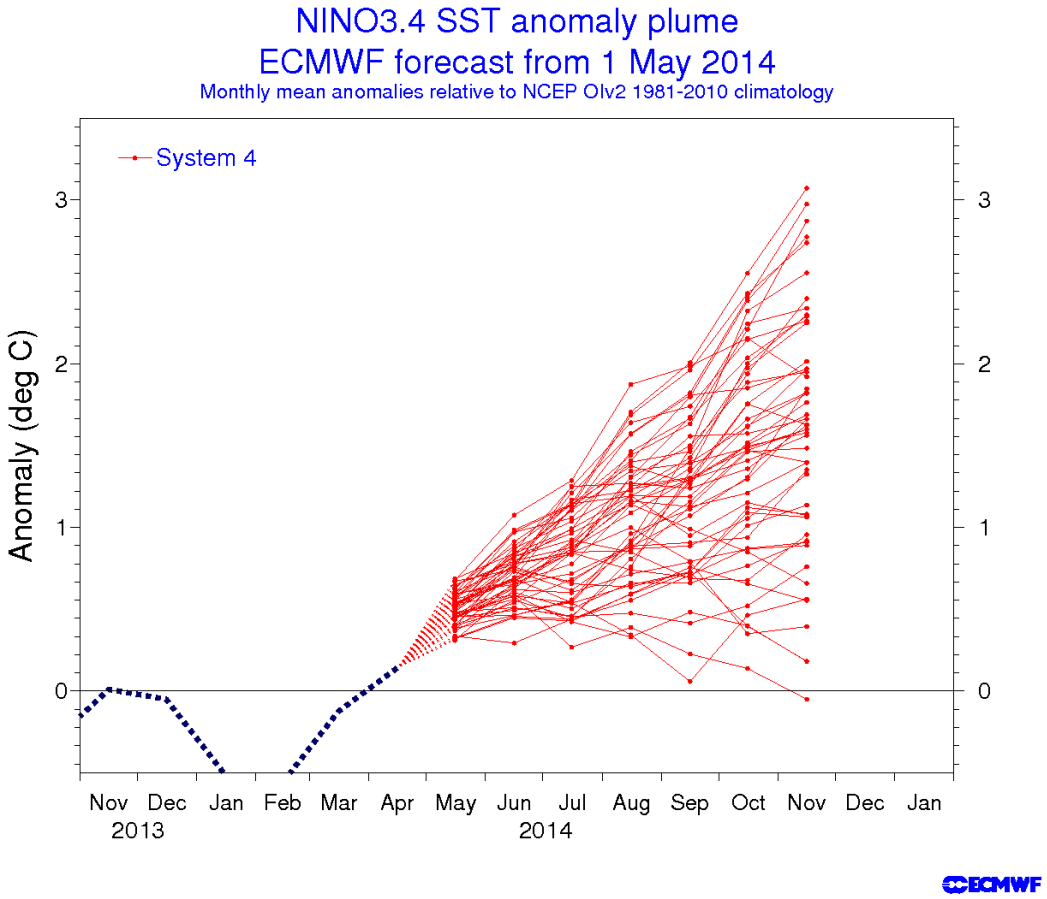
SEAS5 verification includes the amplitude ratio, which should be used *a posteriori* to interpret the Nino plumes. This is important for forecasts of March, April and May.

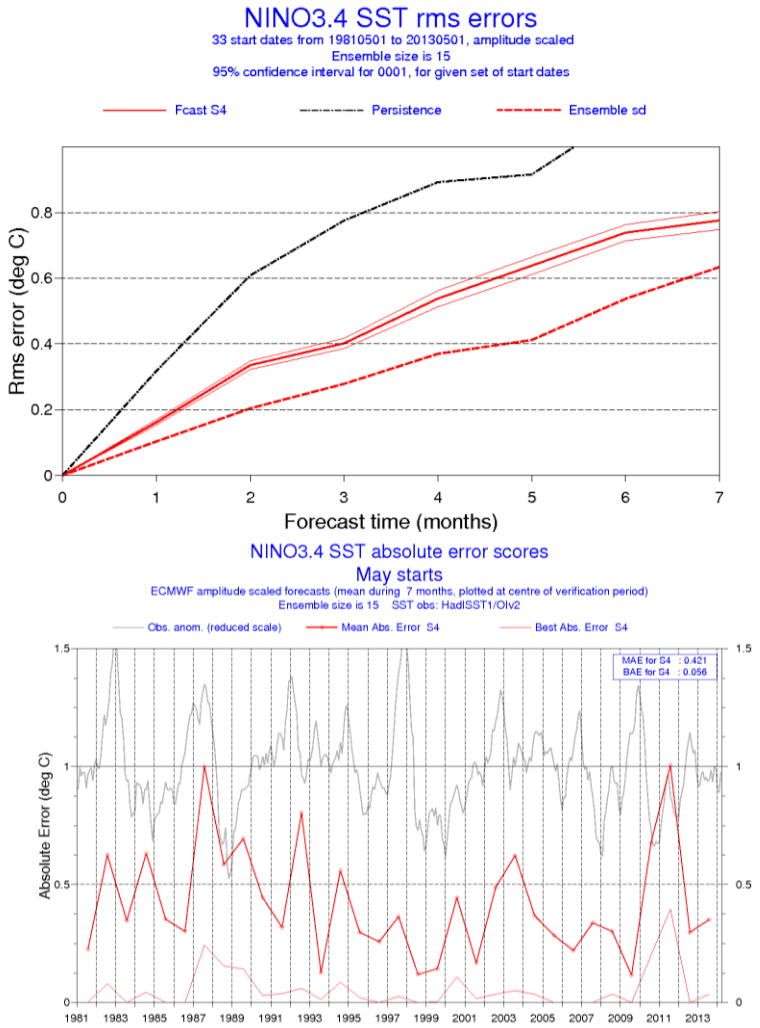






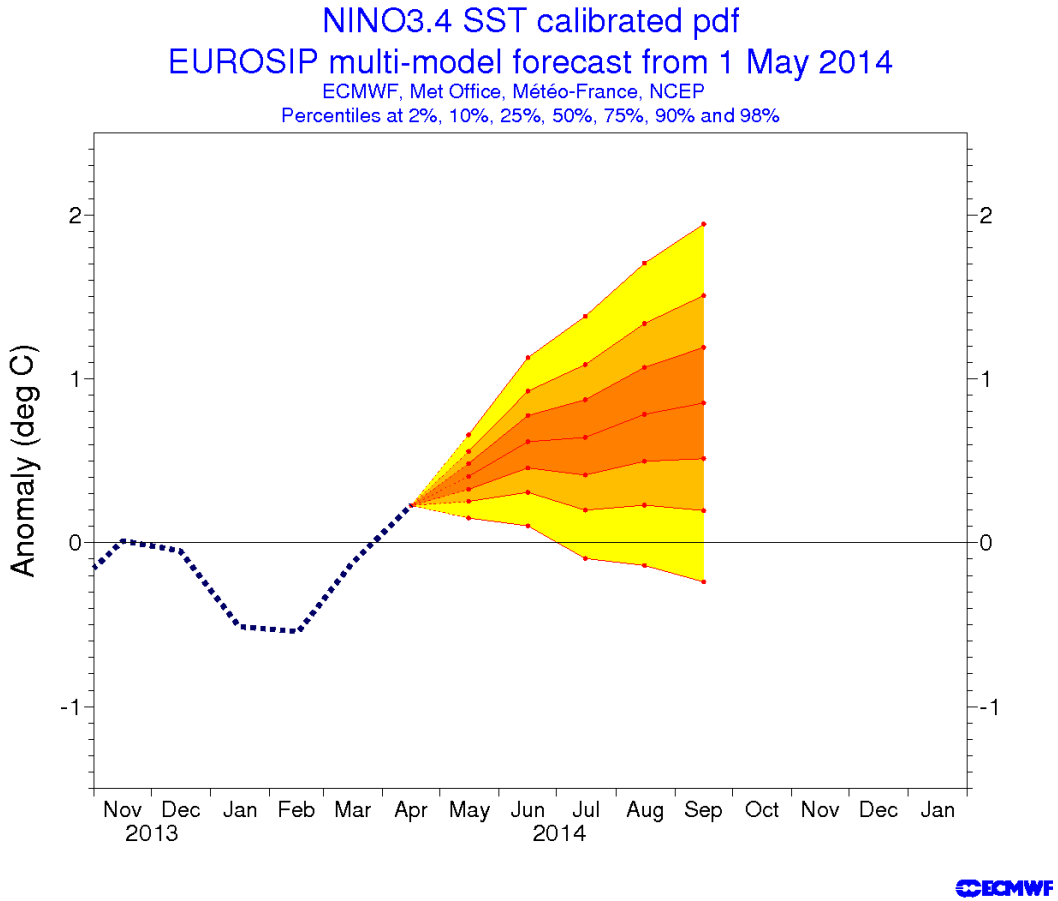
Example of single model Nino SST forecast with simple calibration (bias and variance adjustment)

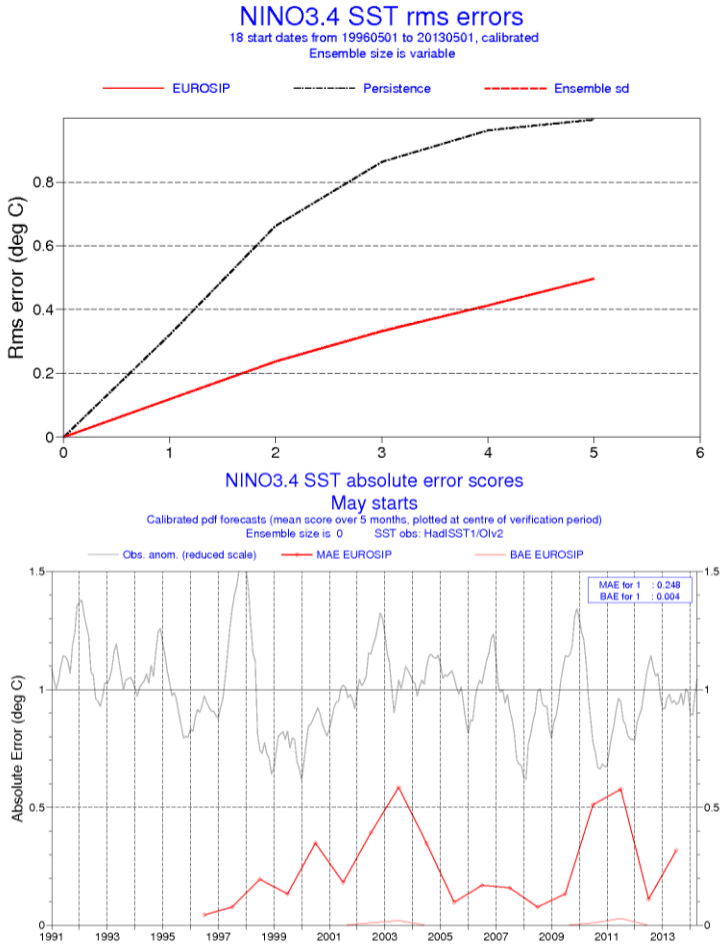




Past performance

Example of multi-model Nino SST forecast with fuller calibration





Past performance

Each model bias corrected and variance adjusted; spread of multi-model combination than adjusted to match past performance

Summary on multi-model and calibration

- What are the relative benefits/costs of both approaches?
  - Both multi-model and a reforecast calibration approach improve probabilistic predictions, in particular for (biased and under-dispersive) near-surface parameters
- What are the mechanisms behind the improvements?
  - Re-forecast calibration is effective at correcting local mis-representations in the model, and can estimate forecast uncertainty well, particularly in the medium-range
  - Multi-model approach can reduce forecast error as well as increase spread; it is particularly powerful at longer leads; sometimes it pays to be selective in which models are used
- Which is the “better” approach?
  - A combination of the two approaches is the most powerful, for example in C3S seasonal
  - For the medium range, a single well-tuned ensemble system is enough, and is easier to implement operationally

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
- ECMWF leaves most calibration to its users
- Calibration cannot guarantee reliable seasonal forecasts

References and further reading

- Glahn, H. R., & Lowry, D. A., 1972: The Use of Model Output Statistics (MOS) in Objective Weather Forecasting, *Journal of Applied Meteorology and Climatology*, **11**(8), 1203-1211

- 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.

- Hamill, T.M. and M. Scheuerer, 2018: Probabilistic Precipitation Forecast Postprocessing Using Quantile Mapping and Rank-Weighted Best-Member Dressing. *Mon. Wea. Rev.*, **146**, 4079-4098.

- 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.