NOAA Earth System
Research Laboratory

Calibrated probabilistic 2-meter temperature and precipitation forecasts using GFS and ECMWF reforecasts

Tom Hamill and Jeff Whitaker

NOAA / ESRL / PSD, Boulder, CO

Renate Hagedorn

ECMWF, Reading, England

References for GFS reforecast calibration

Hamill, T. M., J. S. Whitaker, and X. Wei, 2003: Ensemble re-forecasting: improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.*, **132**, 1434-1447.
http://www.cdc.noaa.gov/people/tom.hamill/reforecast_mwr.pdf

Hamill, T. M., J. S. Whitaker, and S. L. Mullen, 2005: Reforecasts, an important dataset for improving weather predictions. *Bull. Amer. Meteor. Soc.*, **87**, 33-46.
http://www.cdc.noaa.gov/people/tom.hamill/refcst_bams.pdf

Whitaker, J. S, F. Vitart, and X. Wei, 2006: Improving week two forecasts with multi-model re-forecast ensembles. *Mon. Wea. Rev.*, **134**, 2279-2284.
http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/Manuscripts/multimodel.pdf

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev.*, in press.
http://www.cdc.noaa.gov/people/tom.hamill/reforecast_analog_v2.pdf

Wilks, D. S., and T. M. Hamill, 2006: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.*, in press. http://www.cdc.noaa.gov/people/tom.hamill/WilksHamill_emos.pdf

Hamill, T. M. and J. S. Whitaker, 2006: White Paper. "Producing high-skill probabilistic forecasts using reforecasts: implementing the National Research Council vision." Available at
http://www.cdc.noaa.gov/people/tom.hamill/whitepaper_reforecast.pdf .

Hamill, T. M., and J. S. Whitaker, 2007: Ensemble calibration of 500-hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. Mon. Wea. Rev., in press. Available at
http://www.cdc.noaa.gov/people/tom.hamill/Calibration_z500t850t2m_v2.pdf

NOAA’s reforecast data set

- **Model**: T62L28 NCEP GFS, circa 1998

- **Initial States**: NCEP-NCAR Reanalysis II plus 7 +/- bred modes.

- **Duration**: 15 days runs every day at 00Z from 19781101 to now. (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst/week2*).

- **Data**: Selected fields (winds, hgt, temp on 5 press levels, precip, t2m, u10m, v10m, pwat, prmsl, rh700, heating). NCEP/NCAR reanalysis verifying fields included (Web form to download at *http://www.cdc.noaa.gov/reforecast*). Data saved on 2.5-degree grid.

- Here, use only the subset of data overlapping with ECMWF reforecast data set.

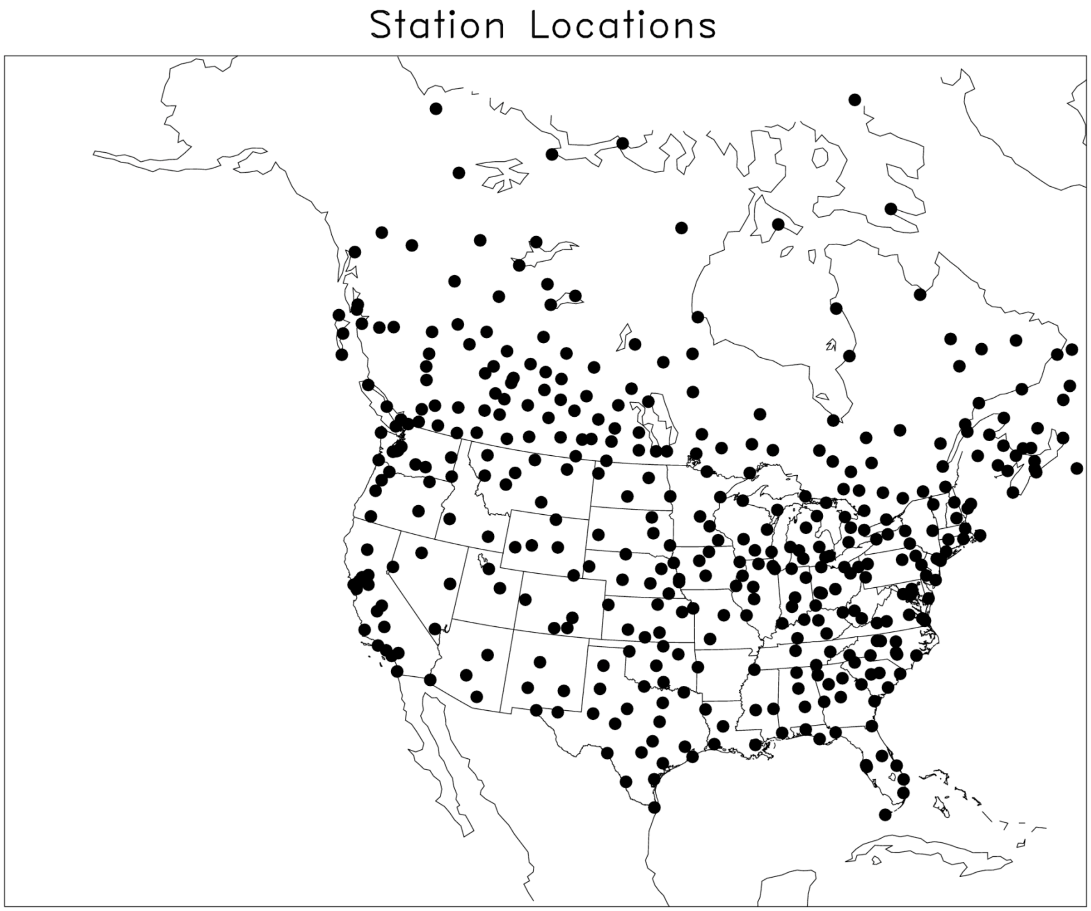
ECMWF’s reforecast data set

- **Model**: 2005 version of ECMWF model; T255 resolution.
- **Initial Conditions**: 15 members, ERA-40 analysis + singular vectors
- **Dates of reforecasts**: 1982-2001, Once-weekly reforecasts from 01 Sep - 01 Dec, 14 total. So, 20*14 ensemble reforecasts = 280 samples.
- **Data** sent to NOAA / ESRL : T_{2M} ensemble over most of North America, excluding Alaska. Saved on 1-degree lat / lon grid. Forecasts to 10 days lead.

Questions

- Is calibration using a large reforecast data set as useful with a current state-of-the-art model as with 10-year old model?
- How much benefit can be achieved by calibrating forecasts with smaller training data sets?
- How different is calibrating precipitation than calibrating temperature?

Observation locations for 2-meter temperature calibration



Uses stations from NCAR's DS472.0 database that have more than 96% of the yearly records available, and overlap with the domain that ECMWF sent us.

Calibration Procedure: “NGR”

“Non-homogeneous Gaussian Regression”

- **Reference**: Gneiting et al., *MWR*, **133**, p. 1098
- **Predictors**: ensemble mean and ensemble spread
- **Output**: mean, spread of calibrated normal distribution

$$f^{CAL}\left(\bar{\mathbf{x}}, \sigma\right) \sim N\left(a + b\bar{\mathbf{x}}, c + d\sigma\right)$$

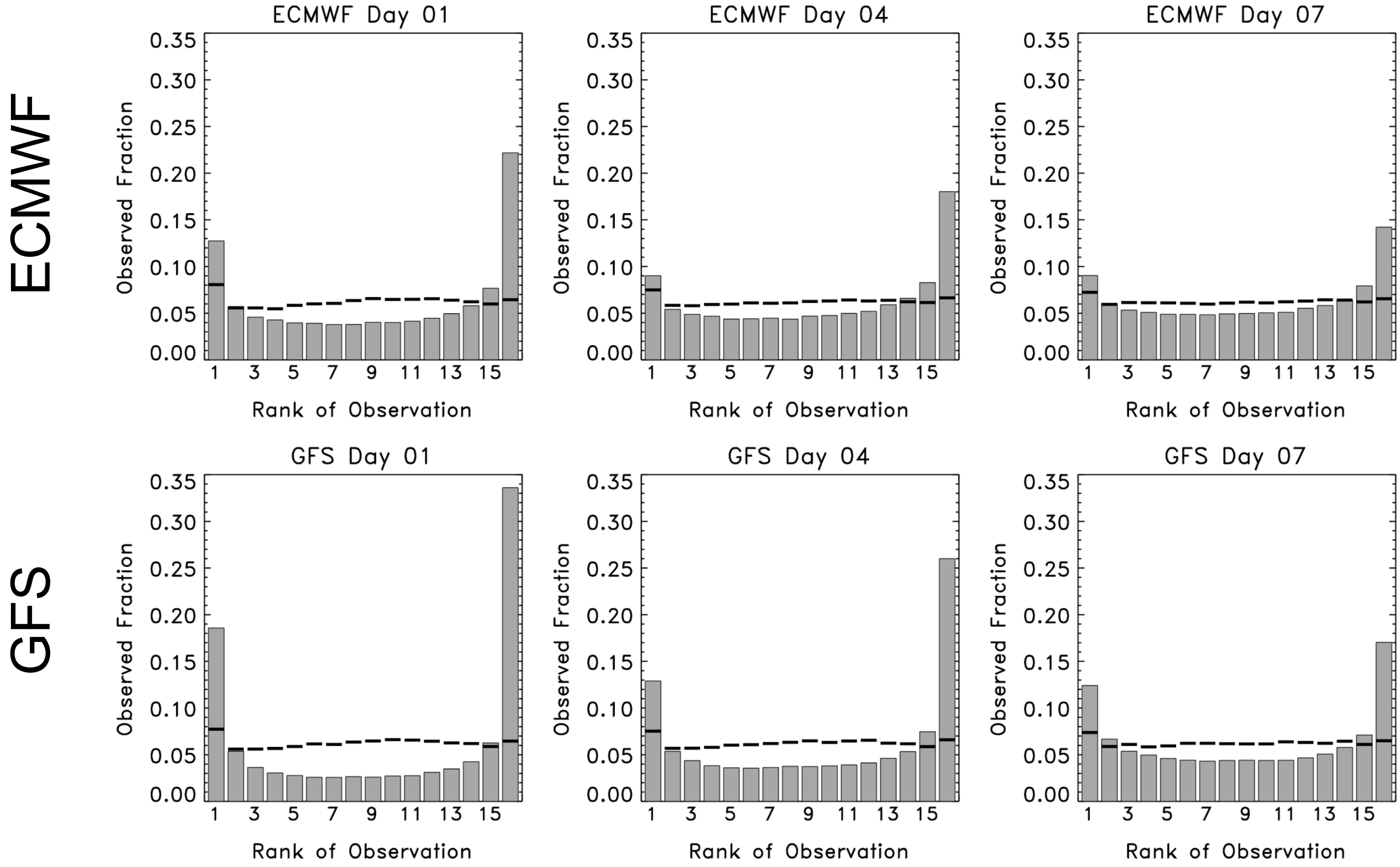
- **Advantage**: leverages possible spread/skill relationship appropriately. Large spread/skill relationship, c ≈ 0.0, *d* ≈1.0. Small, *d* ≈ 0.0
- **Disadvantage**: iterative method, slow…no reason to bother (relative to using simple linear regression) if there’s little or no spread/skill relationship.

Training Data for
Non-homogeneous Gaussian Regression
(all cross validated)

- **01 Sep**: *01 Sep*, 08 Sep, 15 Sep
- **08 Sep**: 01 Sep, *08 Sep*, 15 Sep, 22 Sep
- **15 Sep**: 01 Sep, 08 Sep, *15 Sep*, 22 Sep, 29 Sep
-
-
-
- **17 Nov**: 03 Nov, 10 Nov, *17 Nov*, 24 Nov, 01 Dec
- **24 Nov**: 10 Nov, 17 Nov, *24 Nov*, 01 Dec
- **01 Dec**: 17 Nov, 24 Nov, *01 Dec*

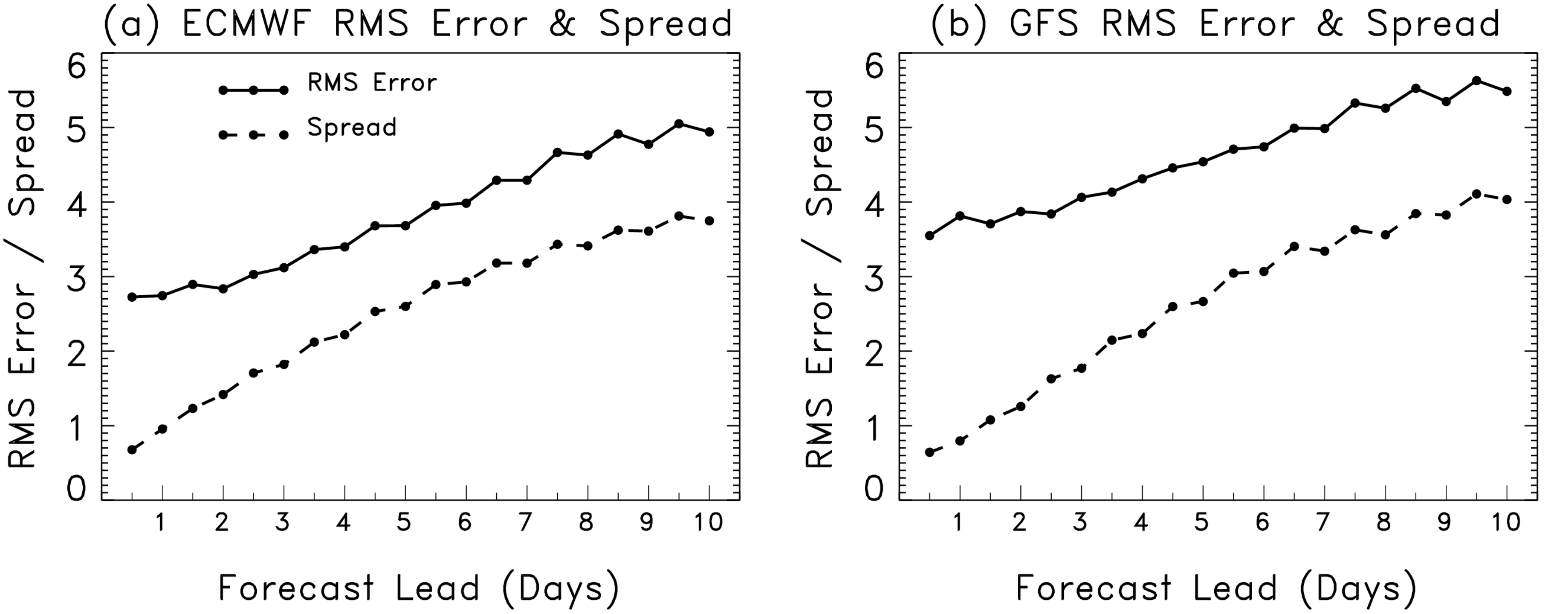
Use a centered training data set for weeks 3 - 12, uncentered for weeks 1, 2, 13, and 14

Rank histograms



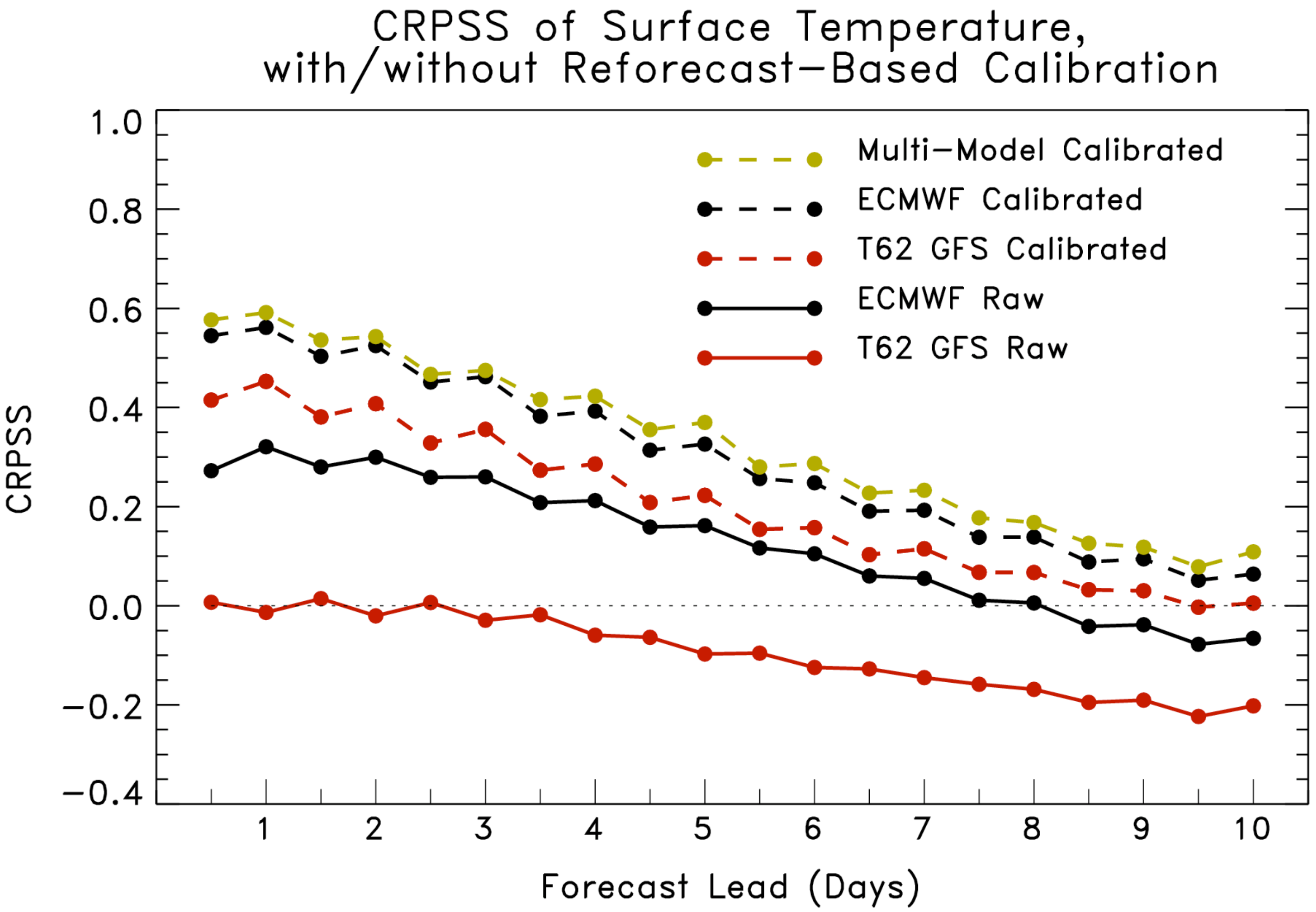
Members randomly perturbed by 1.0K to account for observation error; probably a bit small for GFS on its coarser 2.5^{o} grid, which would make their histograms slightly more uniform. Ref: Hamill, *MWR*, **129**, p. 556**.**

Spread vs. RMS error



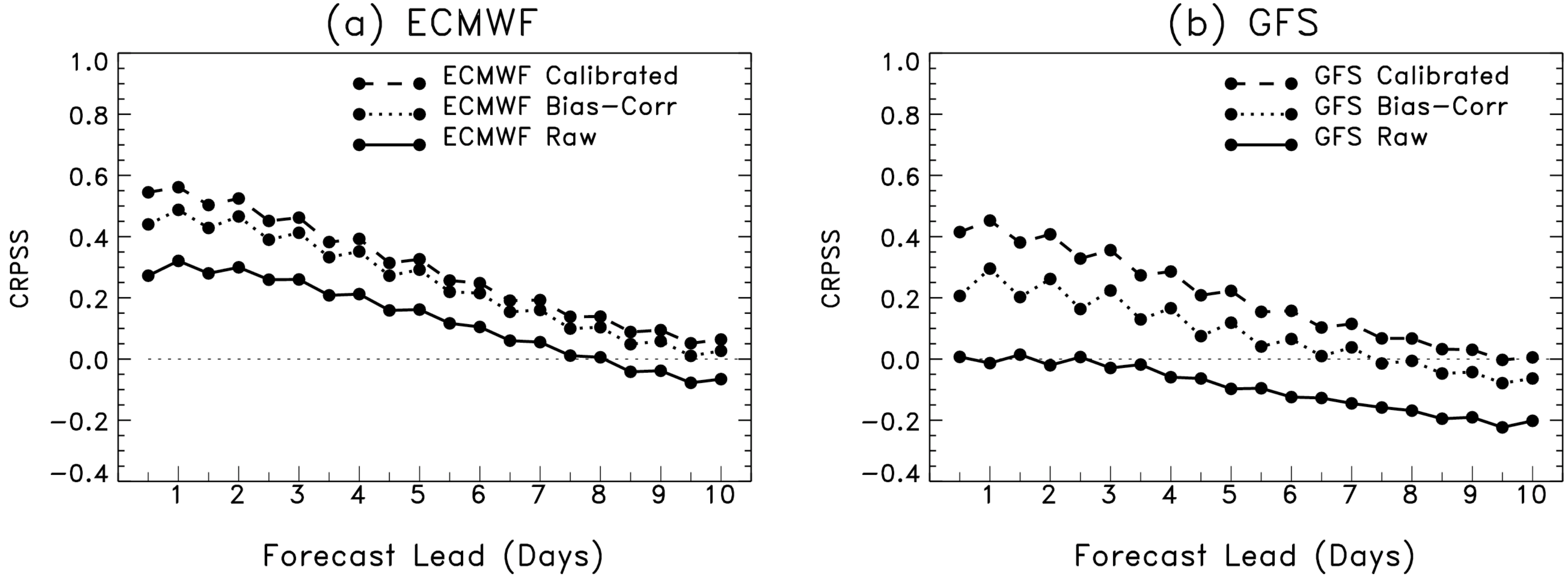
While rank histograms of GFS and ECMWF were not remarkably different, here one can see that RMS errors of the ECMWF system are much lower. Note that surface temp has much bigger inconsistency than is typically shown for, say, 500 hPa geopotential.

Skill after calibration



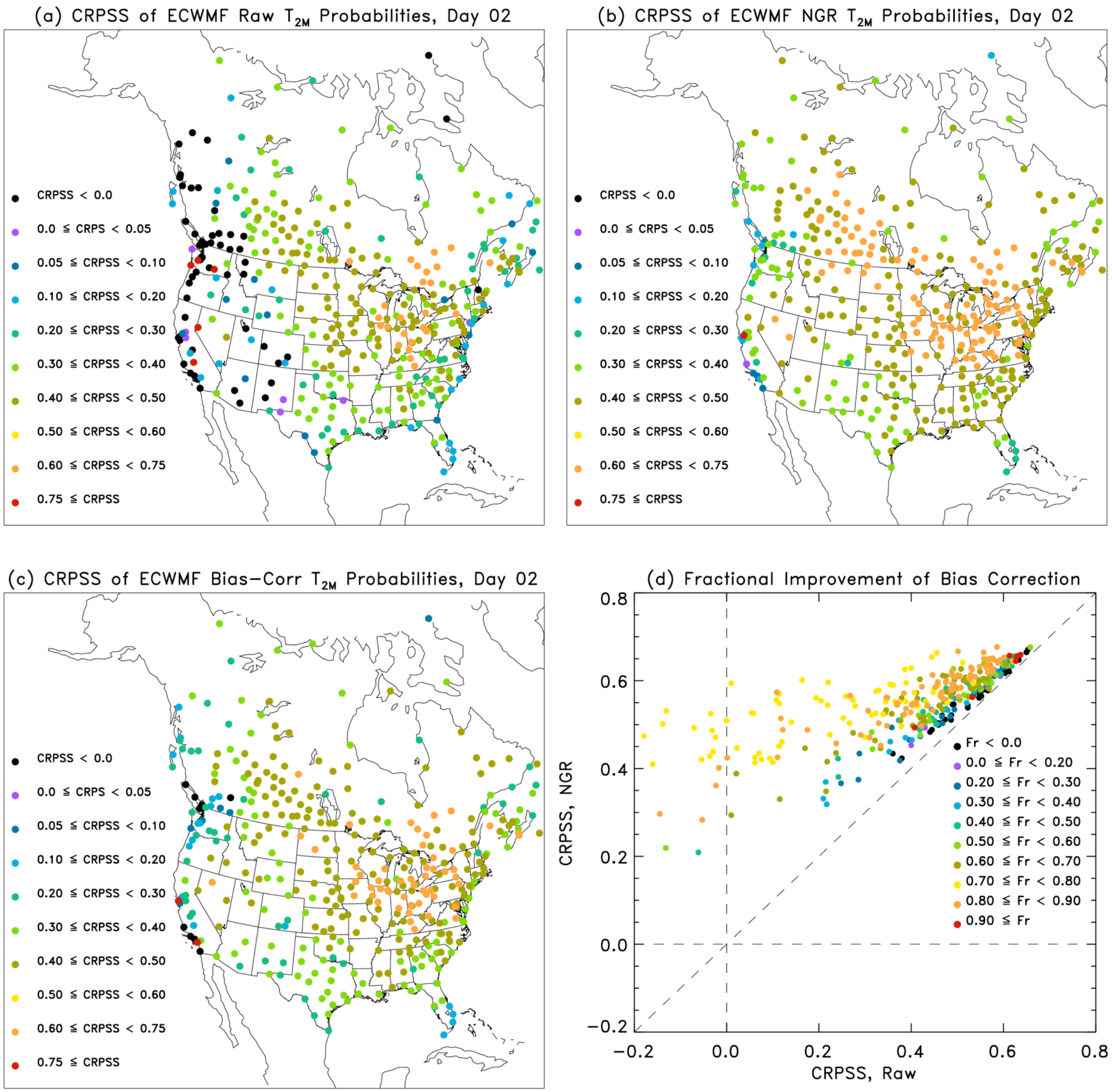
Notes: (1) GFS calibrated > ECMWF raw; (2) Still significant benefit from calibration of ECMWF; (3) multi-model slightly better than ECMWF.

How much from simple bias correction?



For ECMWF, ~ 60 percent of total improvement at short leads, ~70 percent at longer leads.

Where was skill improvement largest?



(a) CRPSS of ECWMF Raw T_{2M} Probabilities, Day 02

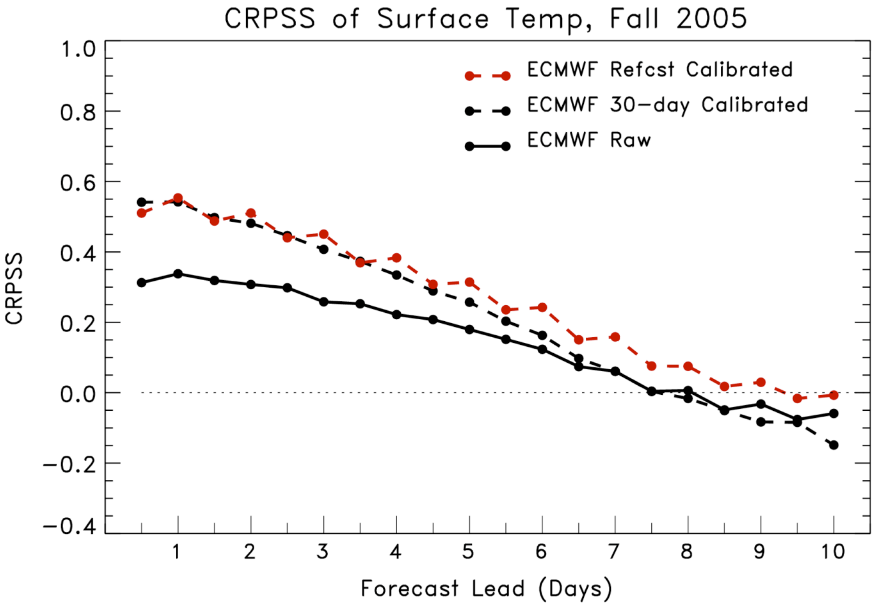
(b) CRPSS of ECWMF NGR T_{2M} Probabilities, Day 02

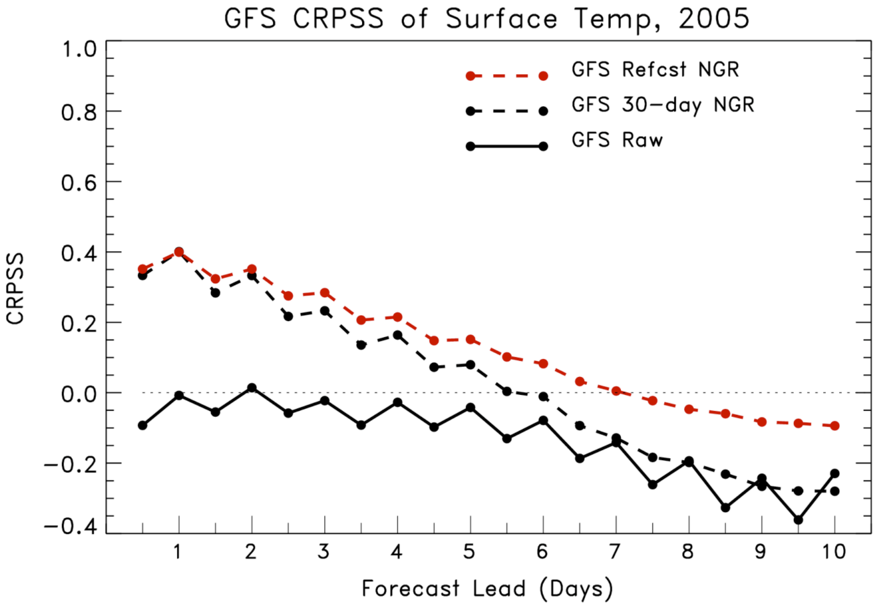
(c) CRPSS of ECWMF Bias-Corr T_{2M} Probabilities, Day 02

(d) Fractional Improvement of Bias Correction

Largest improvements over complex terrain of western US and Canada. Calibration homogenizes skill, bringing up lower-performing stations more than higher-performing ones. Bigger effect of bias correction where raw forecasts are particularly poor.

How much from short 30-day training data sets?



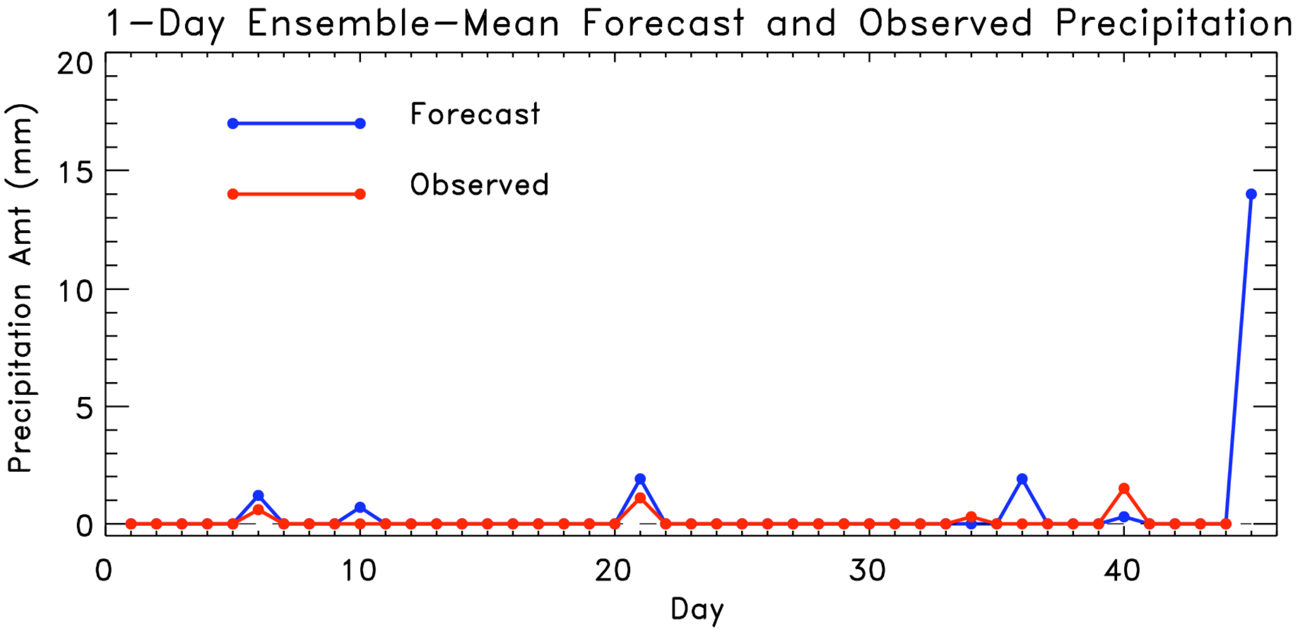


Prior 30-days improves forecasts about as much as long training data set at short leads, but not as much at longer leads.

Calibration of probabilistic precipitation forecasts

Expect precip. calibration to be tougher than temperature.

Want lots of old forecast cases that were similar to today's forecast. Then the difference between the observed and forecast on those days can be used to calibrate today's forecast.



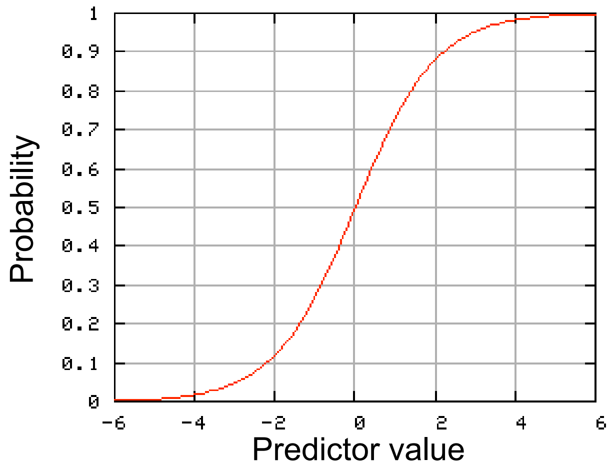
Forecast Calibration: Logistic Regression

Given predictors $x_1, \dots, x_N$ (such as the ensemble-mean and spread), find regression coefficients

$\beta_0, \beta_1, \dots, \beta_N$ for the equation

$$P(obs > T) = 1. - \frac{1}{1 - \exp\left(\beta_0 + \beta_1 x_1 + \dots + \beta_N x_N\right)}$$

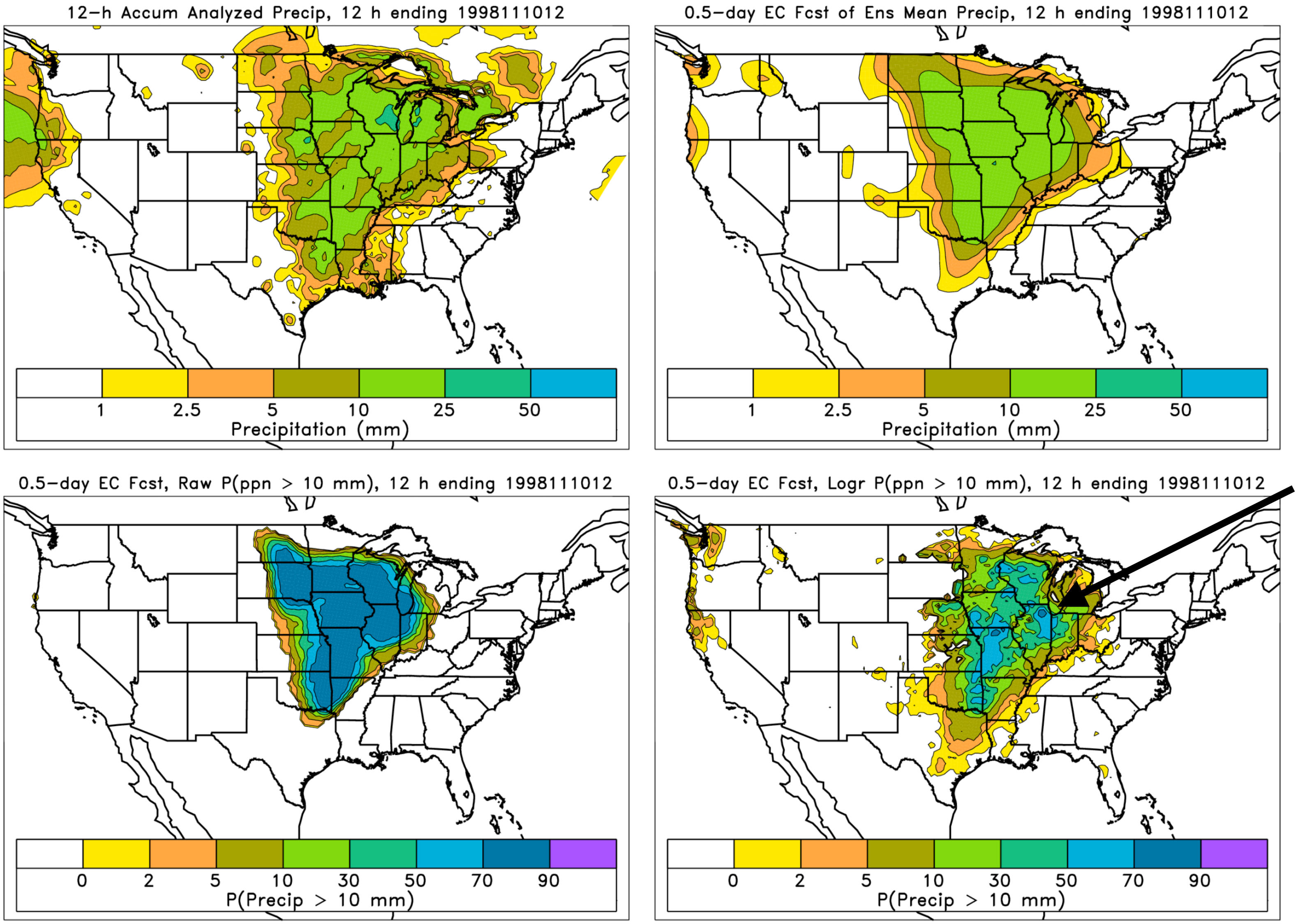
This generates an S-shaped curve (here for one predictor)



Logistic regression for precipitation: training method

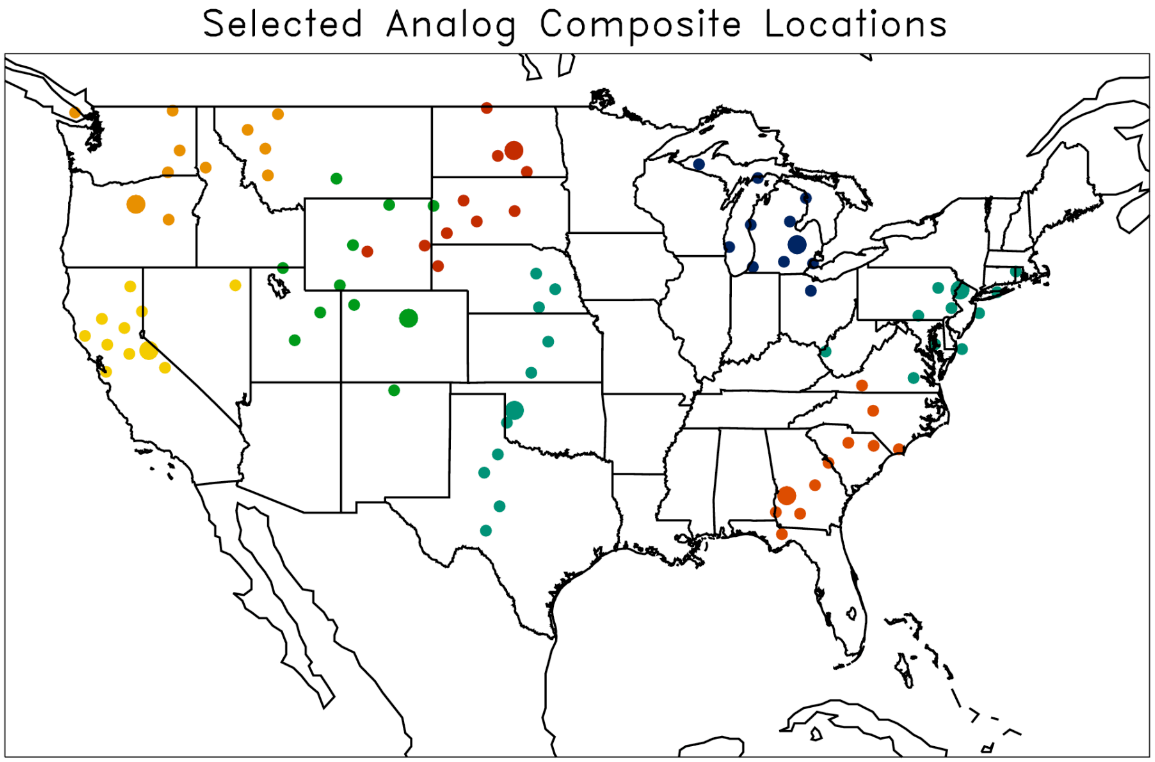
- Cross-validate (for example, 1982 forecasts use 1981, 1983-2001).
- Use all fall season data together, unlike temperature (1 Sep forecasts use 1 Sep - 1 Dec training data). [seasonal biases assumed less important than training sample size]
- Sole predictor: (ens. mean precip)$^{0.25}$

Not enough ECMWF training data to provide stable logistic regression coefficients.


12−h Accum Analyzed Precip, 12 h ending 1998111012
1
2.5
5
10
25
50
Precipitation (mm)
0.5−day EC Fcst of Ens Mean Precip, 12 h ending 1998111012
1
2.5
5
10
25
50
Precipitation (mm)
0.5−day EC Fcst, Raw P(ppn > 10 mm), 12 h ending 1998111012
0
2
5
10
30
50
70
90
P(Precip > 10 mm)
0.5−day EC Fcst, Logr P(ppn > 10 mm), 12 h ending 1998111012
0
2
5
10
30
50
70
90
P(Precip > 10 mm)

Note lumpiness of probabilities when original field was smooth

Increasing logistic regression sample size by compositing data from different locations



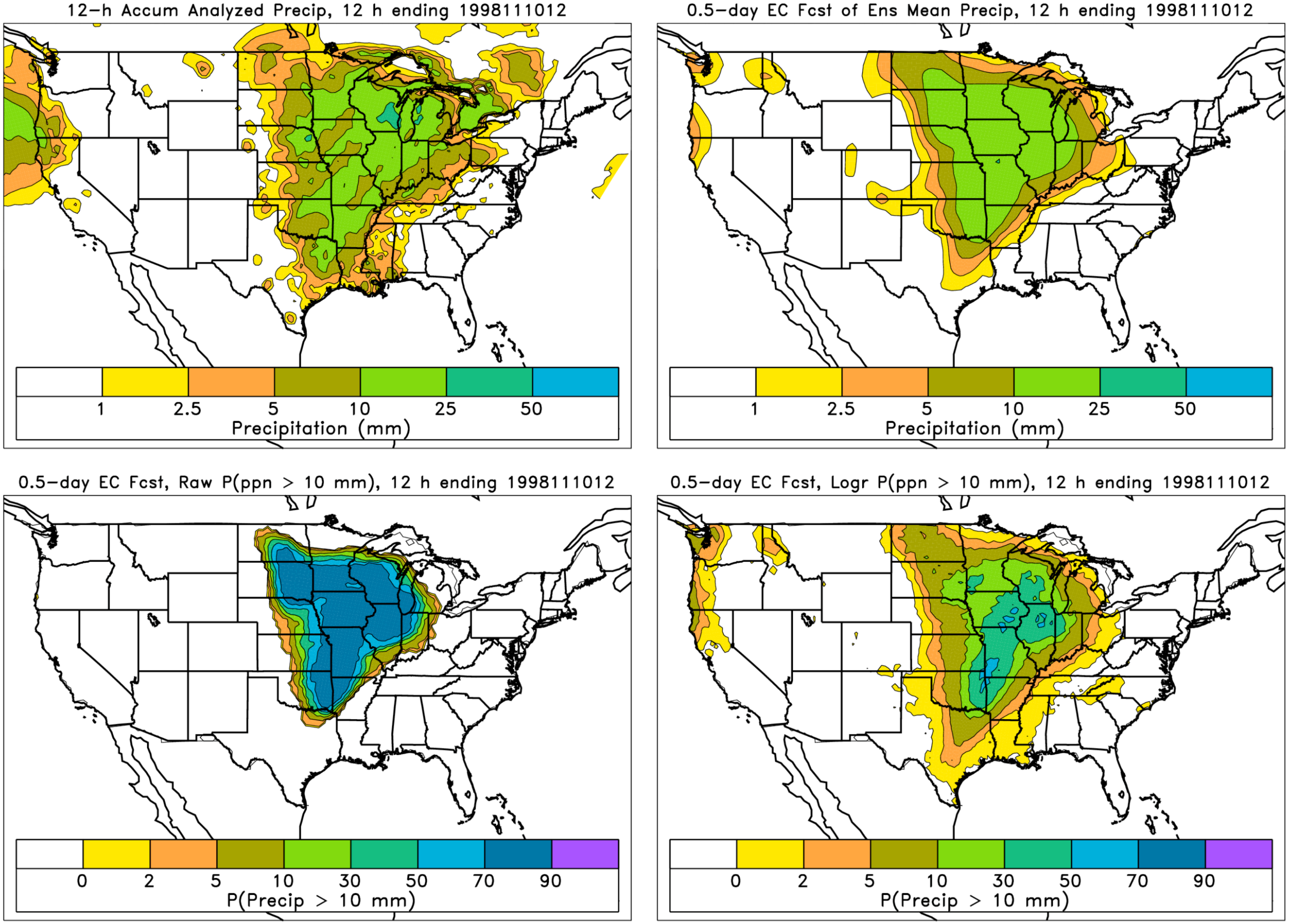
Big dot: location to perform logistic regression.

Small dots: grid points with similar observed climatologies.

Constrained so that the analog composite locations can't be too near to each other.

Sub-optimal (what if forecast climatologies differ? What if forecast/observed correlations differ? These not accounted for in choosing analog locations.)

Calibrated forecasts after compositing



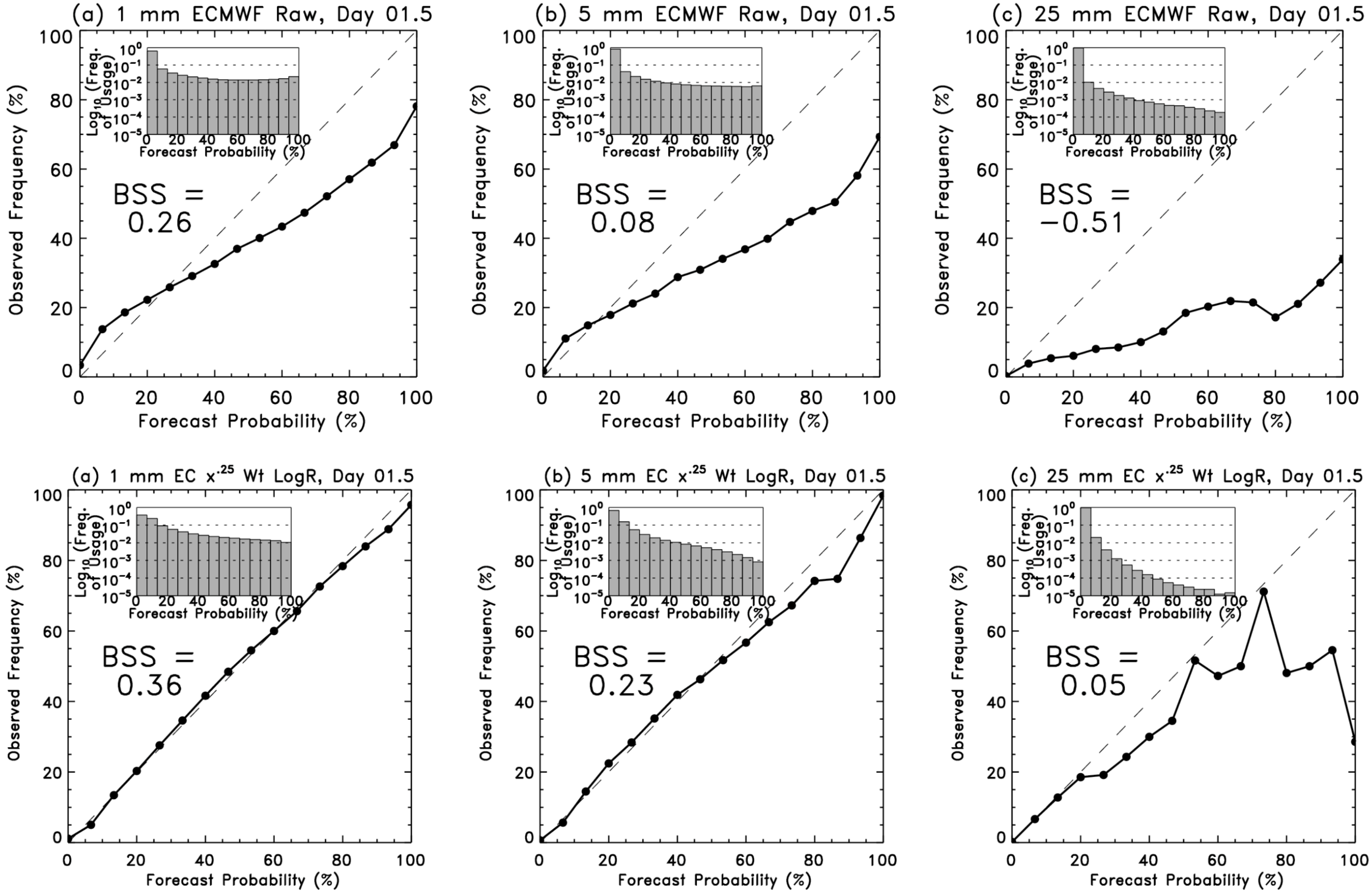
12−h Accum Analyzed Precip, 12 h ending 1998111012

0.5−day EC Fcst of Ens Mean Precip, 12 h ending 1998111012

0.5−day EC Fcst, Raw P(ppn > 10 mm), 12 h ending 1998111012

0.5−day EC Fcst, Logr P(ppn > 10 mm), 12 h ending 1998111012

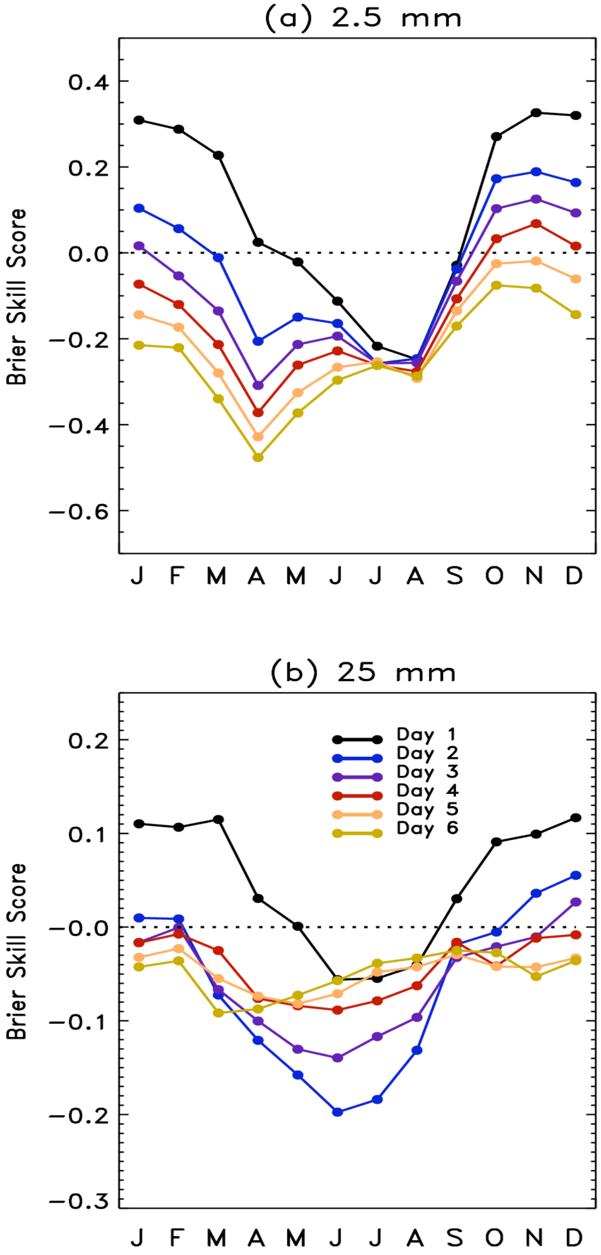
Reliability Diagrams



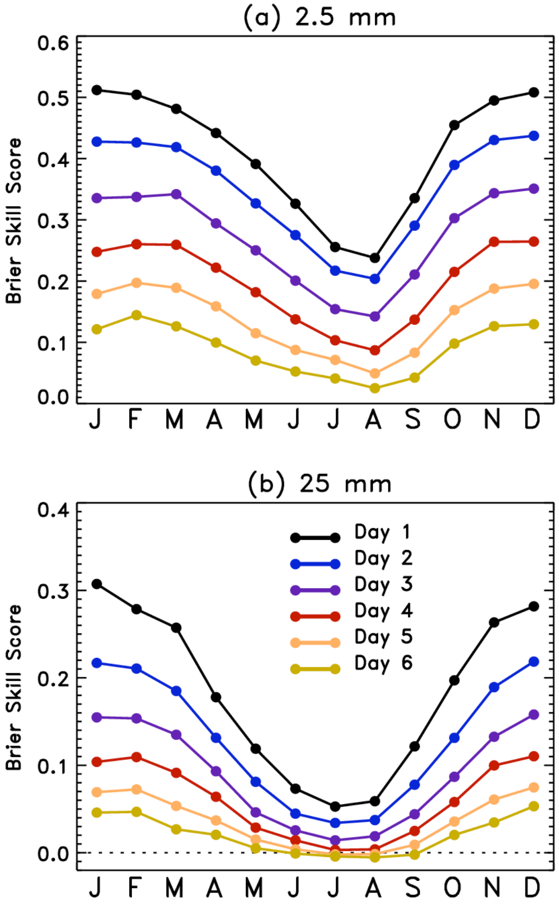
Precipitation calibration - work to be done

- What can be done with small, 30-day training data sets?
- Is relative improvement from ECMWF as large as it was with GFS?

Ensemble Relative Frequency

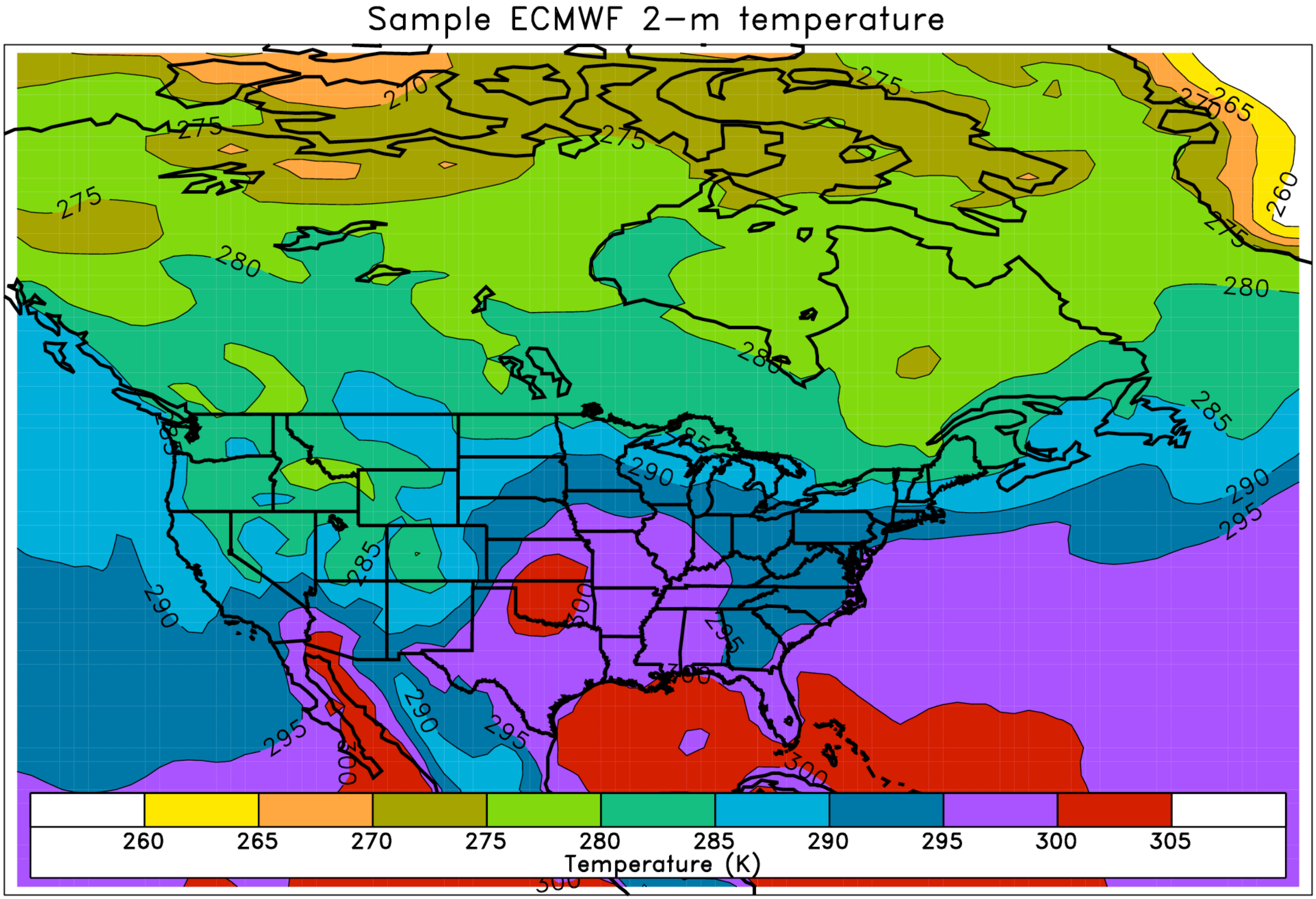


Basic Analog Technique



Verified over 25 years of forecasts; skill scores use conventional method of calculation which may overestimate skill
(Hamill and Juras 2006, QJRMS, Oct).

ECMWF domain sent to us for reforecast tests



Continuous Ranked Probability Score (CRPS) and Skill Score (CRPSS)

$$CRPS^{f}_{i,j,k} = \int_{-\infty}^{+\infty} \left[F_{i,j,k}(y) - F^{o}_{i,j,k}(y) \right]^{2} dy$$

$$i = 1, \ldots, \#\, case\ days$$

$$j = 1, \ldots, \#\, years\ of\ reforecasts$$

$$k = 1, \ldots, \#\, station\ locations$$

$$F_{i,j,k}(y)\ is\ forecast\ CDF\ at\ value\ y$$

$$F^{o}_{i,j,k}(y)\ is\ obs\ CDF\ at\ value\ y\ (Heaviside)$$

$$CRPSS = 1.0 - \frac{\overline{CRPS}^{f}}{\overline{CRPS}^{c}}$$

(This conventional way of calculating CRPSS exaggerates skill if some samples have more climatological spread than others. Will use a modified version where we calculate CRPSS separately for 8 different categories of climatological spread and then average them. See Hamill and Juras, October 2006(C), *QJRMS,* and Hamill and Whitaker (2007) *MWR,* to appear, tinyurl.com/29oy8s)