NOAA Earth System
Research Laboratory

Exploring ensemble forecast calibration issues using reforecast data sets

Tom Hamill and Jeff Whitaker

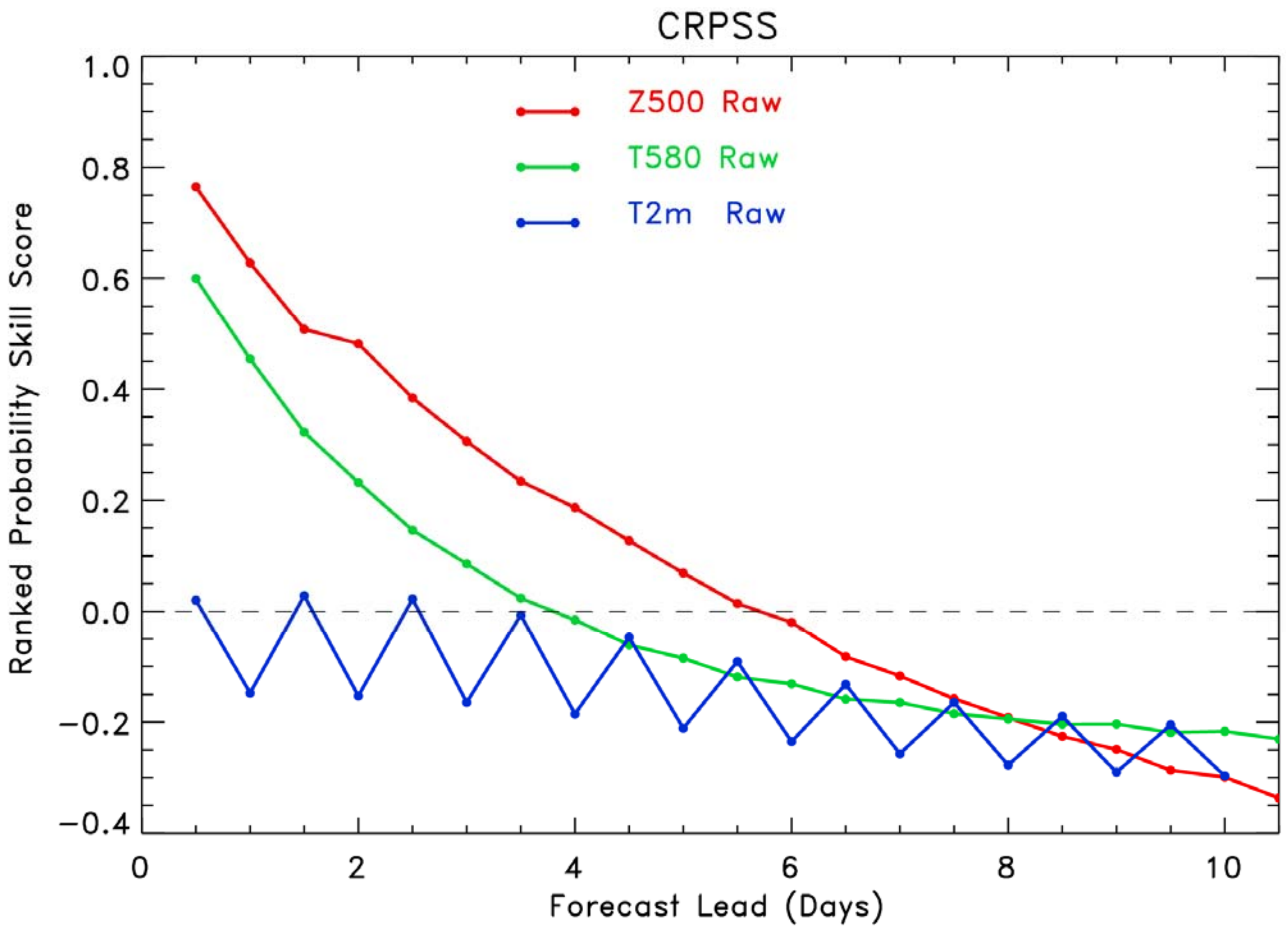
NOAA Earth System Research Lab, Boulder, CO

tom.hamill@noaa.gov ; esrl.noaa.gov/psd/people/tom.hamill

Renate Hagedorn

ECMWF, Reading, England

Skill of 500-hPa Z, 850-hPa T, and 2-m T from raw GFS reforecast ensemble



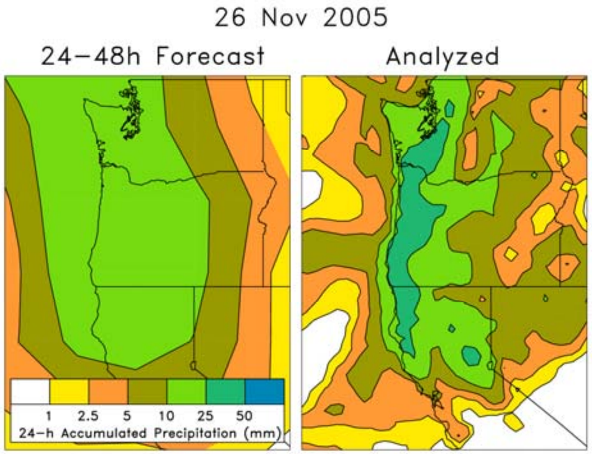
The one variable we probably care about the most, T_{2m}, raw probability forecasts score the worst. Can statistical corrections help?

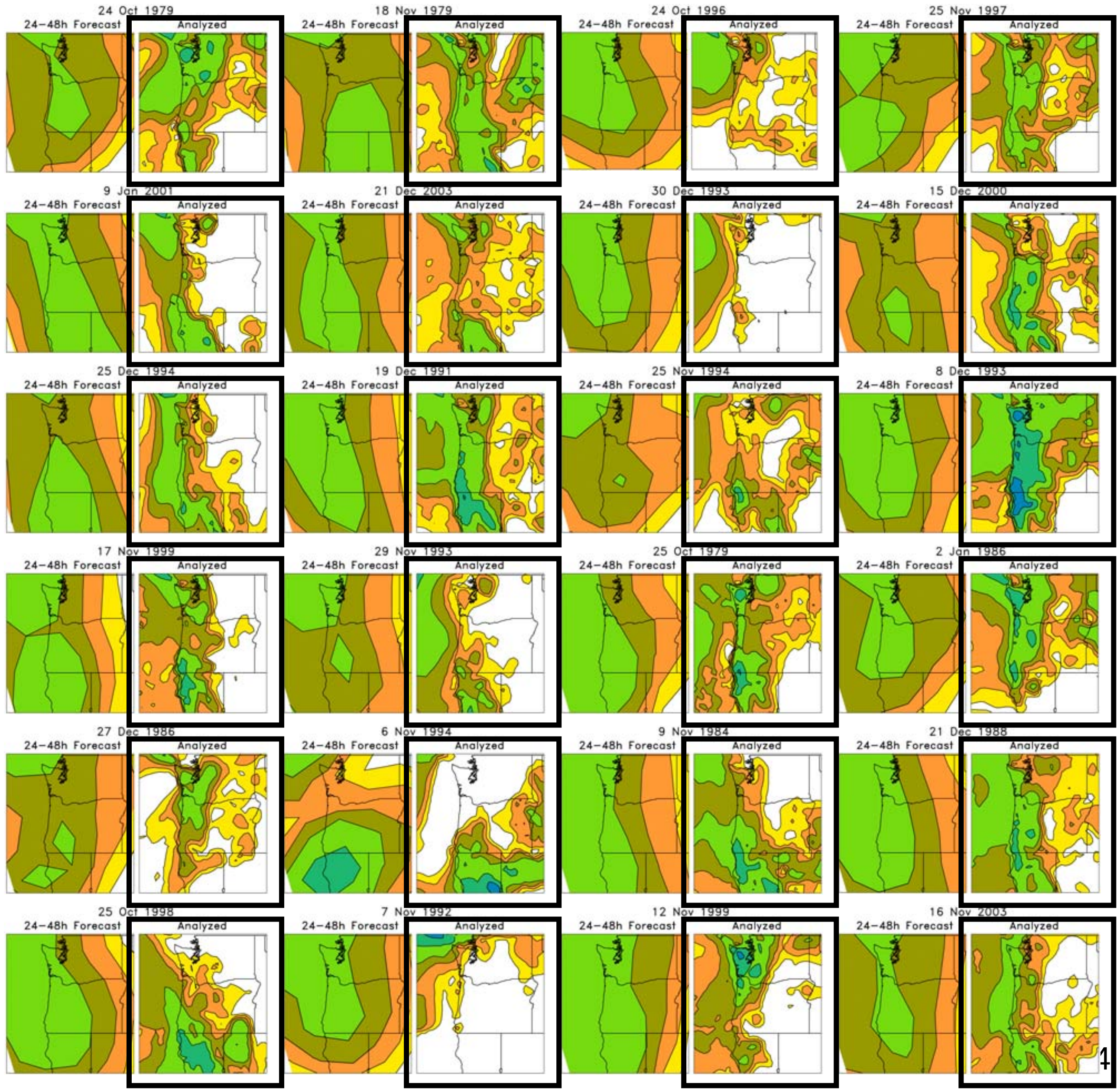
(1979-2004 data; scored using very stringent RPSS that ensures that skill not awarded due to variations in climatology)

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.
- **Experimental precipitation forecast products**: http://www.cdc.noaa.gov/reforecast/narr .

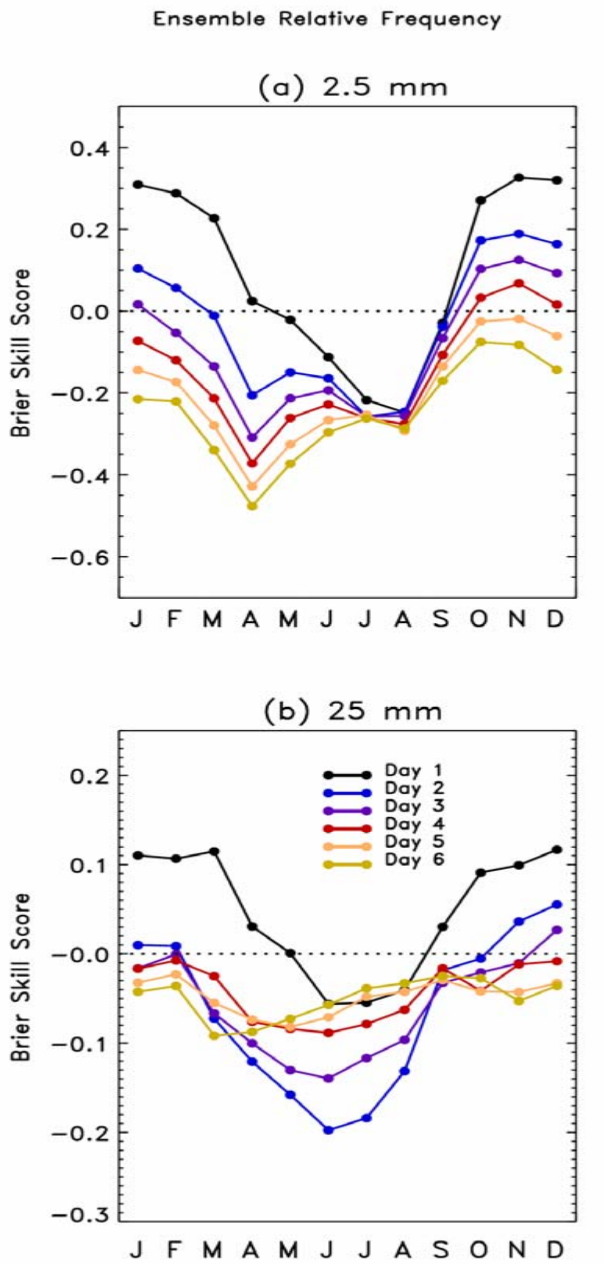
Reforecasts provide lots of old cases for diagnosing and correcting forecast errors.

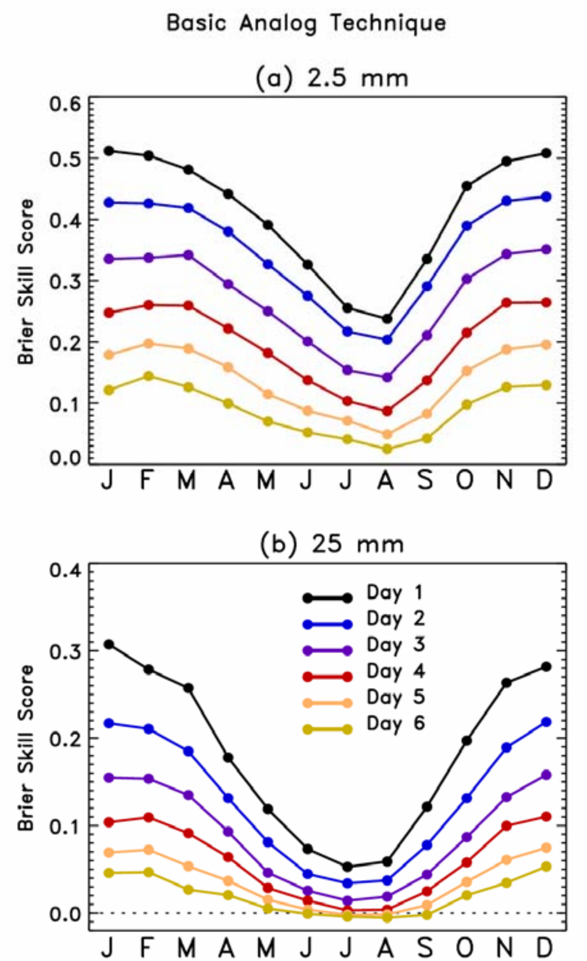




On the left are old forecasts similar to today's ensemble-mean forecast. The data on the right, the analyzed precipitation conditional upon the forecast, can be used to statistically adjust and downscale the forecast.

Before	After





Example of the benefit of reforecasts

Verified over 25 years of forecasts; skill scores use conventional method of calculation which may overestimate skill (Hamill and Juras 2006). Rest of talk uses more stringent method.

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 weeks total. So, 20y $\times$ 14w ensemble reforecasts = 280 samples.
- **Data** obtained by NOAA / ESRL : T_{2M} and precipitation ensemble over most of North America, excluding Alaska. Saved on 1-degree lat / lon grid. Forecasts to 10 days lead.

Questions

- Benefit of reforecast calibration from state-of-the art ECMWF model as much as with now outdated GFS model?
- How does the skill of probabilistic forecasts from the old GFS, with calibration, compare to the new ECMWF without?
- Are multi-decadal, every-day reforecasts really necessary? Given the computational expense, are much smaller training data sets adequate?

Outline

- A quick detour: examining why forecast skill metrics overestimate skill, and a proposed alternative.
- Calibrating temperature forecasts
- Calibrating precipitation forecasts
- Will reforecasting become operational at NWP centers worldwide?

Overestimating skill: a review of the Brier Skill Score

Brier Score: Mean-squared error of probabilistic forecasts.

$$\overline{BS}^{f} = \frac{1}{n}\sum_{k=1}^{n}\left(p_k^f - o_k\right)^2, \quad o_k = \begin{cases} 1.0 & \text{if } kth \text{ observation} \geq \text{threshold} \\ 0.0 & \text{if } kth \text{ observation} < \text{threshold} \end{cases}$$

Brier Skill Score: Skill relative to some reference, like climatology.
1.0 = perfect forecast, 0.0 = skill of reference.

$$BSS = \frac{\overline{BS}^{f} - \overline{BS}^{ref}}{\overline{BS}^{perfect} - \overline{BS}^{ref}} = \frac{\overline{BS}^{f} - \overline{BS}^{ref}}{0.0 - \overline{BS}^{ref}} = 1.0 - \frac{\overline{BS}^{f}}{\overline{BS}^{ref}}$$

Overestimating skill: example

5-mm threshold

Location A: $P^f = 0.05$, $P^{clim} = 0.05$, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.05-0)^2} = 0.0$$

Location B: $P^f = 0.05$, $P^{clim} = 0.25$, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.25-0)^2} = 0.96$$

Locations A and B:

why not 0.48?

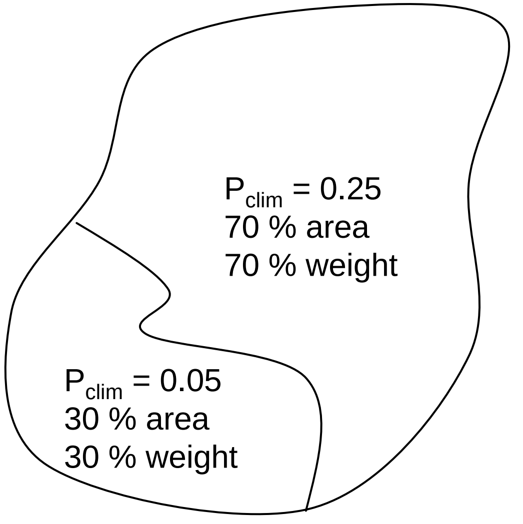
$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2 + (.05-0)^2}{(.25-0)^2 + (.05-0)^2} = 0.923$$

for more detail, see Hamill and Juras, QJRMS, Oct 2006 (c)

An alternative *BSS*

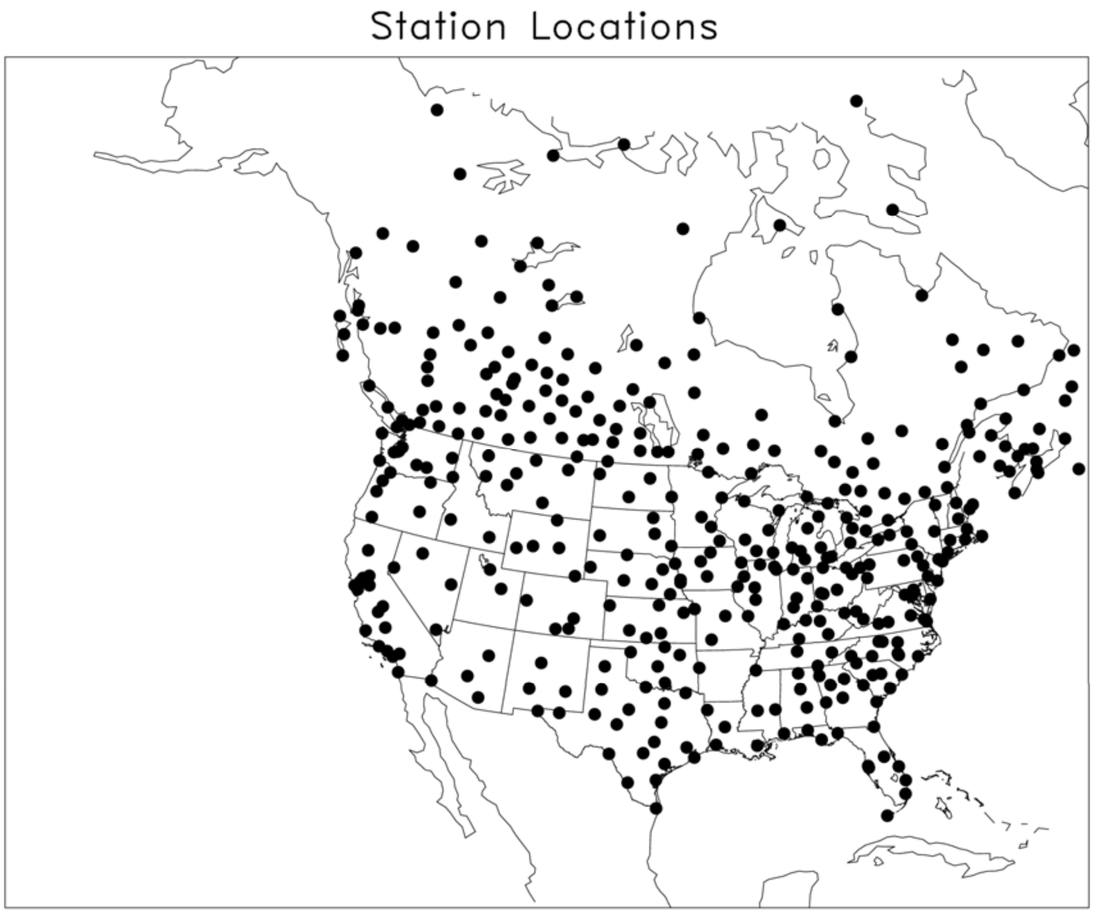
Say *m* overall samples, and *k* categories where climatological event probabilities are similar in this category. $n_s(k)$ samples assigned to this category. Then form BSS from weighted average of skills in the categories.

$$BSS = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \left(1 - \frac{\overline{BS}^f(k)}{\overline{BS}^{clim}(k)} \right)$$



(for more details on all of this, see Hamill and Juras, *QJRMS*, October C, 2006)

Observation locations for temperature calibration



Produce probabilistic forecasts at stations.

Use 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.

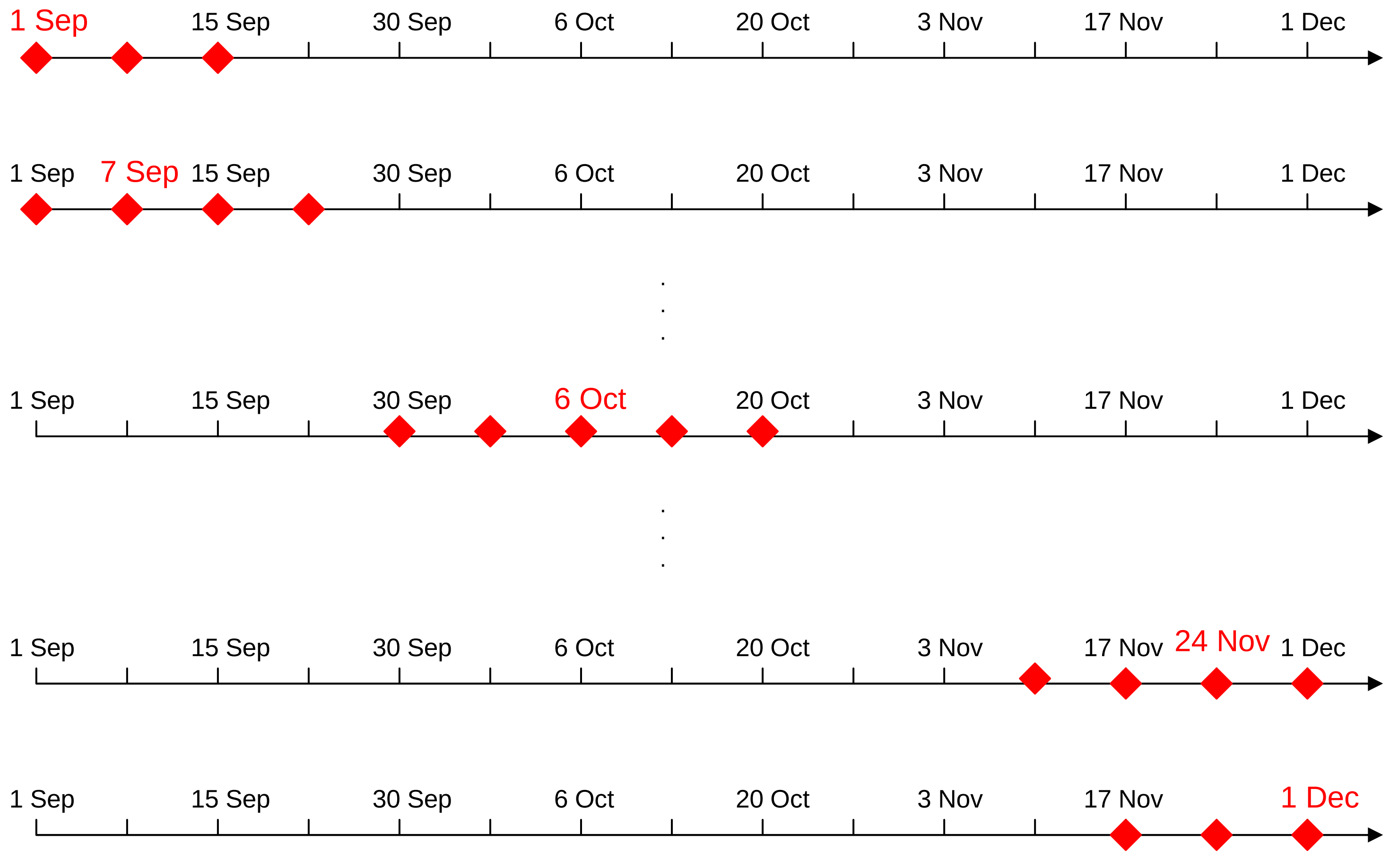
Calibration Procedure: “NGR”
“Non-homogeneous Gaussian Regression”

- **Input predictors**: ensemble mean and ensemble spread
- **Output**: mean, spread of calibrated normal distribution

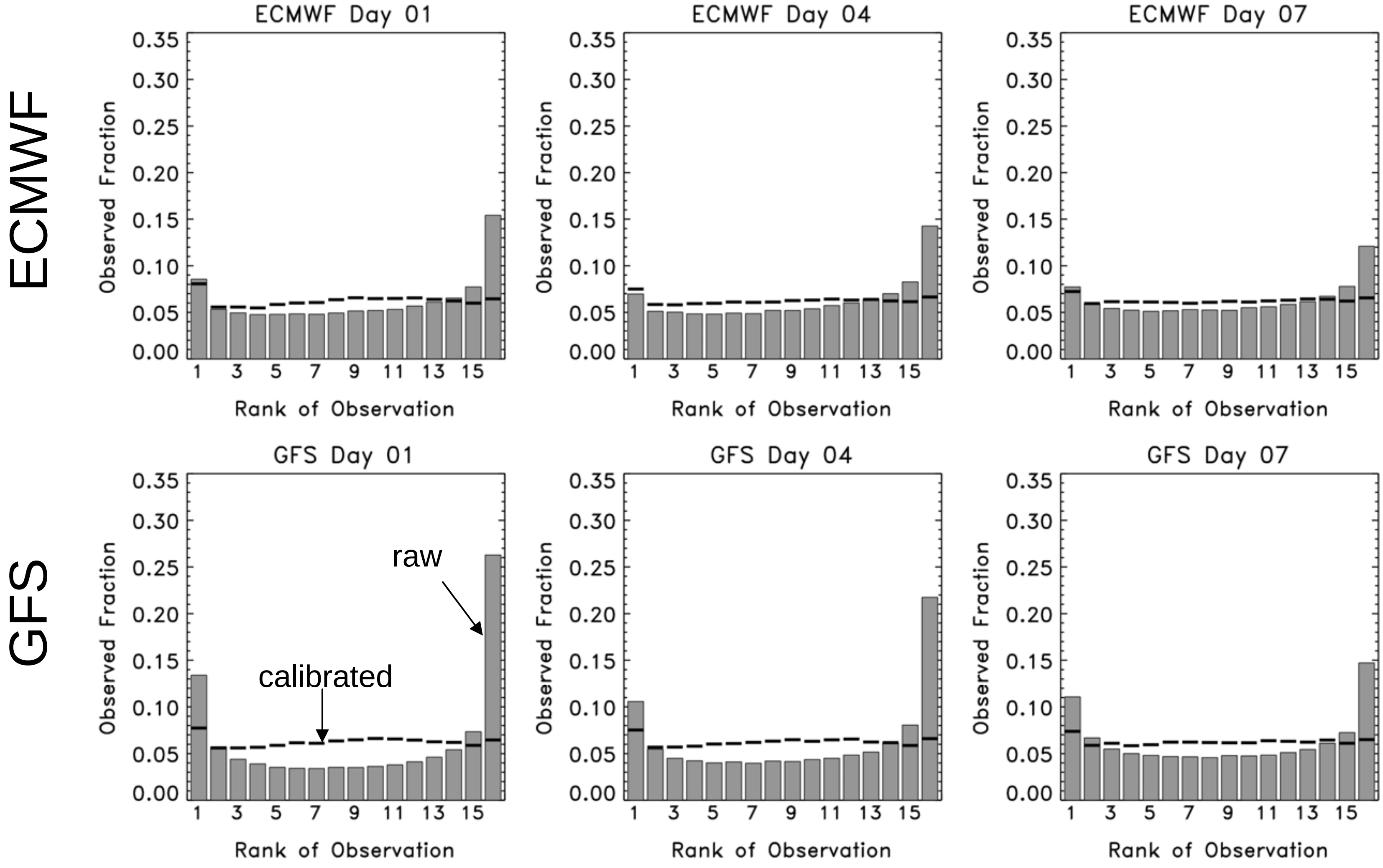
$$f^{CAL} \sim N\left(a + b\bar{\mathbf{x}}, c + d\boldsymbol{\sigma}\right)$$

- **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**: reforecasts +/- 2 weeks within date of interest.
- **Reference**: Gneiting et al., *MWR*, **133**, p. 1098. Shown in Wilks and Hamill (*MWR*, **135**, p. 2379) to be best of common calibration methods for surface temperature using reforecasts.

What training data to use, given inter-annual variability of forecast bias?

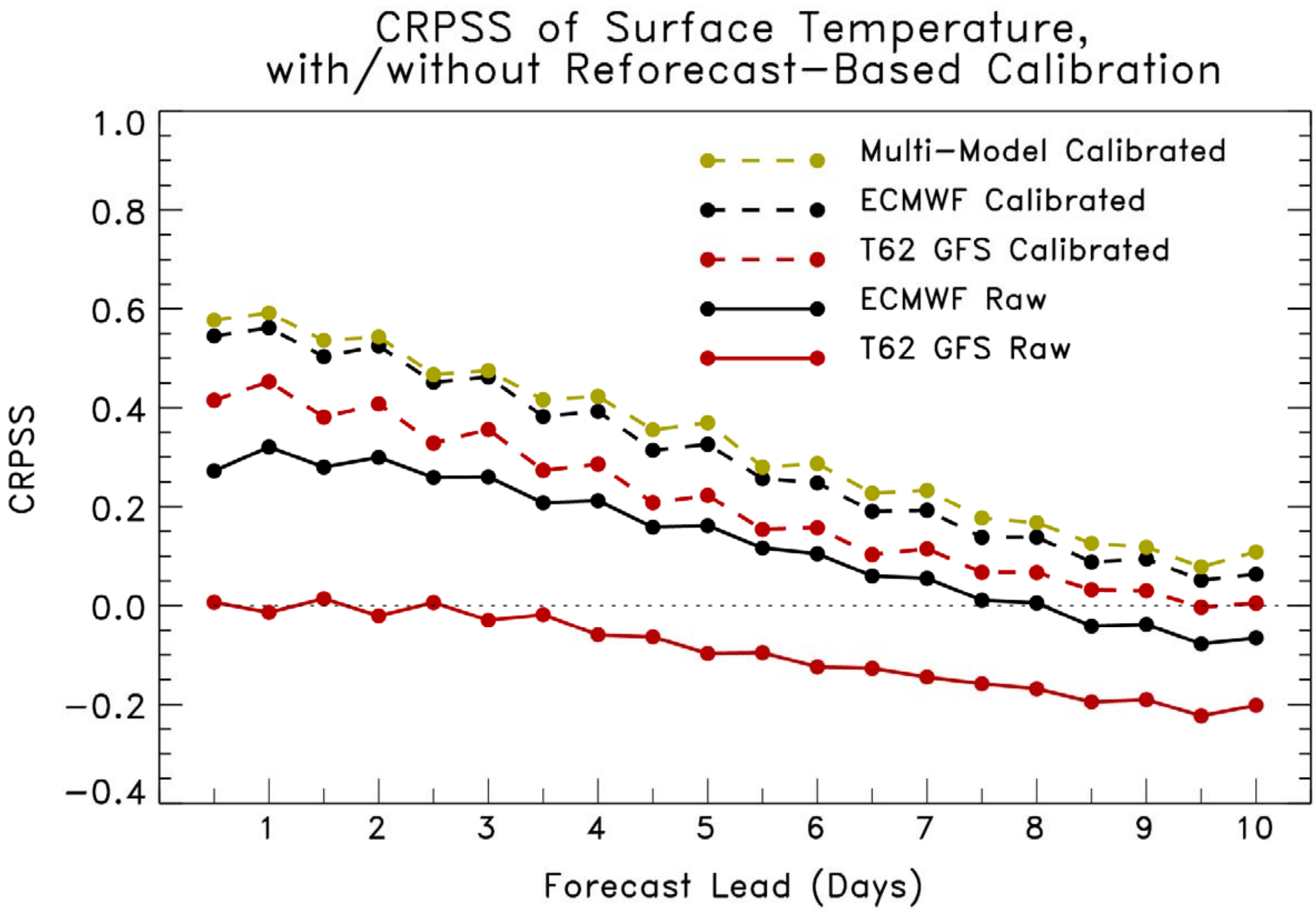


Rank histograms, before & after



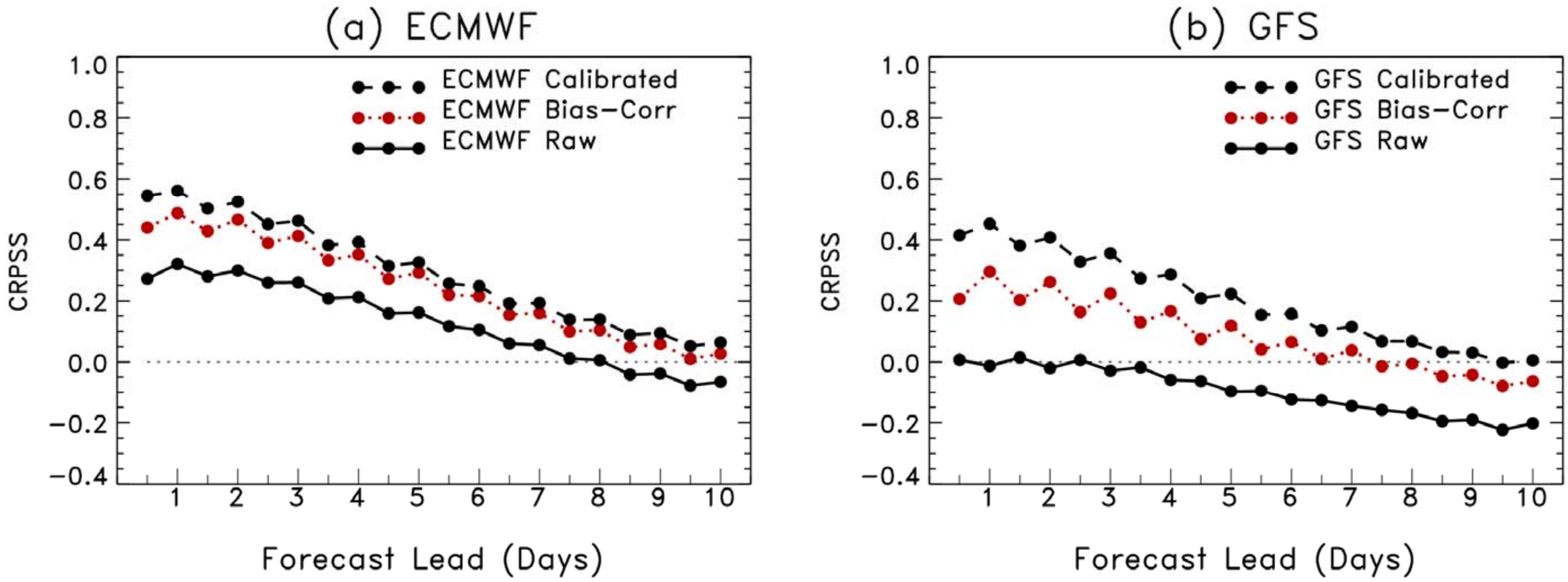
Members randomly perturbed by 1.5K to account for observation error; probably a bit small for GFS on its coarser 2.5º grid, which if perturbed by larger amount would make their histograms slightly more uniform. Ref: Hamill, *MWR*, **129**, p. 556**.**

ECMWF, raw and post-processed



Note: 5th and 95th percentile confidence intervals very small, 0.02 or less, so not plotted

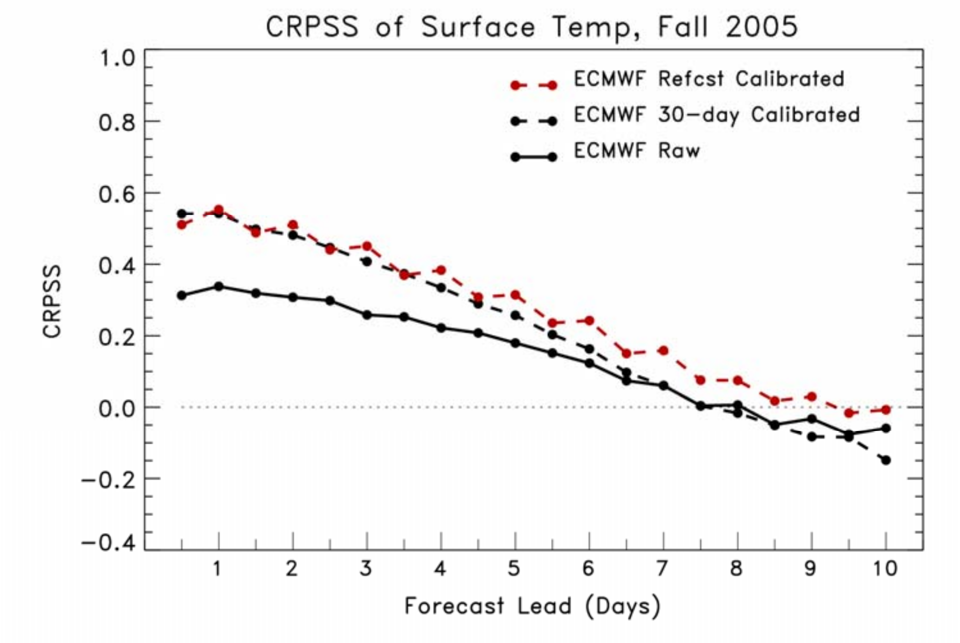
How much from simple bias correction?



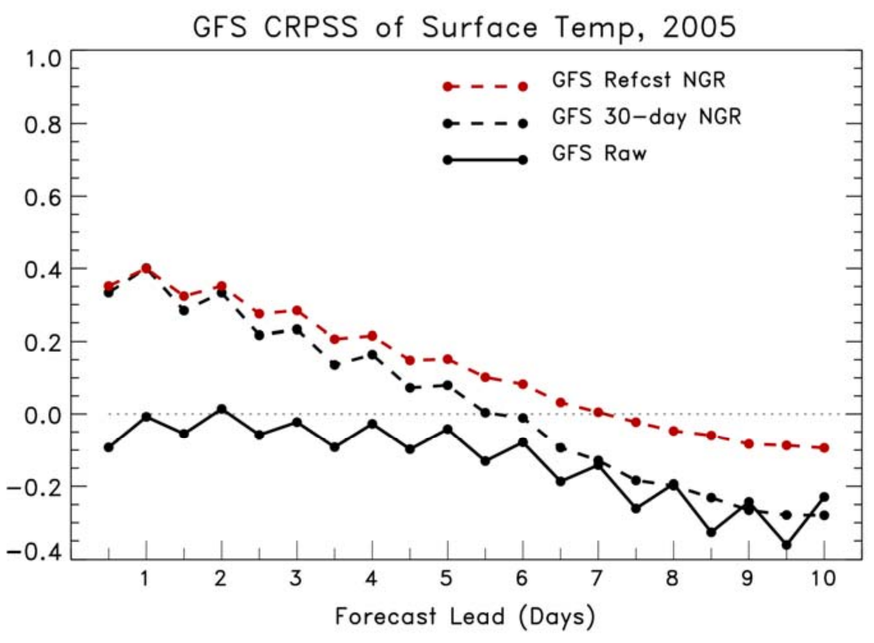
~ 60 percent of total improvement at short leads, 70 percent at longer leads.

How much from short training data sets?

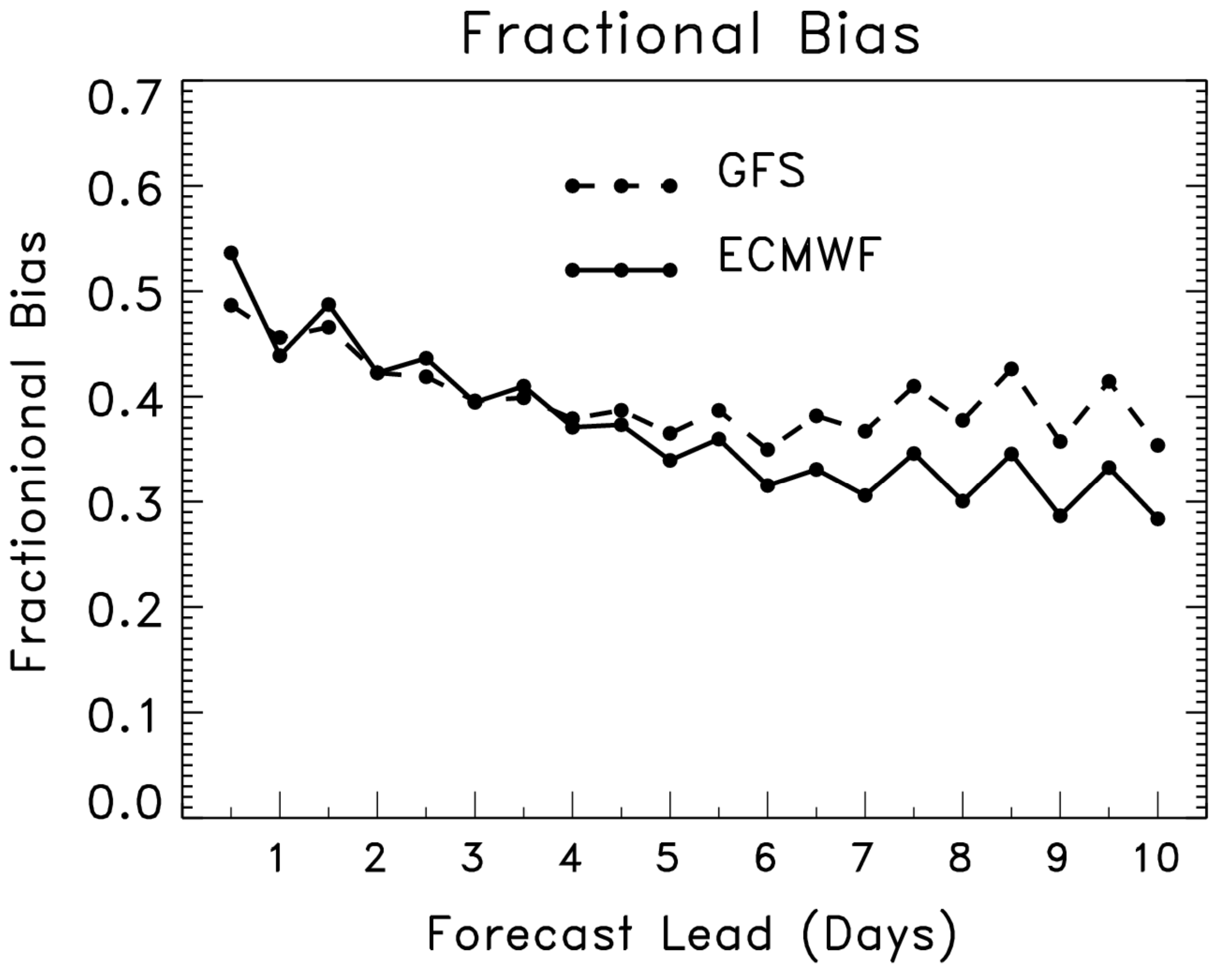
ECMWF



GFS



Note: (1) that ECMWF reforecasts use 3D-Var initial condition, 2005 real-time forecasts use 4D-Var. This difference may lower skill with reforecast training data set. (2) No other predictors besides forecast T_{2m}; perhaps with, say, soil moisture as additional predictor, reforecast calibration would improve relative to 30-day.



This measures the percentage of the forecast error that can be attributed to a long-term mean bias, as opposed to random errors due to chaos. Random errors are a larger percentage at long leads.

Precipitation calibration

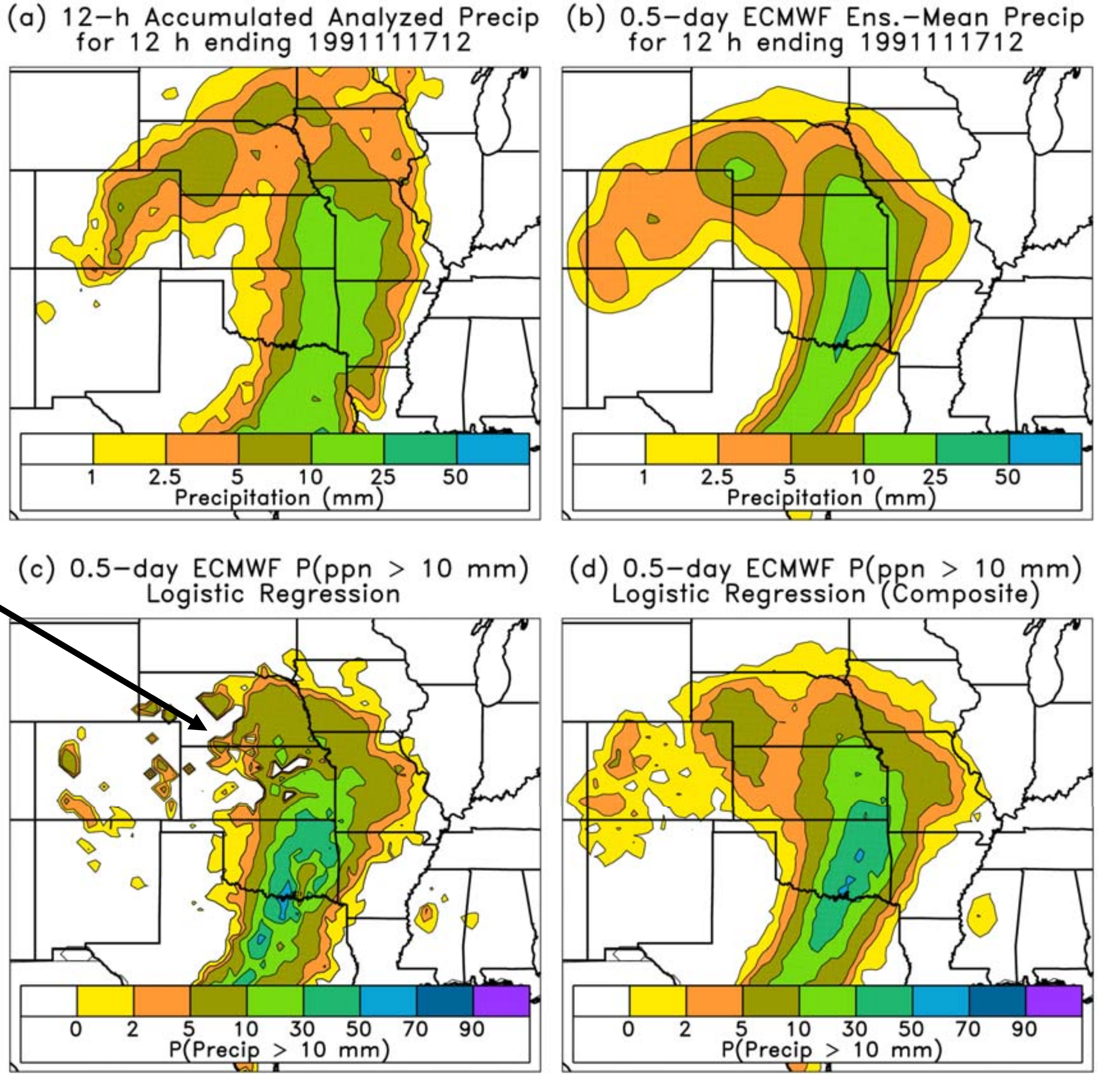
- North American Regional Reanalysis (NARR) CONUS 12-hourly data used for training, verification. ~32 km grid spacing.
- Logistic regression for calibration here

$$P(O > T) = 1.0 - \frac{1.0}{1.0 + \exp\left\{\beta_0 + \beta_1\left(\bar{x}^f\right)^{0.25} + \beta_2\left(\sigma^f\right)^{0.25}\right\}}$$

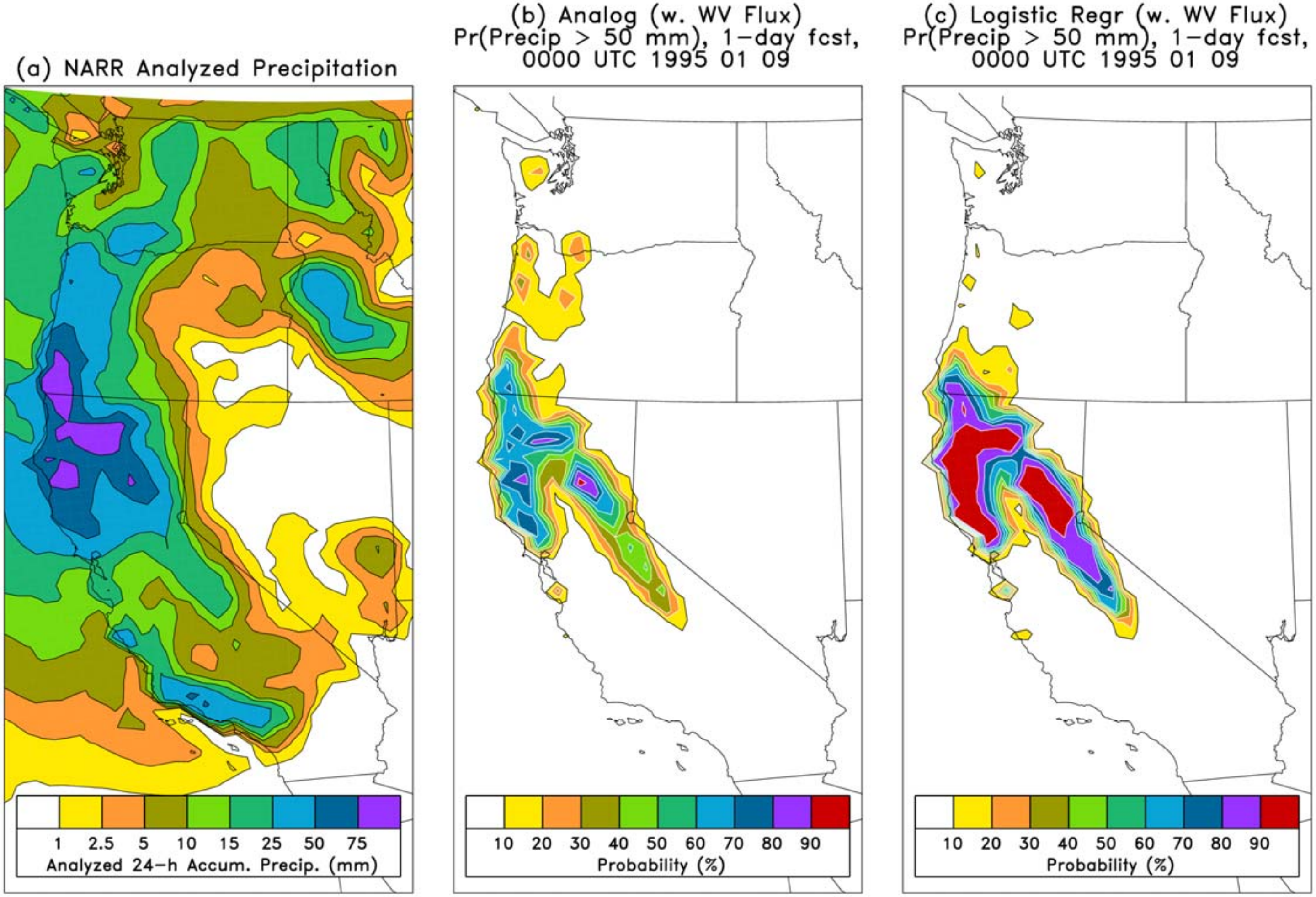
- More weight to samples with heavier forecast precipitation to improve calibration for heavy-rain events.
- Unlike temperature, throw Sep-Dec training data together.

Problem: patchy probabilities when grid point X trained with only grid point X's forecasts / obs

Even 20 years of weekly forecast data (260 samples after cross-validation) is not enough for stable regression coefficients, especially at higher precipitation thresholds.



Logistic regression similar to analog …

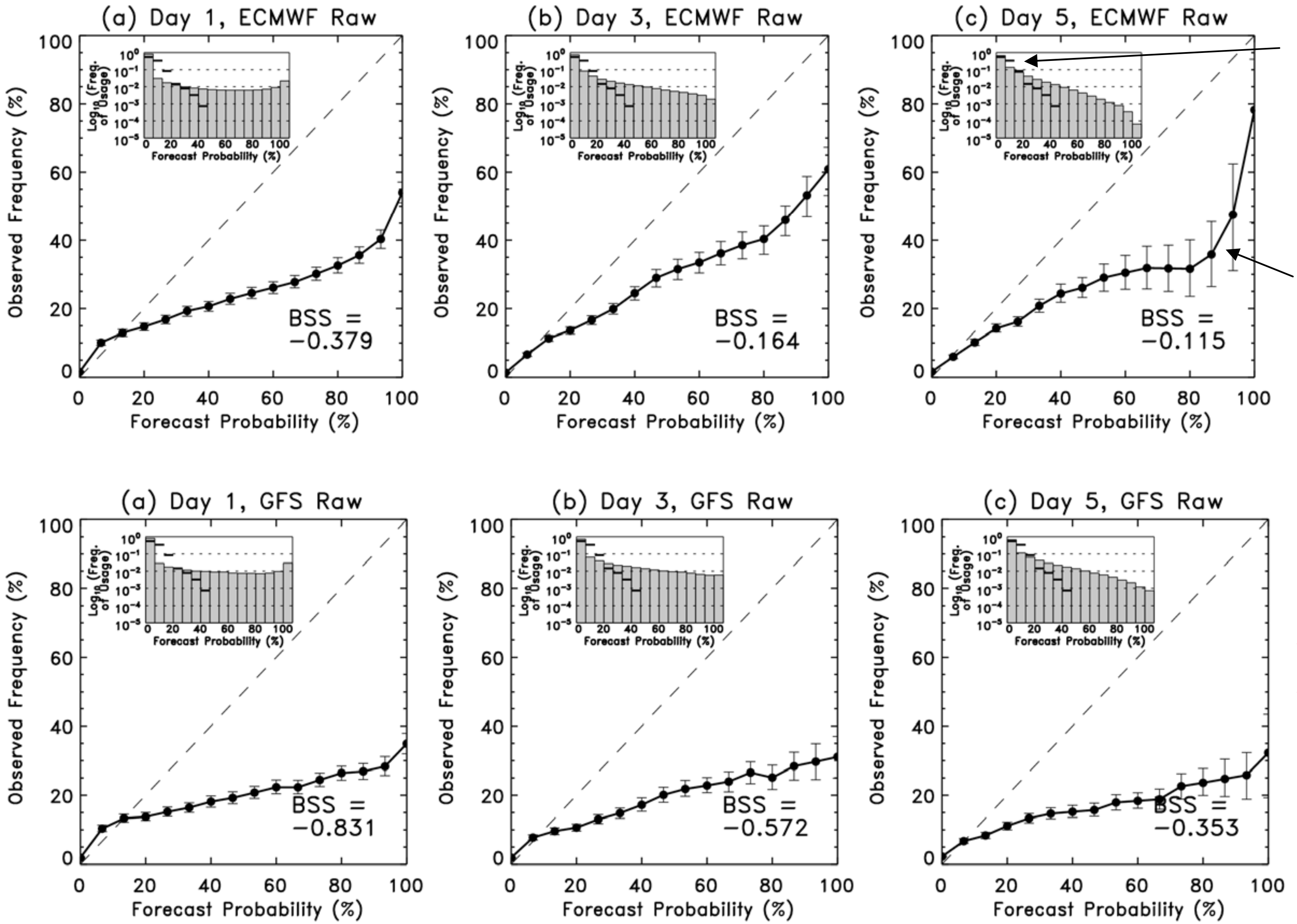


…though it tends to forecast higher probabilities

Training data sets tested

- “Weekly” - use 1x weekly, 20-year reforecasts for training data. Sep-Dec cases all thrown together. X-validated.
- “30-day” - for 2005 only, where forecasts available every day, train using the prior available 30 days.
- “Full” (GFS only) - use 25 years of daily reforecasts. X-validated.

5-mm reliability diagrams, raw ensembles



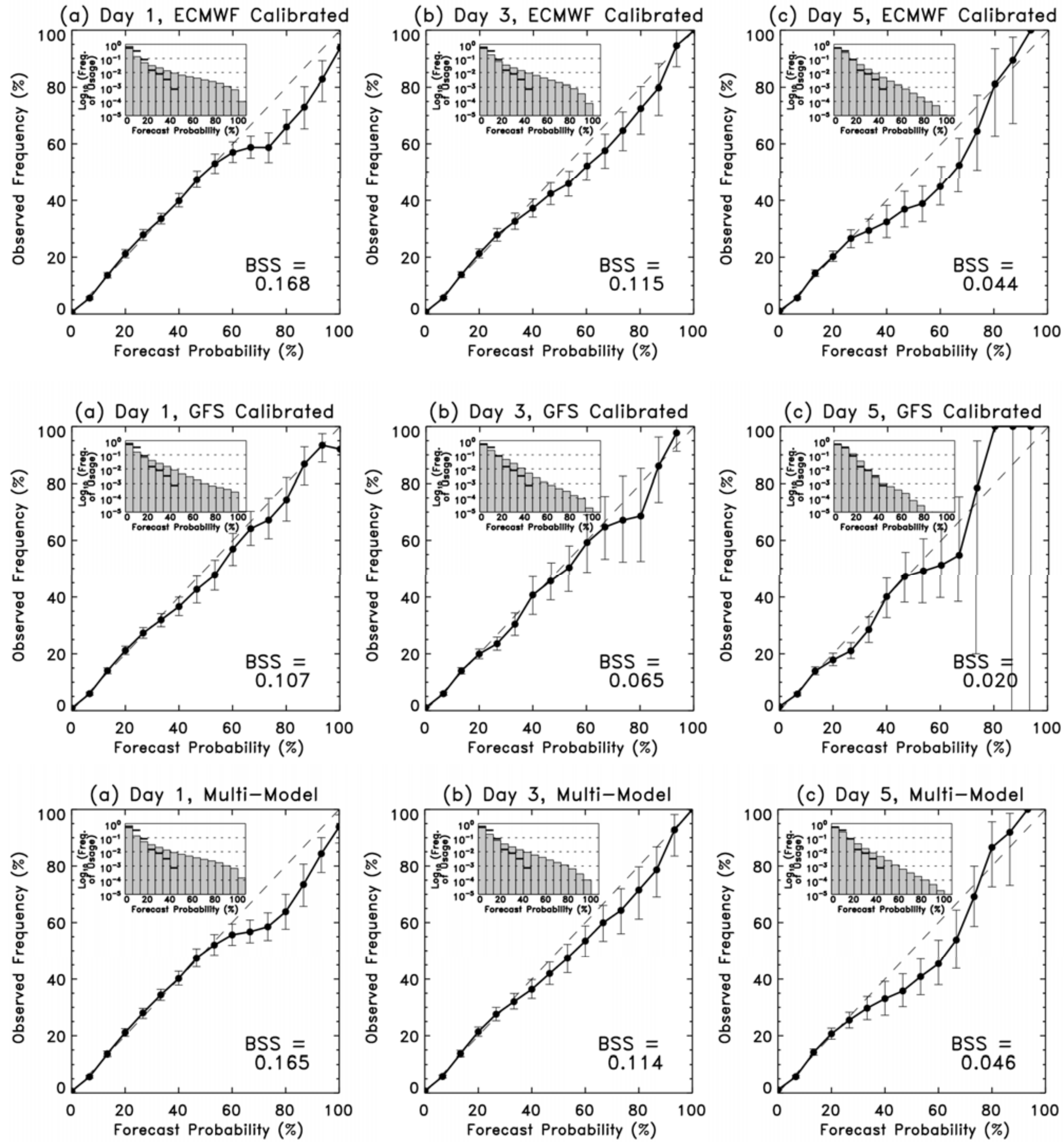
horizontal lines indicate distribution of climatology

error bars from block bootstrap

Raw forecasts have poor skill in this strict BSS

5-mm reliability diagrams, calibrated

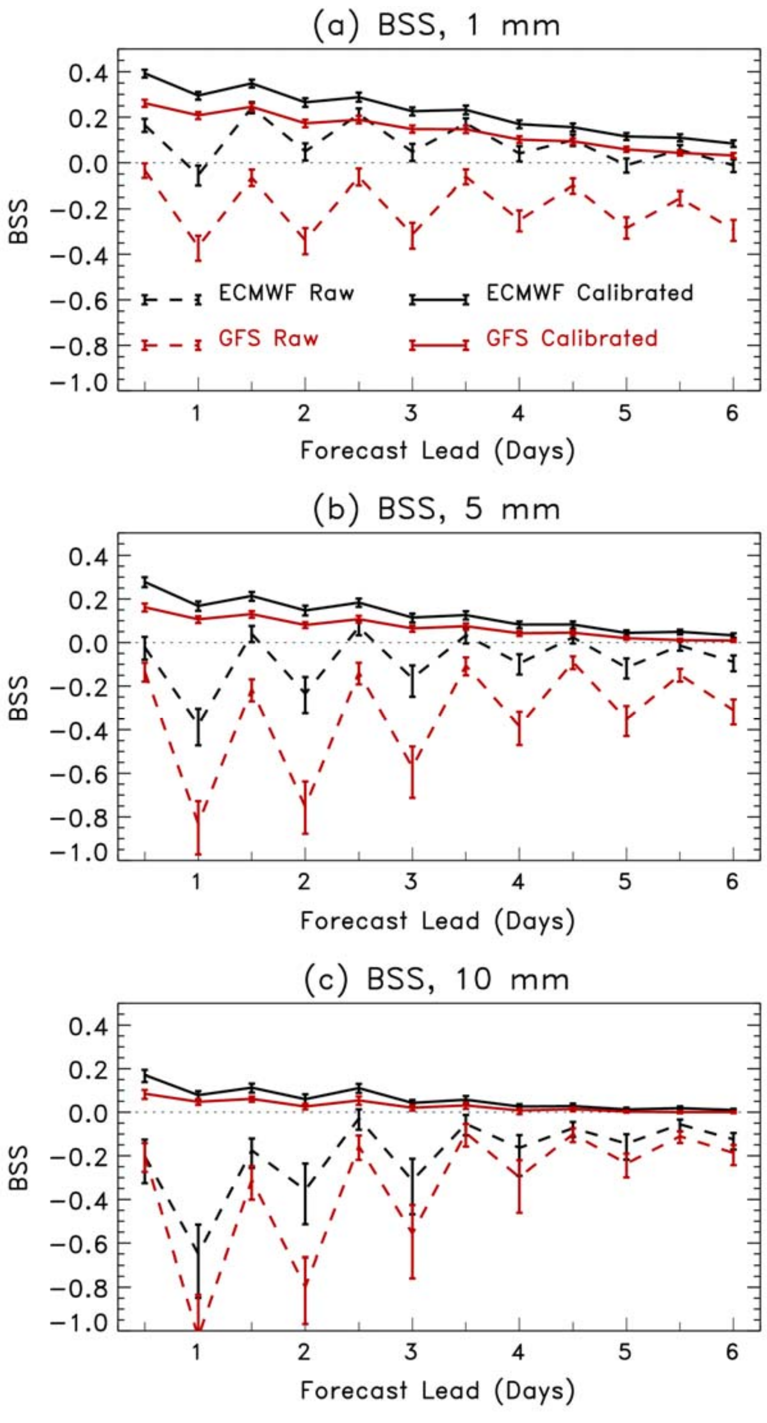
In some respects GFS forecasts look more calibrated but the frequency of usage histograms show ECMWF sharper and thus more skillful.



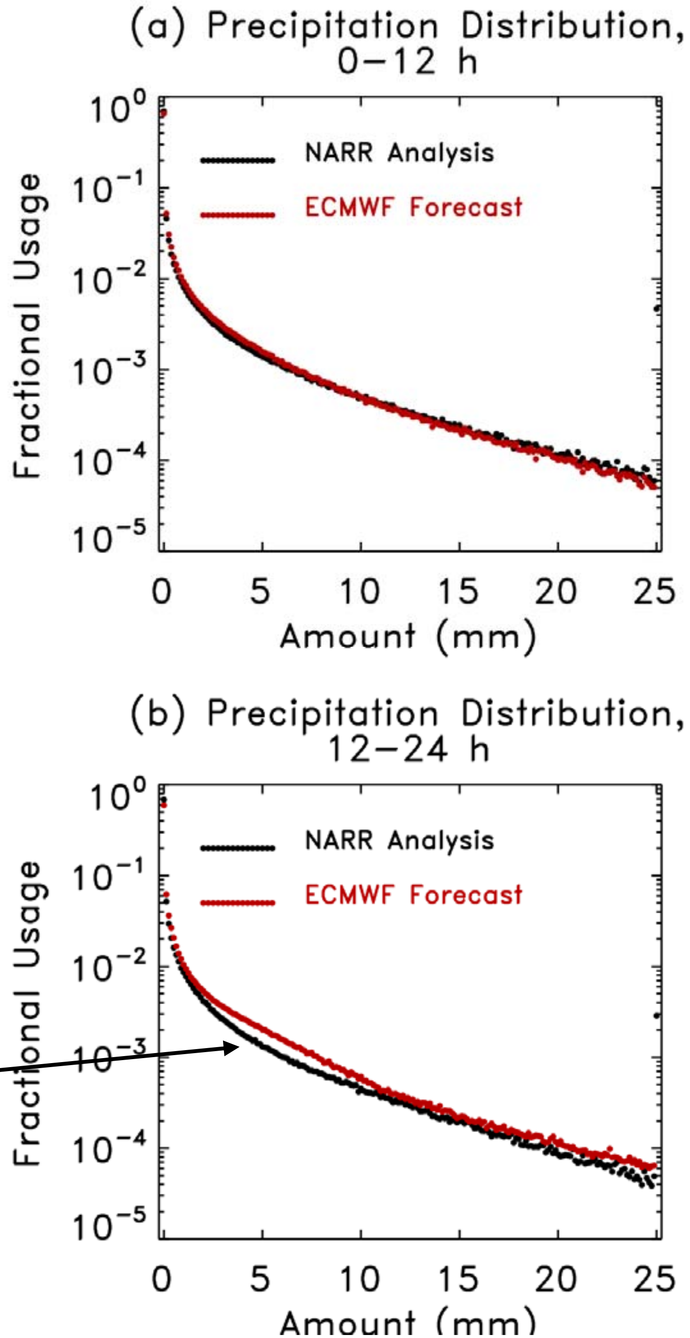
Brier Skill Scores

Notes:

(1) Diurnal oscillation in raw forecast skill
(2) Raw forecast skill poor, especially at higher thresholds
(3) Calibration has substantial positive impact.
(4) ECMWF > GFS skill.
(5) Multimodel not plotted, ~ same as ECMWF calibrated

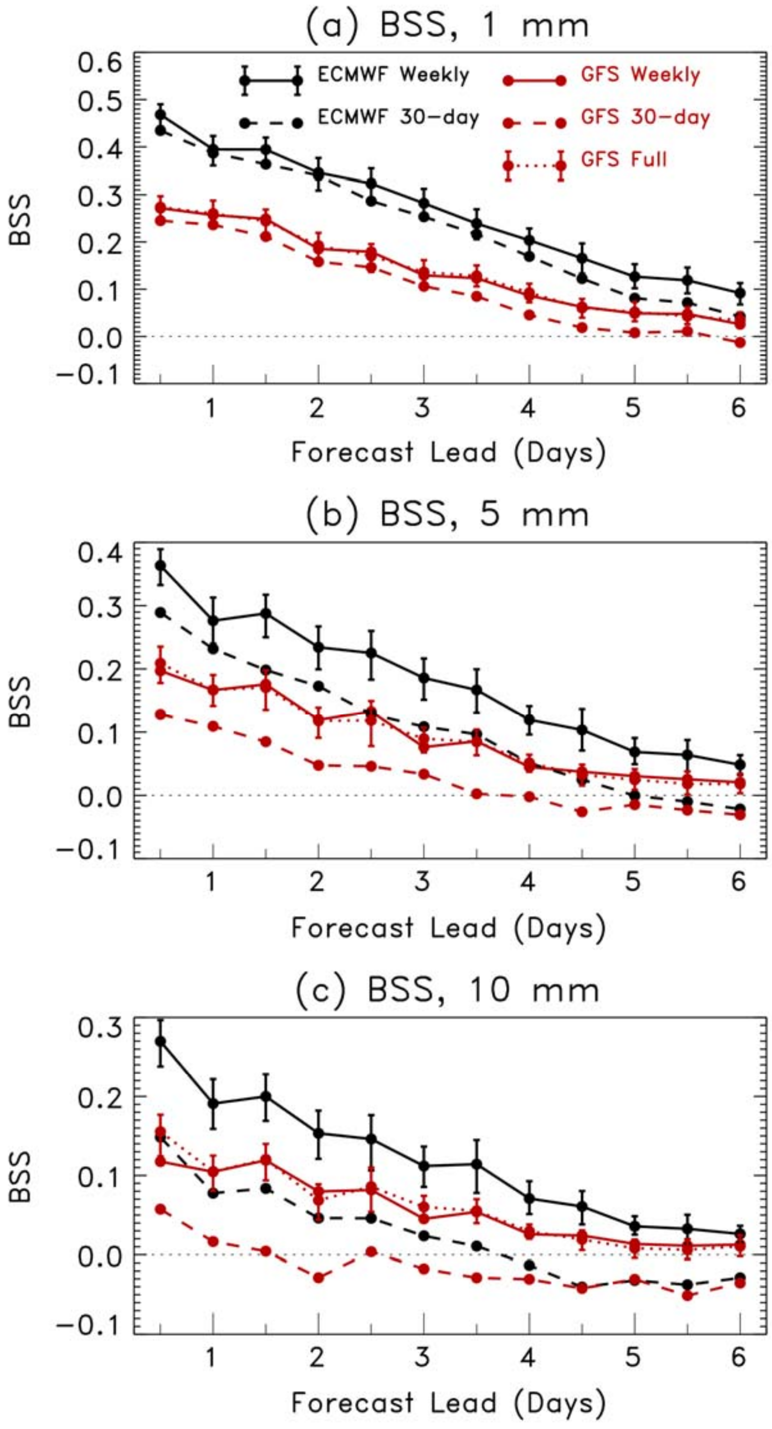


Why are 12Z - 00Z forecasts less skillful?



Over-forecast bias in models during daytime relative to NARR

Precipitation skill with weekly, 30-day, and full training data sets



Notes:

(1) Substantial benefit of weekly relative to 30-day training data sets, especially at high thresholds.

(2) Not much benefit from full relative to weekly reforecasts.

Conclusions

- Still a large benefit from forecast calibration, even with state-of-the-art ECMWF forecast model.
- Temperature calibration:
  - Short leads: a few previous forecasts adequate for calibration
  - Long leads: better skill with long reforecast training data set.
- Precipitation calibration
  - Low thresholds: a few previous forecasts somewhat ok for calibration
  - Larger thresholds: large benefit from large training data set.

Other research issues

- Optimal reforecast ensemble size?
  – Other results suggest ~ 5 members
- Optimal frequency, length of reforecasts data sets?
  – Multi-decadal, but every day may not be necessary
- End-to-end linkages into hydrologic prediction systems.
- New applications (fire weather, severe storms, wind forecasting).

Are operational centers heading toward reforecasting?

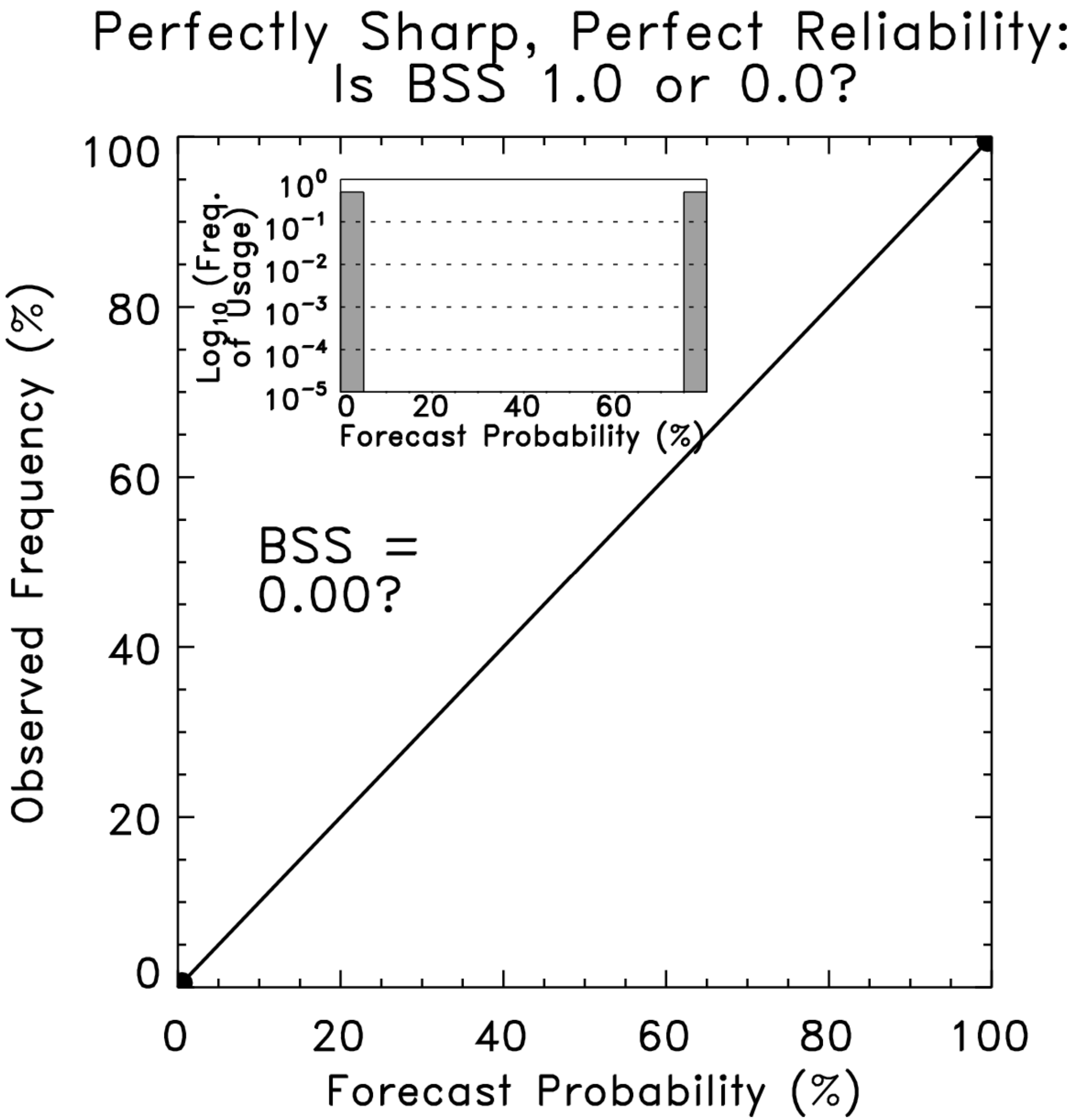
- **NCEP**: tentative plans for 1-member real-time reforecast.
- **ECMWF**: once-weekly, real-time 5-member reforecasts starting ~ early 2008.
- **RPN Canada**: possible ~5-year reforecast data set, delayed by budget and staffing issues.
- **NOAA-ESRL**: seeking computer resources for next-generation reforecast

References

- Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part I: surface temperature. *Mon. Wea. Rev*., submitted. Available at http://tinyurl.com/3axuac

- Hamill, T. M., J. S. Whitaker, and R. Hagedorn, 2007: Probabilistic forecast calibration using ECMWF and GFS ensemble forecasts. Part II: precipitation. *Mon. Wea. Rev*., submitted. Available at http://tinyurl.com/38jgkv

- (and references therein)



This is normally considered the reliability diagram of a perfect forecast. But suppose half the samples are from a location where the forecast probability is always zero, and the other half from a location where the forecast probability is always 1.0. Then even if the forecast is correct in both locations, it's never better than climatology… so skill should = 0.0 !

A thought experiment: two islands

Each island's forecast is an ensemble formed from a random draw from its climatology, ~ N($\pm\ \alpha$,1)

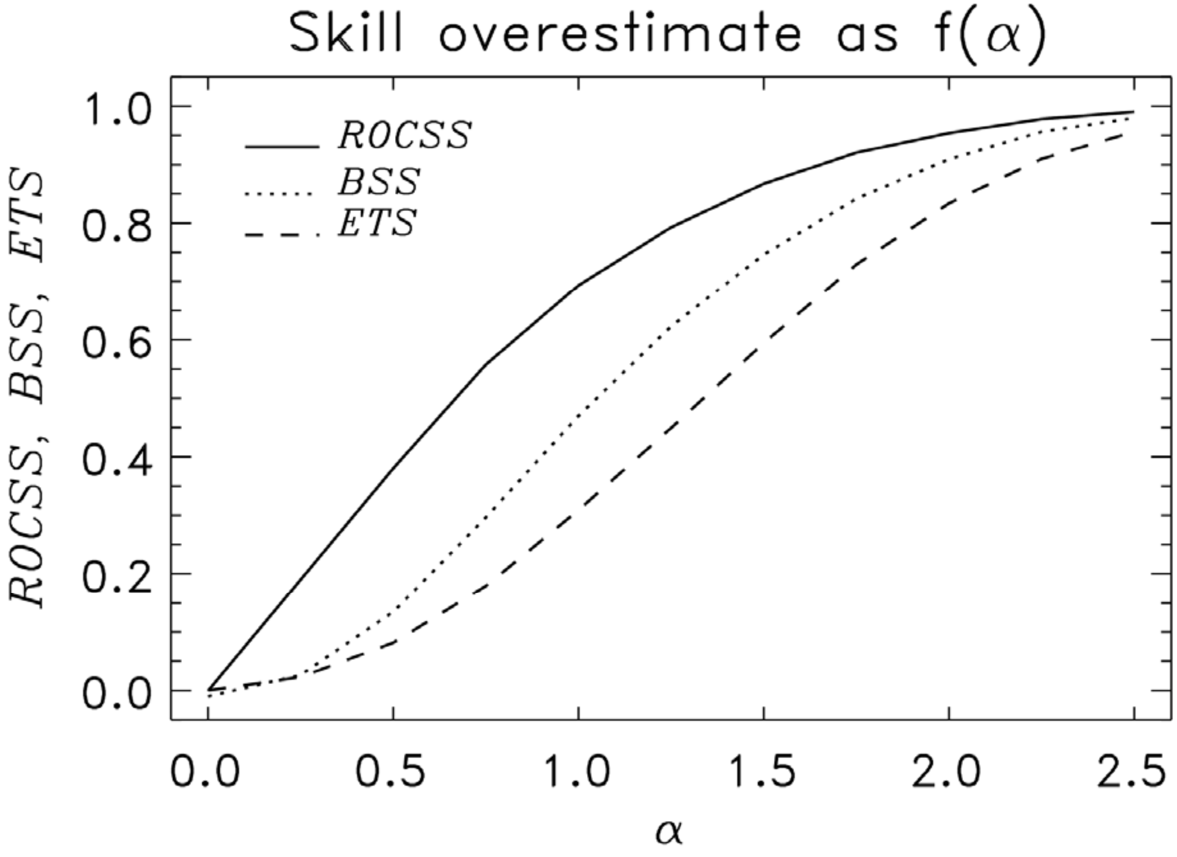
Island 2: ~N($-\alpha$,1)

$\longleftarrow$ As α increases… $\longrightarrow$

Island 1: ~N(α,1)

Expect no skill relative to climatology for the event P(Obs) > 0.0 for common meteorological verification methods like Brier Skill Score, Equitable Threat Score, ROC skill score.

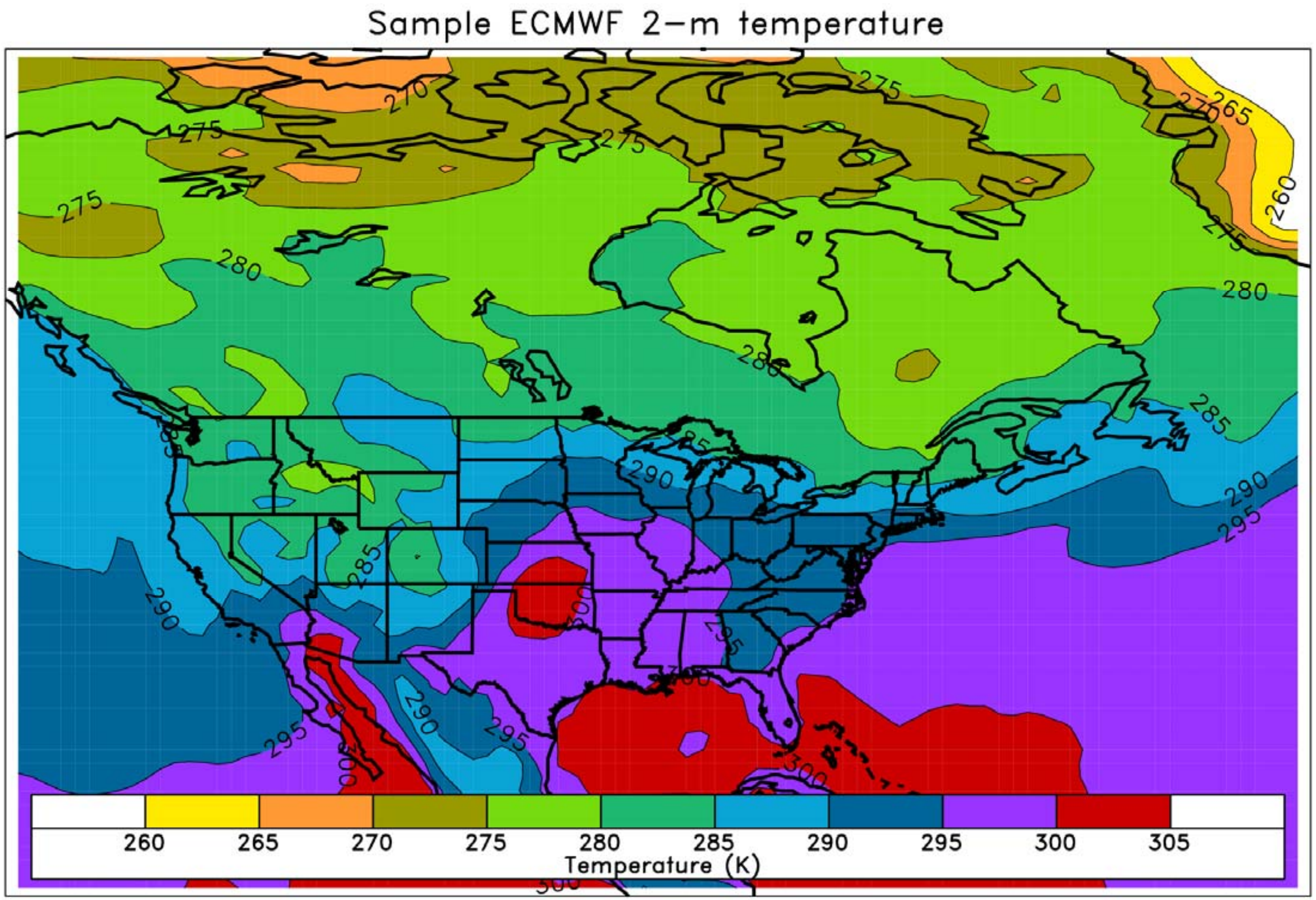
Skill with conventional methods of calculation



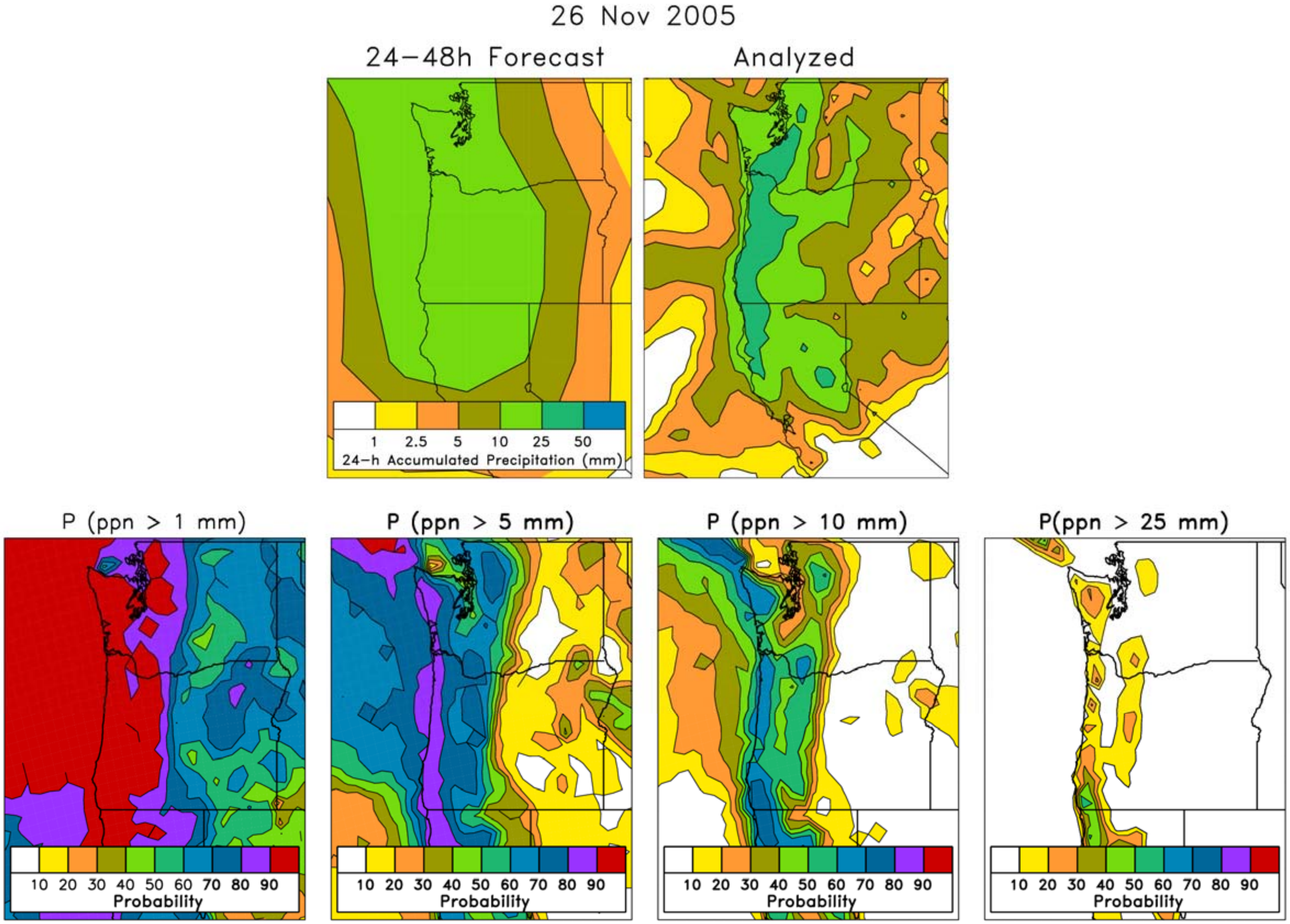
Reference climatology implicitly becomes

N(+α,1) + N(–α,1)   not N(+α,1) OR N(–α,1)

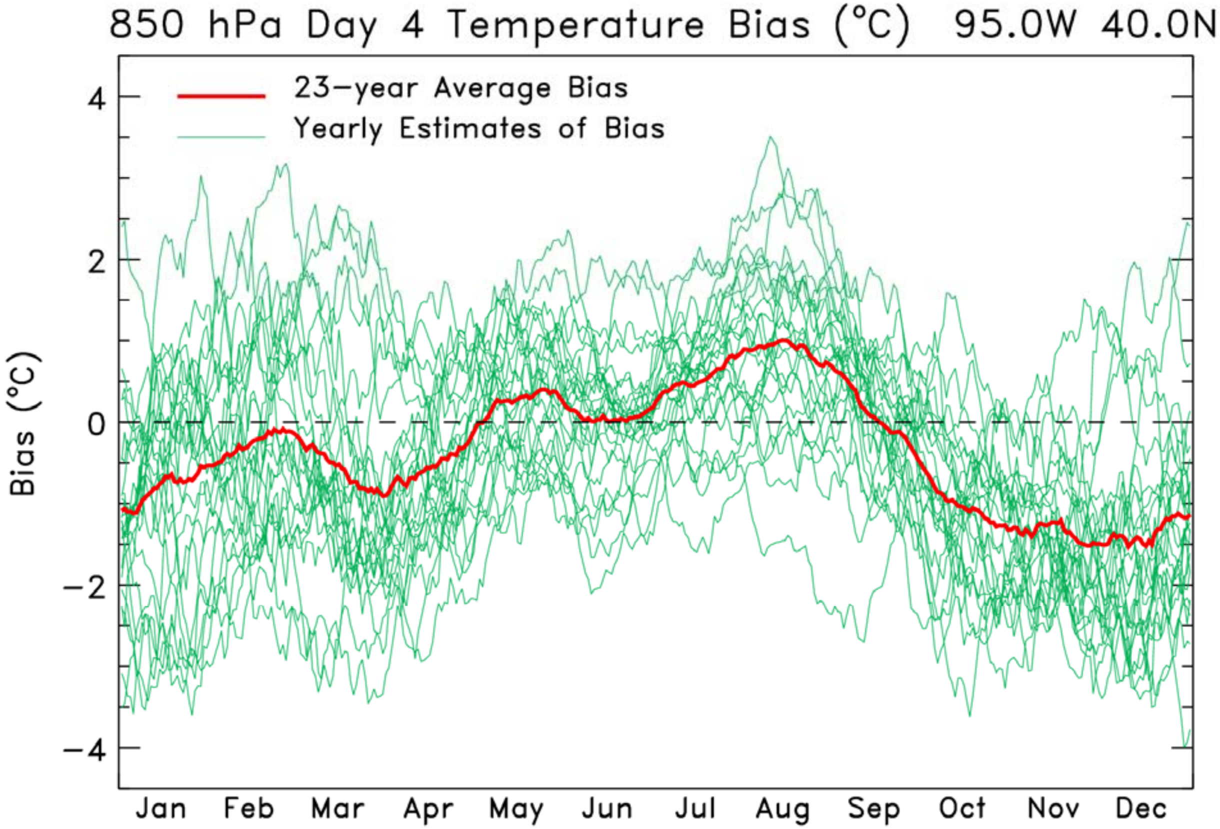
ECMWF domain sent to us for reforecast tests



Downscaled analog probability forecasts



Inter-annual variability of forecast bias



Red curve shows bias averaged over 23 years of data (bias = mean F-O in running 61-day window)

Green curves show 23 individual yearly running-mean bias estimates

Note large inter-annual variability of bias.

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

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

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

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

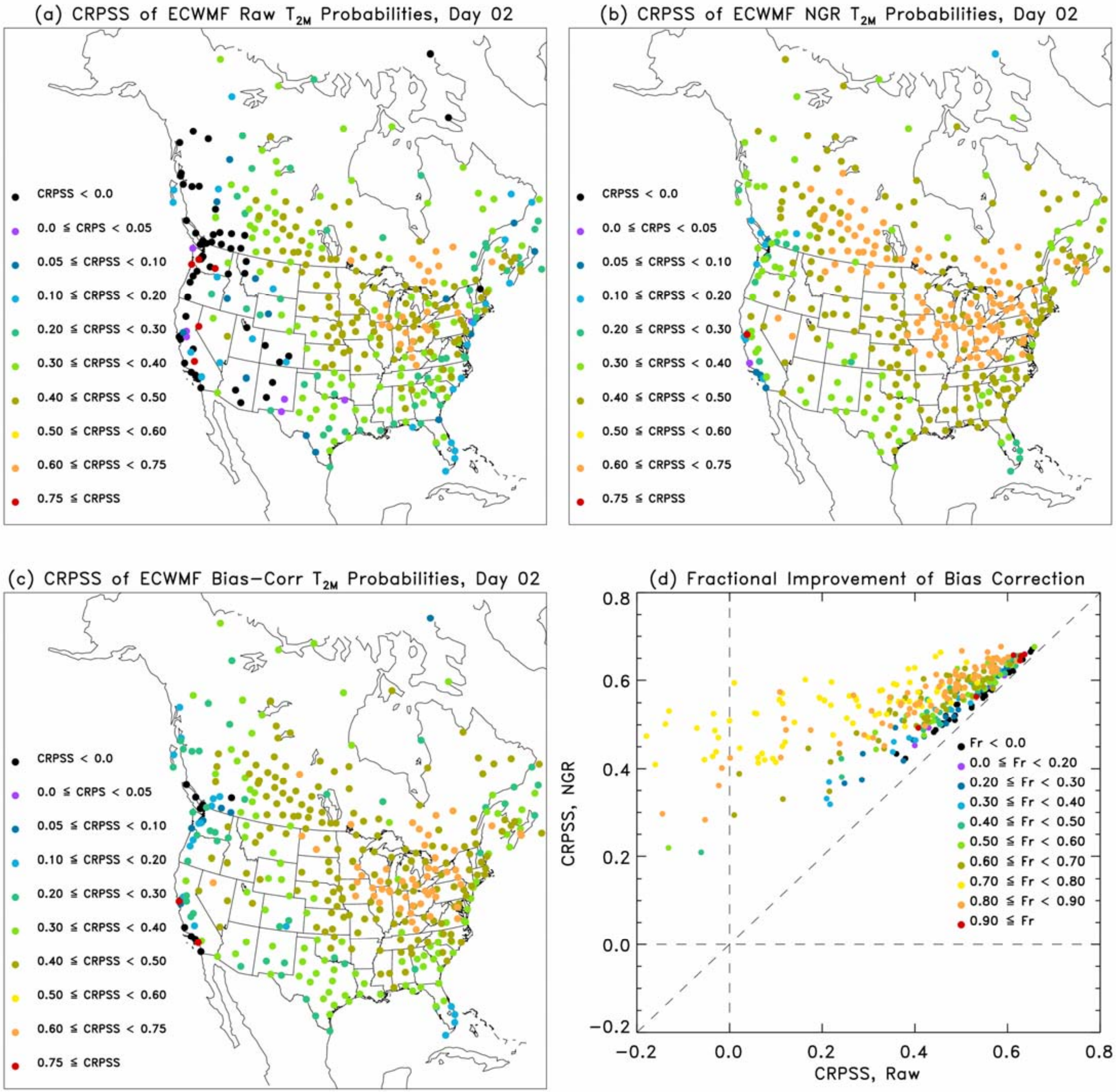
$$k = 1, \ldots, \# \, station \; locations$$

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

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

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

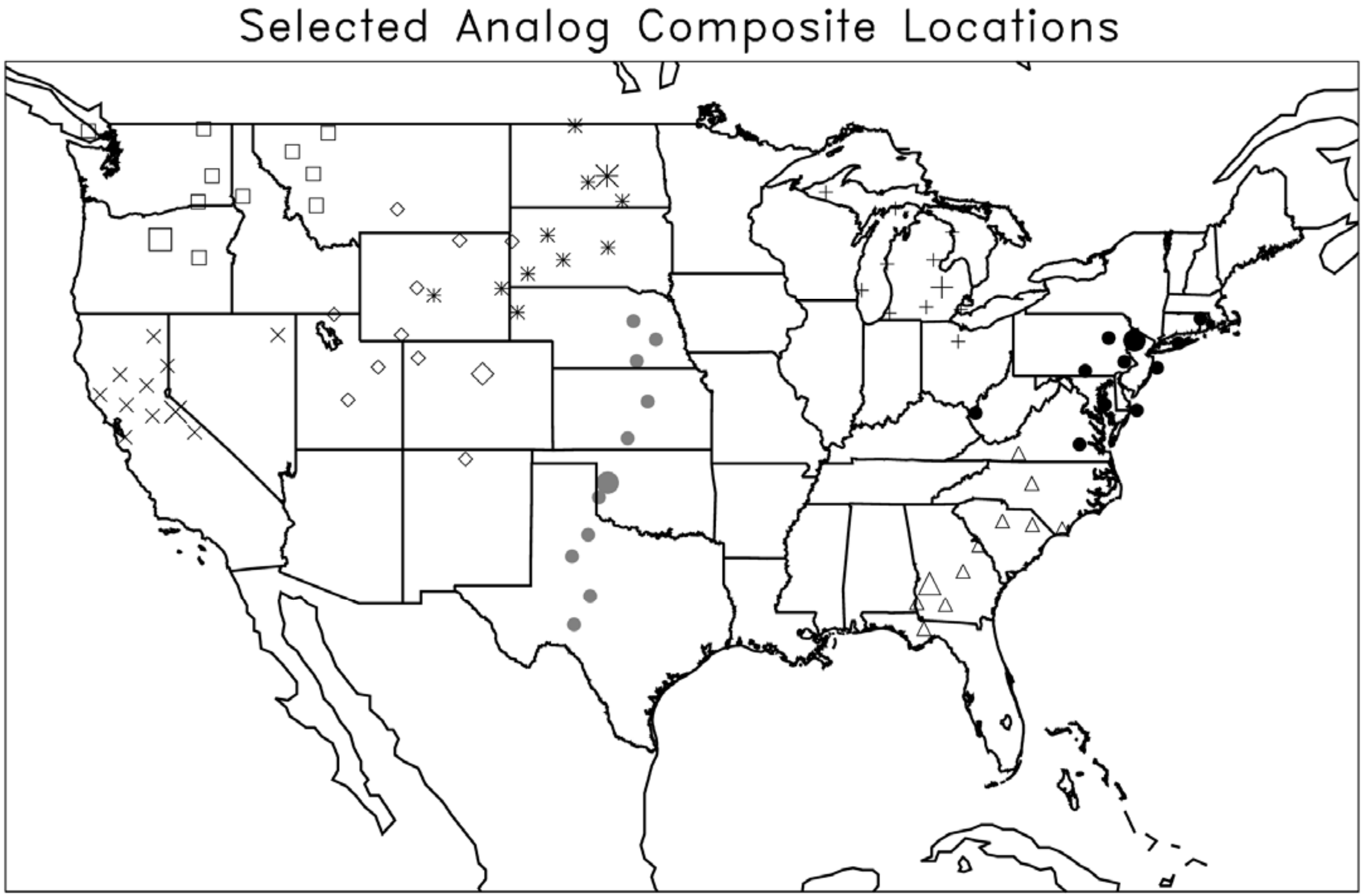
Will use a modified version where we calculate CRPSS separately for 8 different categories of climatological spread and then average them. See Hamill and Juras, January 2007, *QJRMS,* and Hamill and Whitaker Sep. 2007 *MWR.*



ECMWF’s geographical distribution of skill, before and after calibration.

The tide of calibration raises all boats, the sunken ones the most.

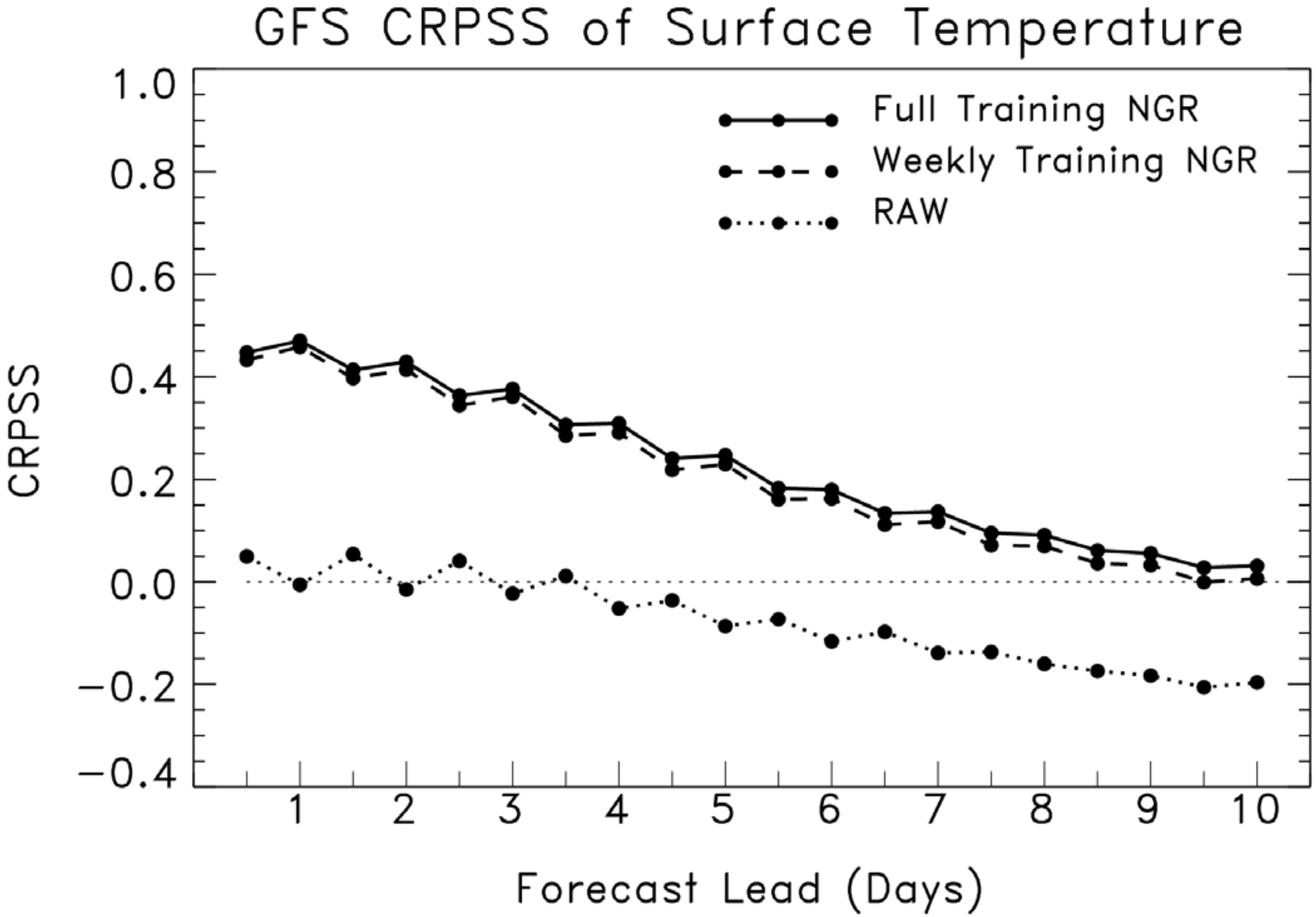
Tested method: add in training data at other grid points that have similar analyzed climatologies



Big symbol:
grid point
where we
do regression

Small symbols:
analog locations
with similar
climatologies

How much from long GFS training data set?



Here GFS reforecasts sampled once per week are compared to those sampled once per day ("full").