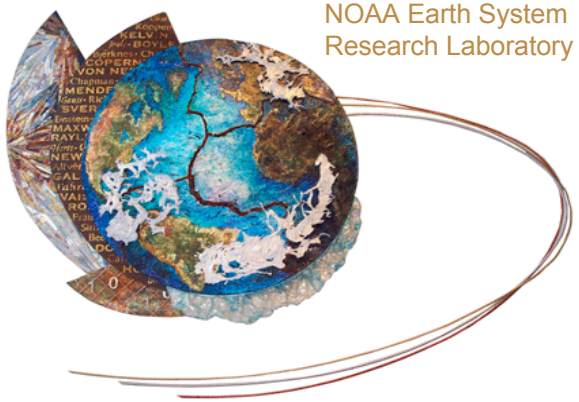
prepared for 2006 NCEP Predictability Meeting, 14 Nov 2006



NOAA Earth System
Research Laboratory

A discussion of the use of reforecasts

Tom Hamill and Jeff Whitaker

NOAA/ESRL, Physical Sciences Division

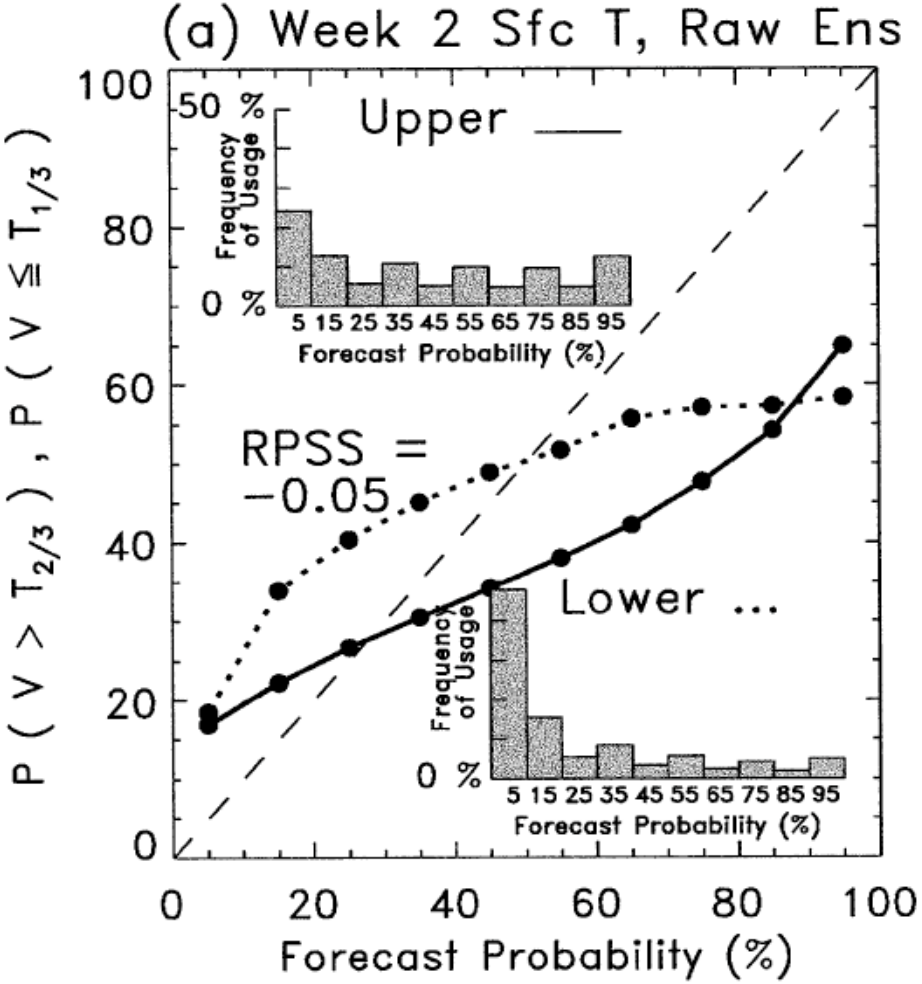
tom.hamill@noaa.gov, (303) 497-3060

Motivation

"*Hydrometeorological services in the United States are an Enterprise effort. Therefore, effective incorporation of uncertainty information will require a fundamental and coordinated shift by all sectors of the Enterprise. Furthermore, it will take time and perseverance to successfully make this shift. As the Nation's public weather service, NWS has the responsibility to take a leading role in the transition to widespread, effective incorporation of uncertainty information into hydrometeorological prediction*."

– From finding 1 of NRC report "Completing the Forecast"

The problem with using raw ensemble forecasts



Probabilistic forecasts from raw ensembles are not very reliable, due to deficiencies in forecast model, ensemble methods.

Users want “sharp” and “reliable” forecasts. Statistical adjustment necessary.

References for reforecasting

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

Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, in press. http://www.cdc.noaa.gov/people/tom.hamill/skill_overforecast_QJ_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, 2006: Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using reforecasts. Mon. Wea. Rev.,submitted. Available at
http://www.cdc.noaa.gov/people/tom.hamill/Calibration_z500t850t2m.pdf

Outline

- Addressing some of the concerns NCEP has about reforecast calibration efficacy.
- Configuration of real-time reforecasts: the compromises between ideal and practical.
- Moving toward operational implementation.
  - Archiving and accessing reforecasts
  - What can we do to hit the ground running?
  - What can various organizations contribute?

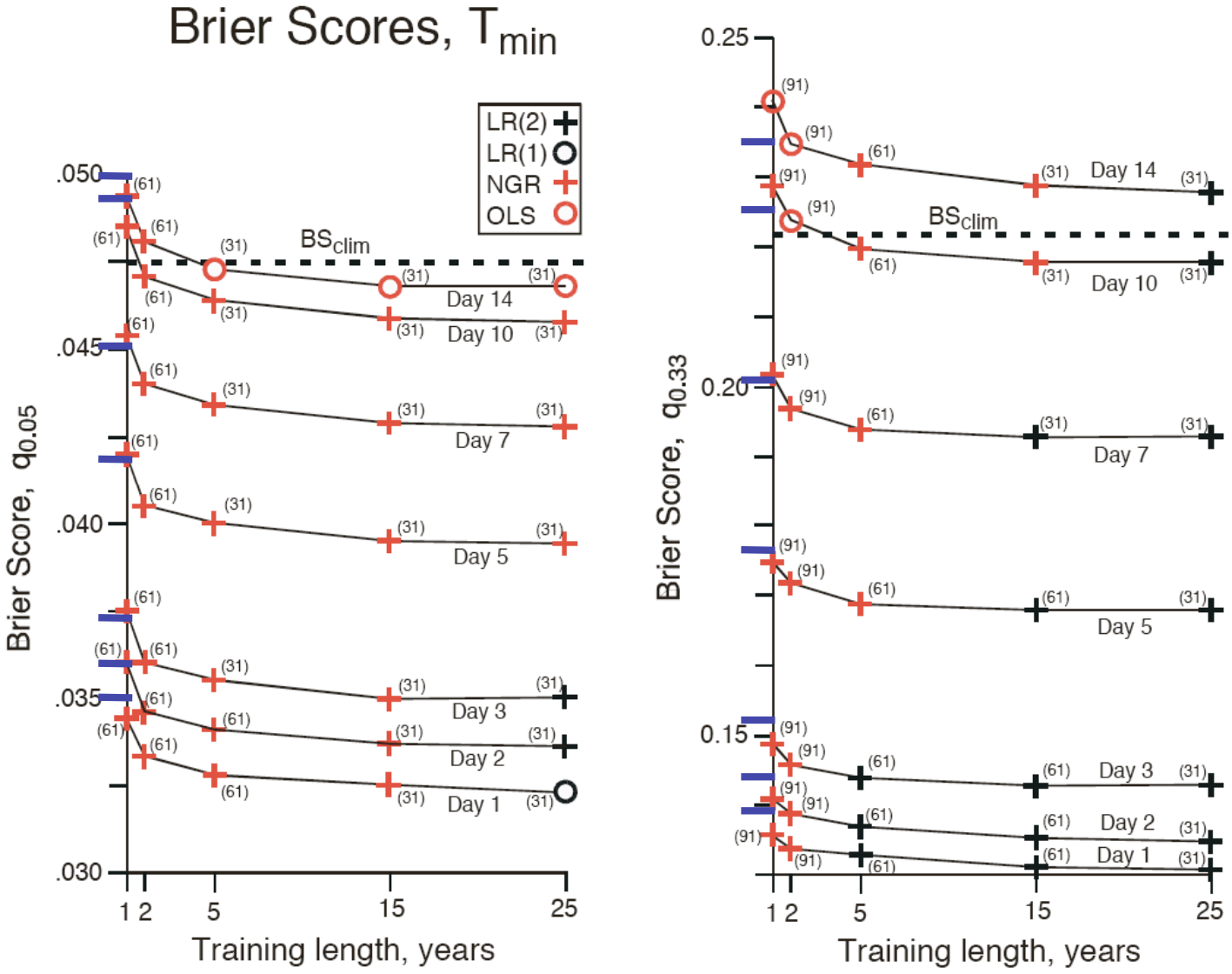
Will emphasize results that I don't think we've already showed many times before.

Questions based on differences in NCEP / ESRL calibration results

- Is a short time series of most recent forecasts really the best for short-range forecast calibration?

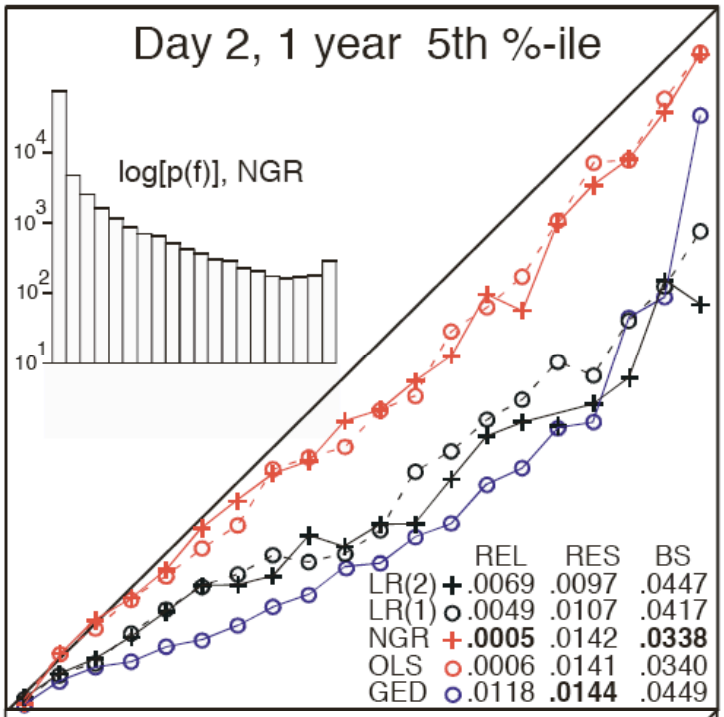
- Is the bulk of the benefit from correction of model bias, or other deficiencies in the ensemble system?

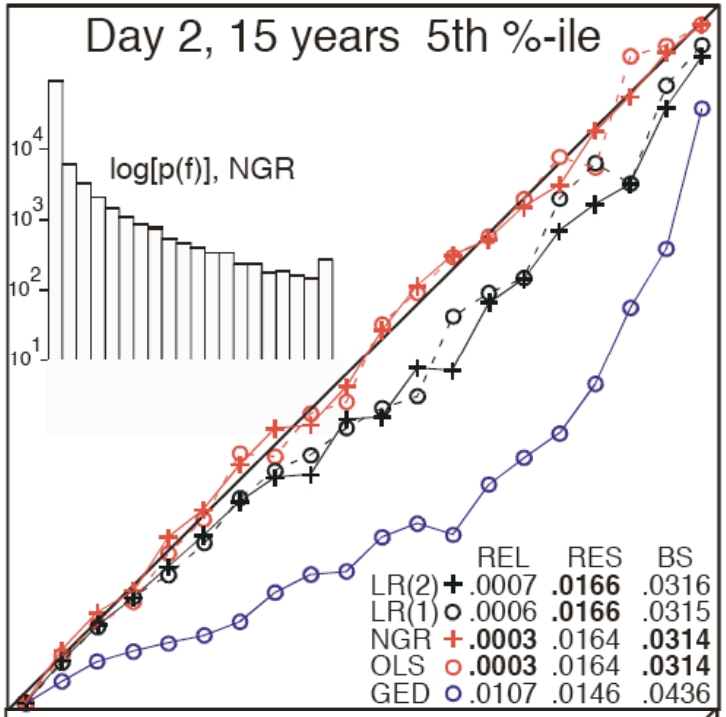
Concerns/issues: Are short training data sets better for short-lead forecasts?

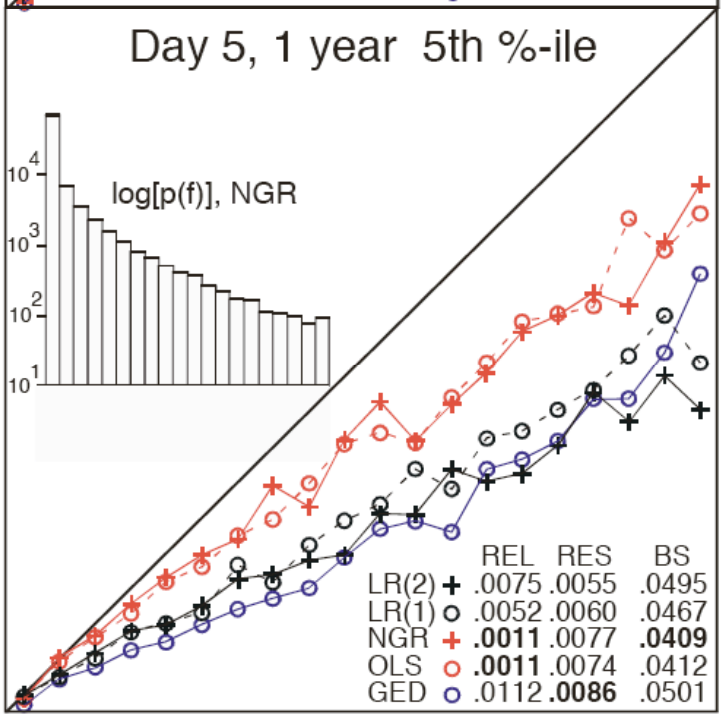


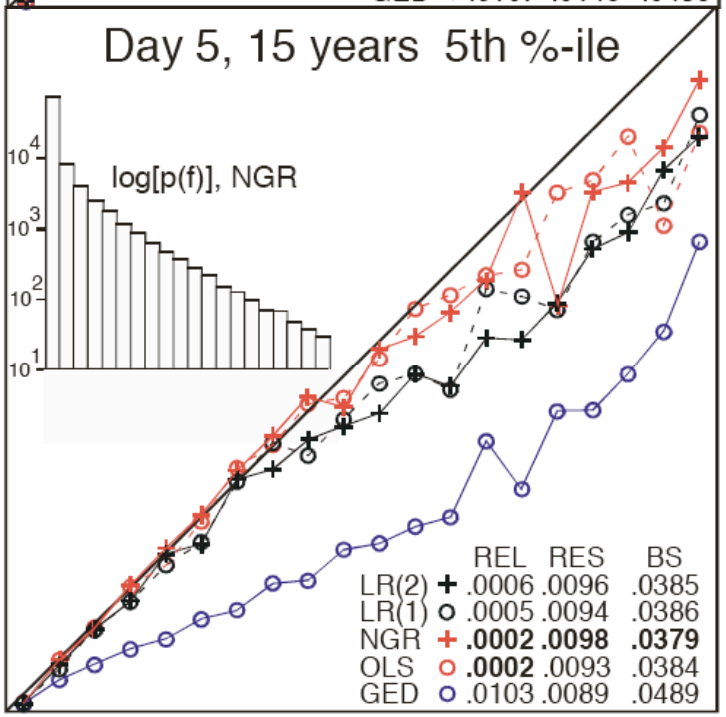
From Wilks and Hamill (2006); intercomparison of logistic regression, nonhomogeneous Gaussian regression, and ordinary least-squares regression for 2-m temperature forecasts.

Blue lines are 40-day training skill.


Day 2, 1 year 5th %-ile
log[p(f)], NGR
REL RES BS
LR(2) .0069 .0097 .0447
LR(1) .0049 .0107 .0417
NGR .0005 .0142 .0338
OLS .0006 .0141 .0340
GED .0118 .0144 .0449

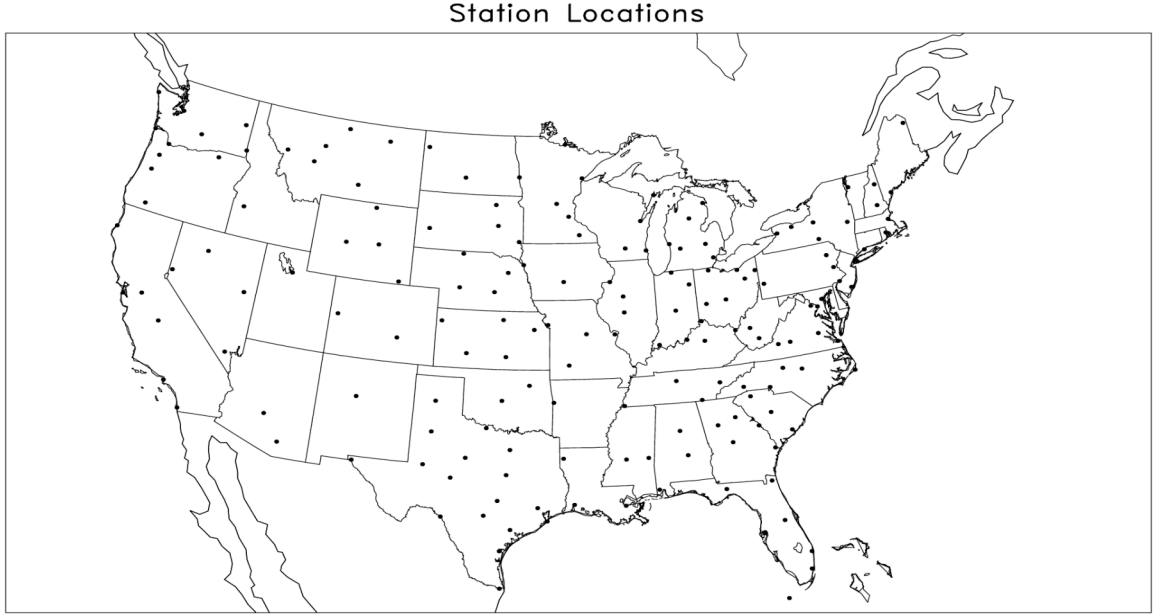

Day 2, 15 years 5th %-ile
log[p(f)], NGR
REL RES BS
LR(2) .0007 .0166 .0316
LR(1) .0006 .0166 .0315
NGR .0003 .0164 .0314
OLS .0003 .0164 .0314
GED .0107 .0146 .0436


Day 5, 1 year 5th %-ile
log[p(f)], NGR
REL RES BS
LR(2) .0075 .0055 .0495
LR(1) .0052 .0060 .0467
NGR .0011 .0077 .0409
OLS .0011 .0074 .0412
GED .0112 .0086 .0501

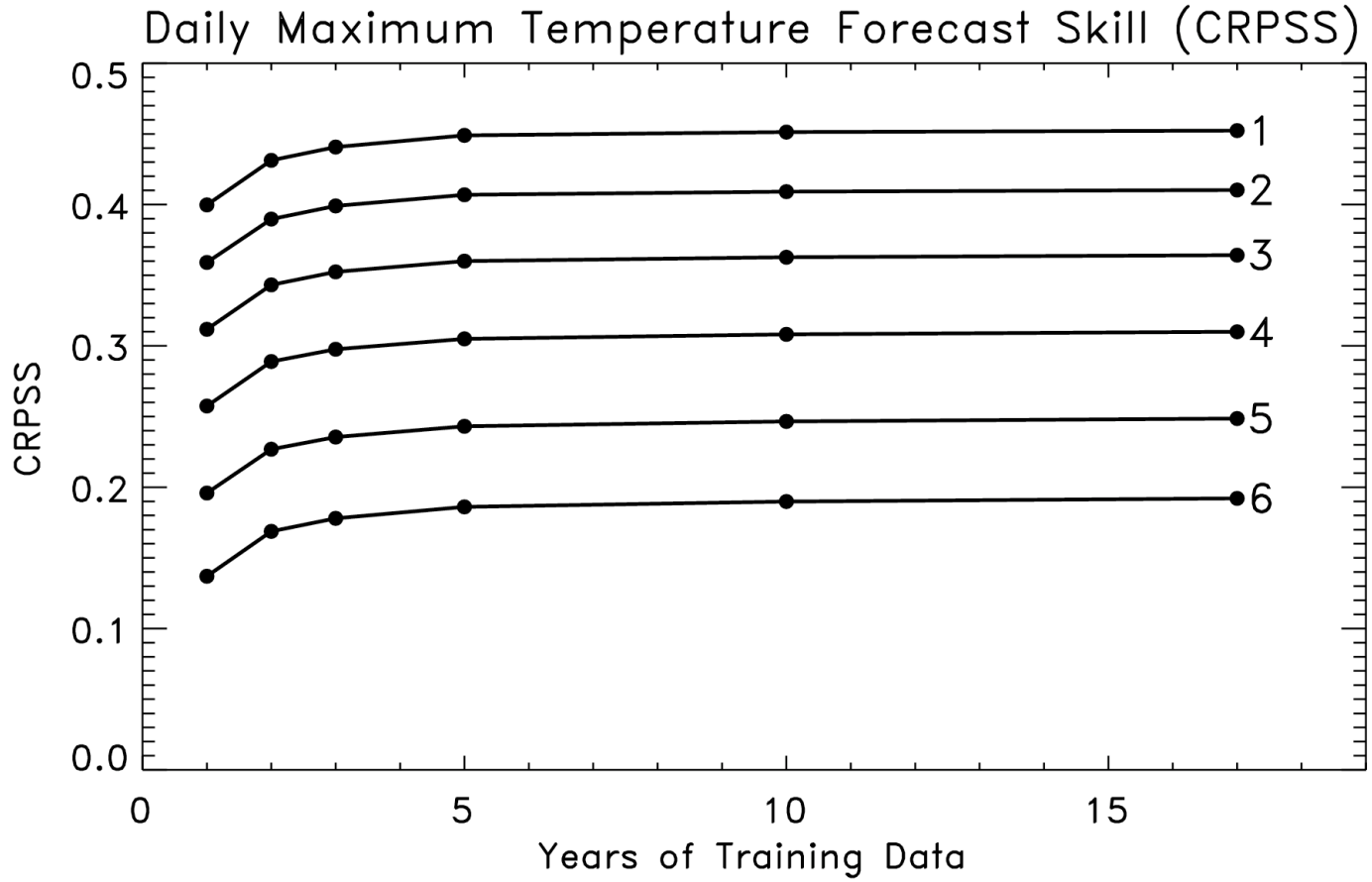

Day 5, 15 years 5th %-ile
log[p(f)], NGR
REL RES BS
LR(2) .0006 .0096 .0385
LR(1) .0005 .0094 .0386
NGR .0002 .0098 .0379
OLS .0002 .0093 .0384
GED .0103 .0089 .0489

T2M skill at Day 1 was so small that I repeated the experiment myself…

- Procedure:
  - Determine ensemble-mean forecast deviation from climatology, observed dev. from climatology.
  - Regress to predict observed deviation.
  - Form Gaussian pdf from regression.



Daily Max Temp CRPSS



Consistent 1-day improvement in lead from fully using reforecasts; (didn't trust T_{max} data before 1987).

A discrepancy…

- Prior NCEP results: better skill for short-lead forecasts using just the most recent forecasts as training data.

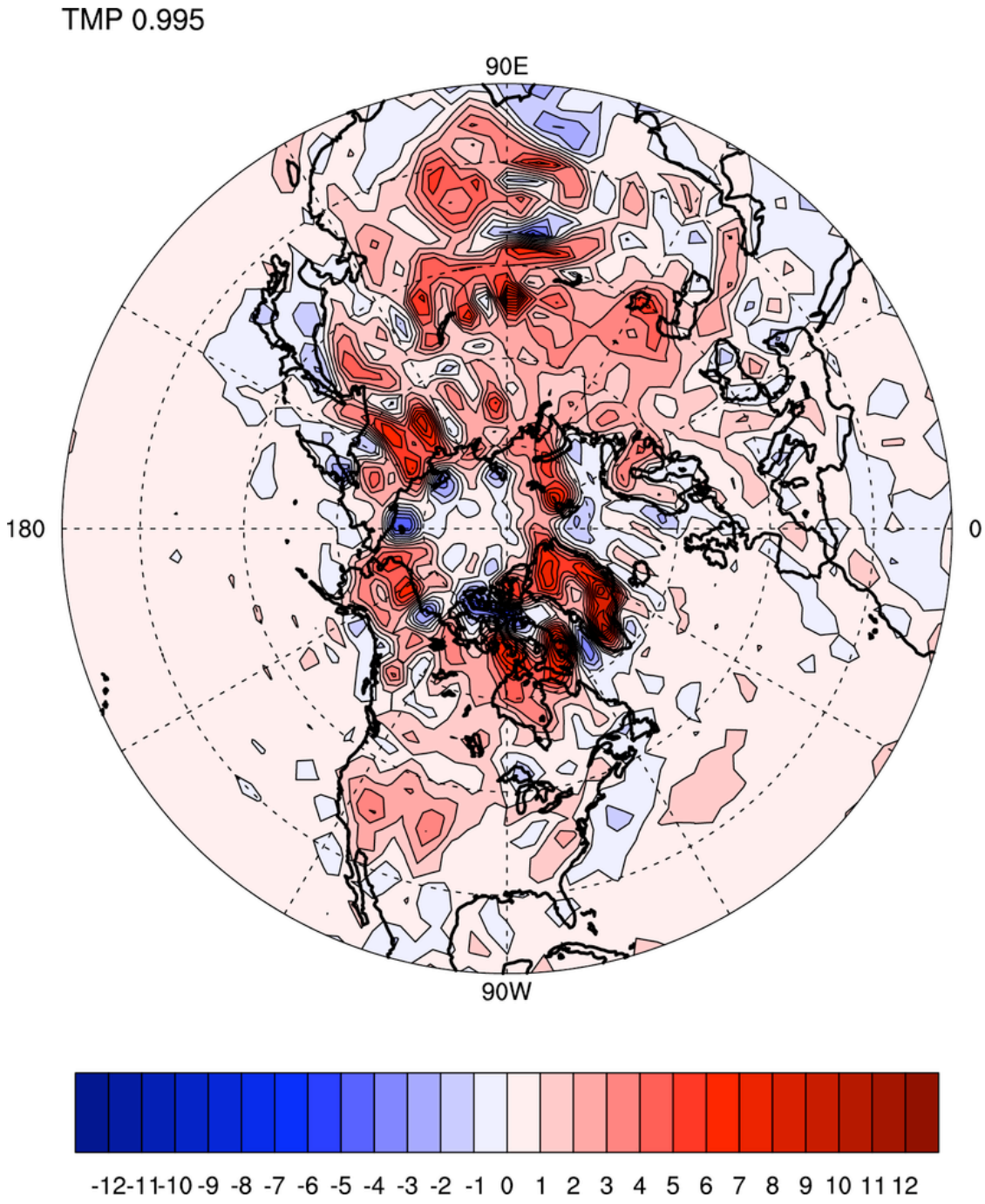
- ESRL results: short training data set provide far less benefit than long ones at all leads.

- Hypothesis: NCEP results may have used different analysis system for verification than for reforecasts. Strong biases between them.

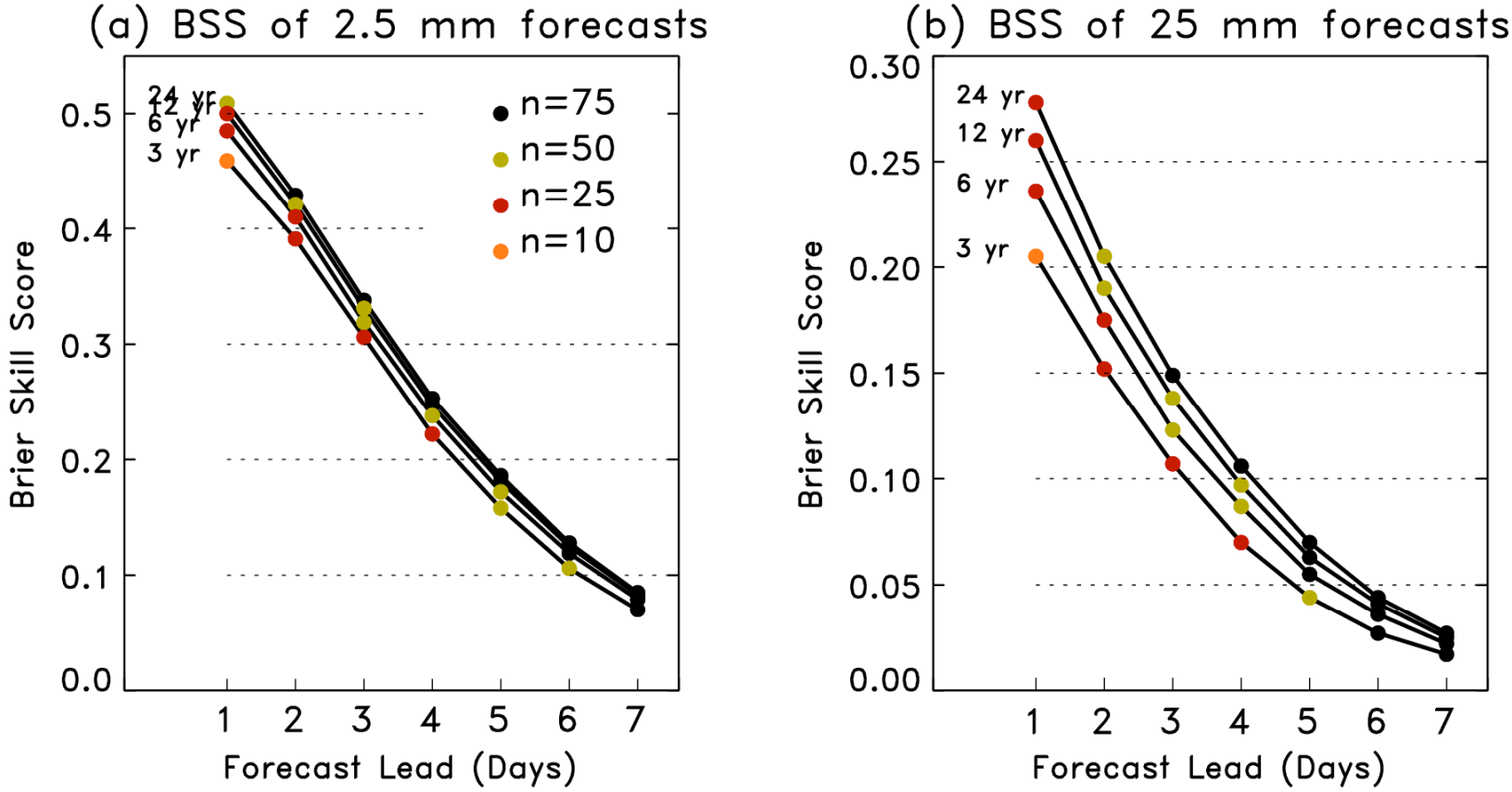
σ = .995, difference in climatology, Operational (→T62) - CDAS, January 2004

Very large differences, due to land-surface treatment and terrain differences in models with different resolutions. Both will be effects that will have to be dealt with if CDAS reanalyses are used with newer version of NCEP GFS.

This may explain NCEP's test results showing better performance using short training data set (since your short training data set used same analysis system as operational)



How does precipitation forecast skill change with training sample size?



Here, 2-step analog approach used; colors of dots indicate which size analog ensemble provided the largest amount of skill.

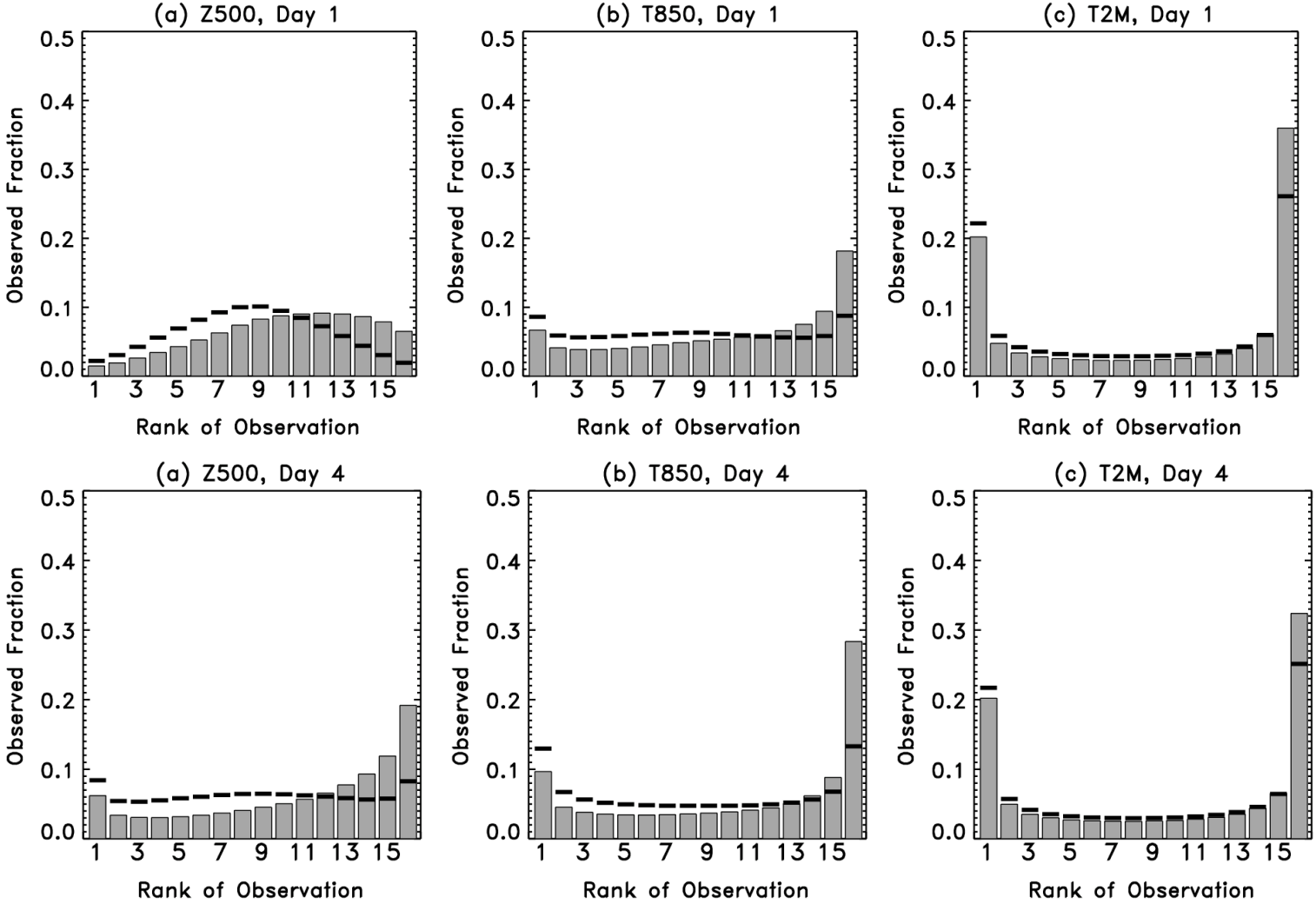
Does the impact of calibration depend upon the variable?

- Decided we wanted to look systematically at the differences in calibration between Z500, T850, T2M, all with long training data sets. Was Z500 “easier” ?

- How much from bias correction vs. full pdf calibration?

- Since NCEP did their calc’s over N. Hemisphere, we’d repeat using that area.

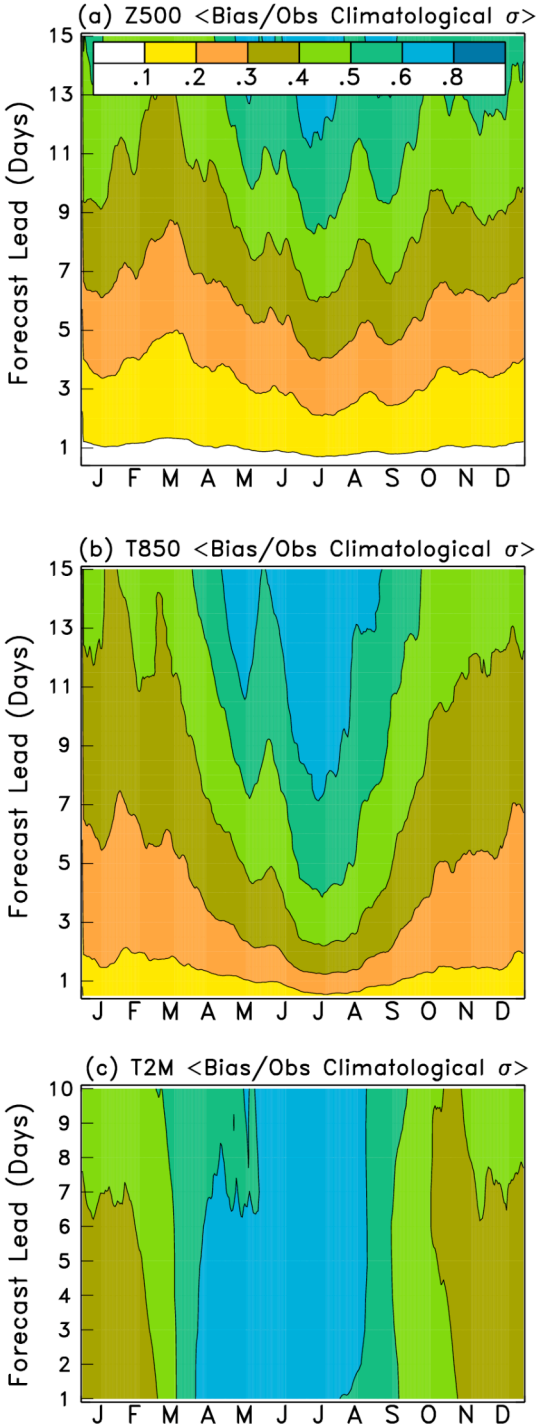
Calibration of Z_{500}, T_{850}, T_{2m}



Perturbed ensemble Z500 by 12.0m, T850 by 0.6C, T2M by 1.5C; solid lines are rank histogram after bias correction

Another piece of evidence why T2M is so much harder to calibrate than Z500.

T2M bias is large relative to the intrinsic climatological variability.



Calibration techniques

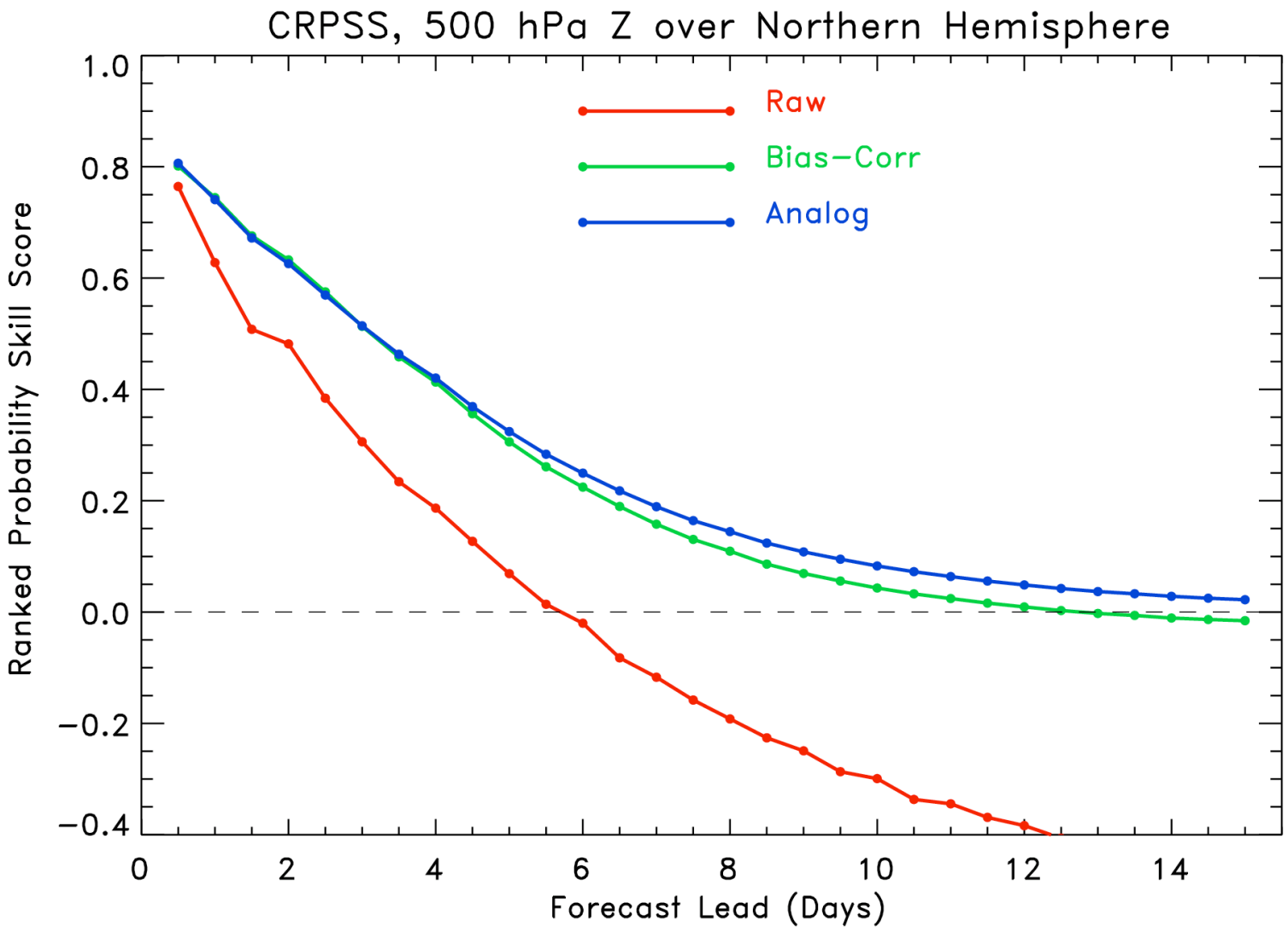
- Uncalibrated : PDF from raw ensemble
- Gross Bias Correction:
  - (1) Calculate Mean B = F - O
  - (2) Corrected ens = raw ens + B
- Analog method
  - Similar to method for precipitation, but now find forecast analogs using only the current grid point's data. 50 members.
  - Wilks and Hamill (2006, MWR, to appear) found that many other calibration methods (e.g. logistic regression, non-homogeneous Gaussian regression) were similar in performance.

Verification of Z_{500}, T_{850}, T_{2m}

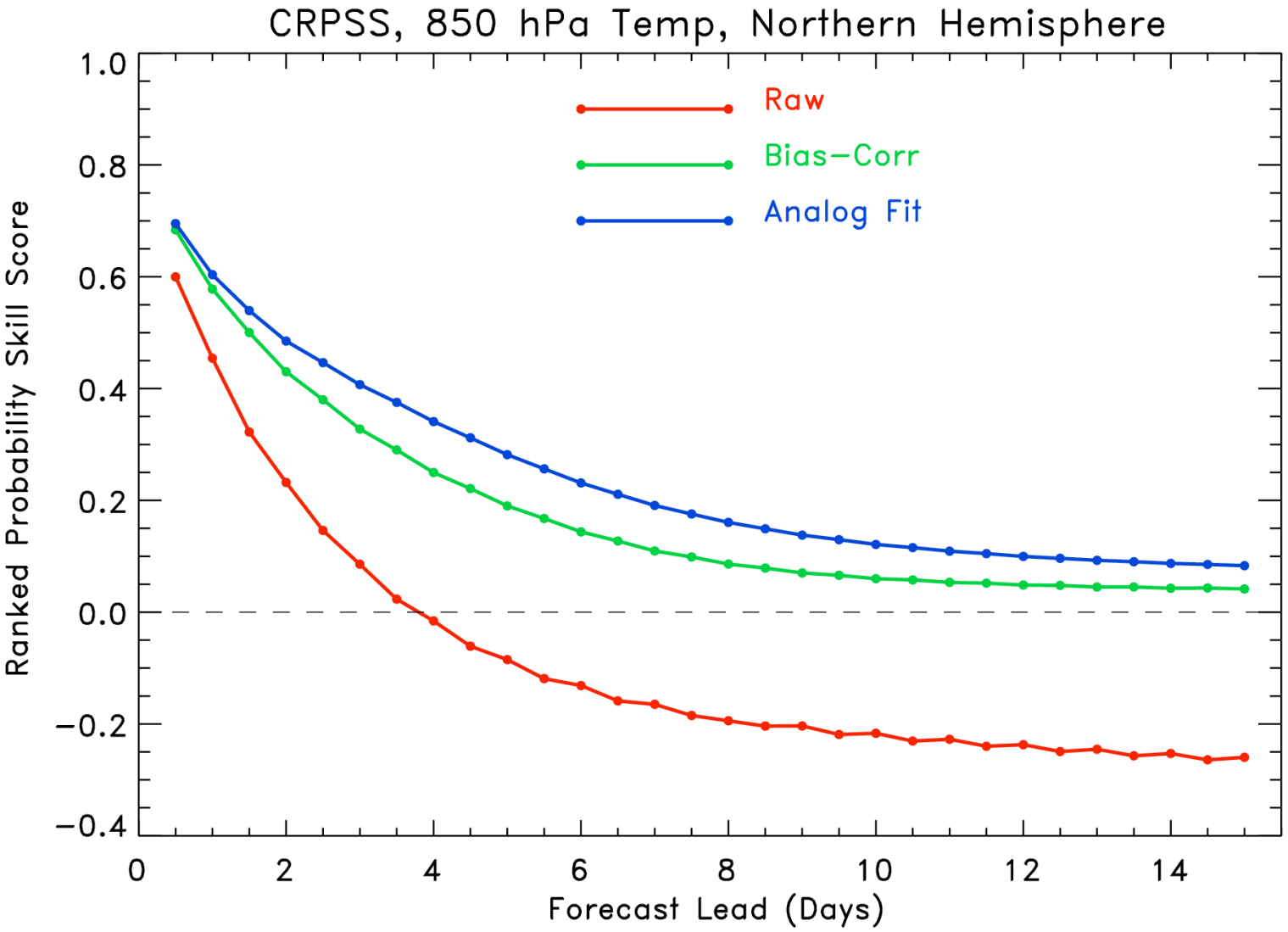
- Northern Hemisphere (Z_{500}, T_{850}); 00Z North American surface obs with > 97% complete record from 1979 - 2004 (T_{2m})

- Use continuous ranked probability skill score (CRPSS; 0=no skill, 1=perfect); use method of calculation in Hamill and Juras (2006, Oct. *QJRMS*) to avoid overestimating skill when climatology varies.

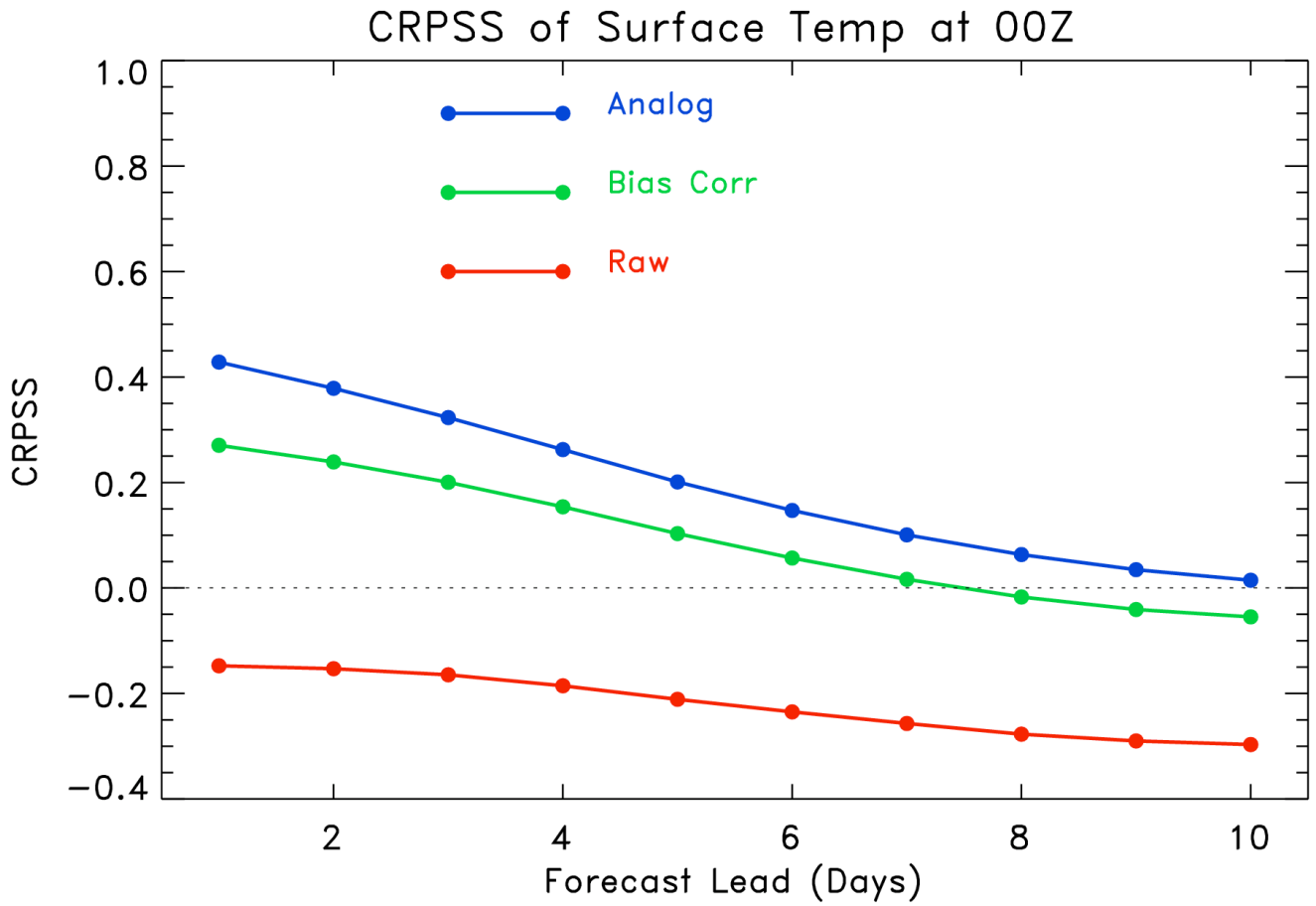
Z_{500} CRPSS


CRPSS, 500 hPa Z over Northern Hemisphere
Raw
Bias−Corr
Analog
Ranked Probability Skill Score
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10
12
14
Forecast Lead (Days)

T_{850} CRPSS


CRPSS, 850 hPa Temp, Northern Hemisphere
Raw
Bias−Corr
Analog Fit
Ranked Probability Skill Score
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
0
2
4
6
8
10
12
14
Forecast Lead (Days)

T_{2m} CRPSS

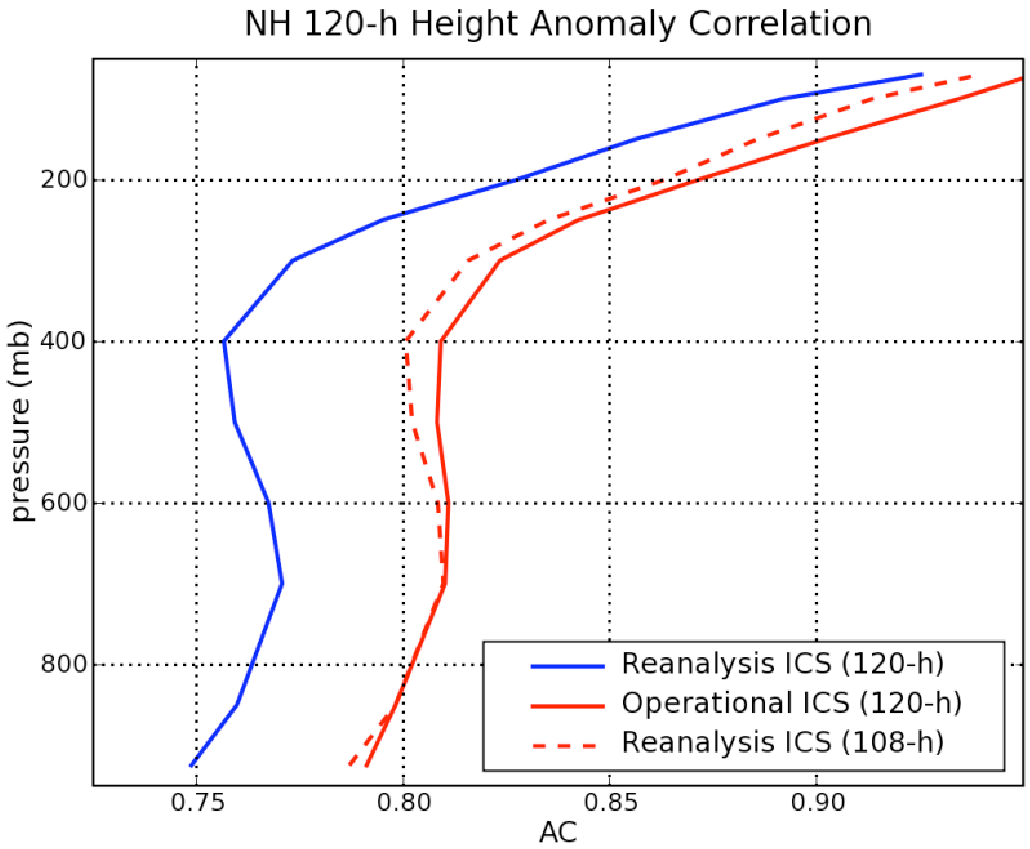


Impact of calibration largest for surface variables

Configuration for reforecasts: between ideal and practical

- **Reanalyses**: why redoing them is an important part of reforecast process.
  - Forecast skill improving partly due to better initial conditions. Reforecasts will leverage that.
  - Different biases in different analysis systems, especially near the surface. Reforecasts with old reanalysis initial conditions thus start with biased initial conditions relative to real-time forecasts.
  - Reforecasts may be the way of highlighting reanalysis importance - doing reanls improves skill
- Large ensemble of reforecasts necessary?

Forecasts based on new vs. old analysis systems



- Want homogeneous characteristics of forecasts; skill the same for 1980's forecasts as 2006 forecasts.

- Part of better skill of current forecasts is the better initial condition.

- Reanalysis would improve skill of old forecasts.

- Reanalyses should use same or similar model as used in reforecasts.

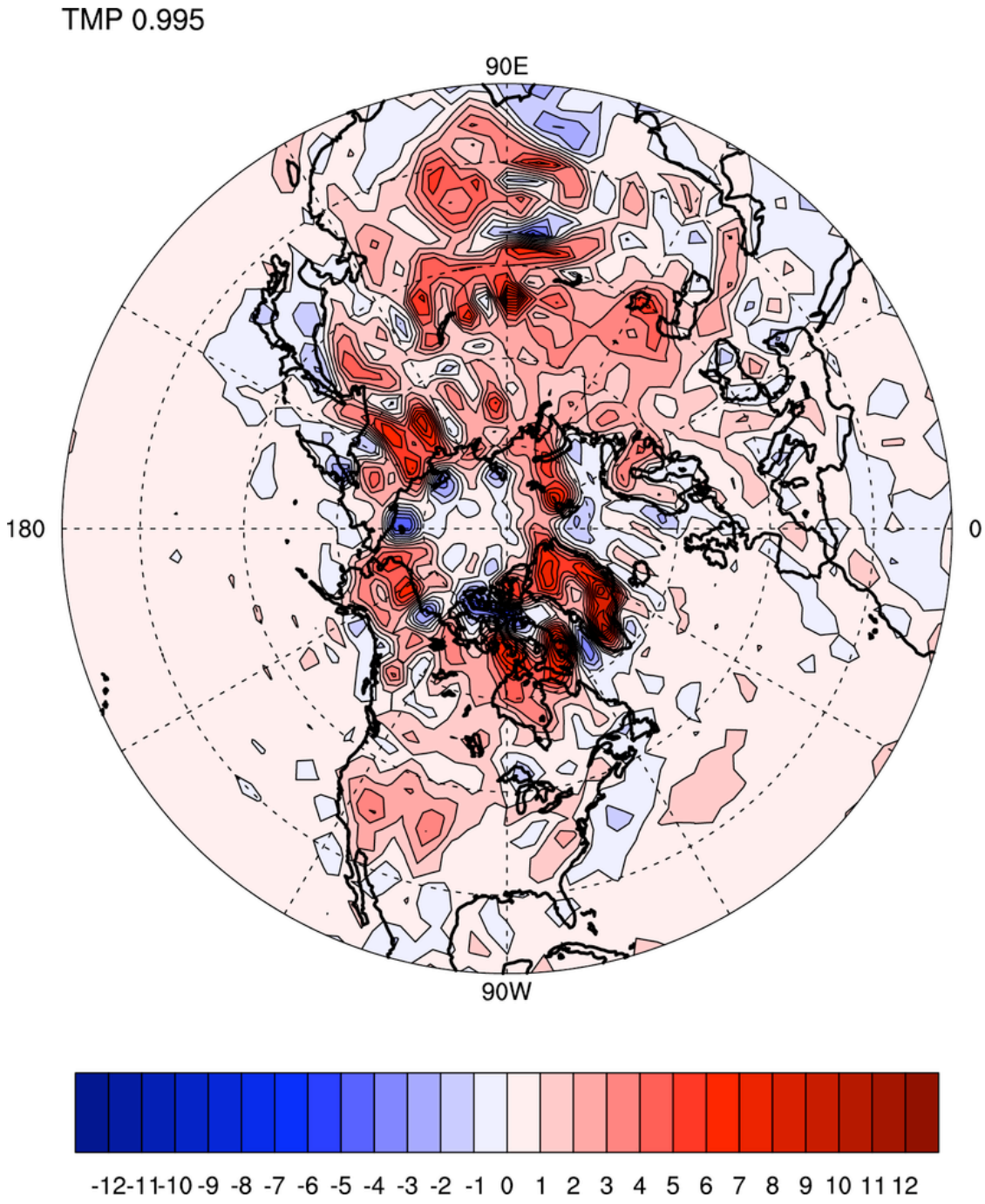
Biases in reanalyses

- Especially near the surface, reanalyses can have substantial bias, reflecting bias of forecast model.
  - $\rightarrow$ Analysis systems don't use much near-surface data, so much information from 1st guess.
  - $\rightarrow$ Near-surface temperatures reflect specific choice of boundary-layer, land-surface parameterizations.
  - $\rightarrow$ Should expect that near-surface temperature forecasts from reanalyses will differ from those from current analysis, especially for short-lead forecasts, before equilibration.

σ = .995, difference in climatology, Operational (→T62) - CDAS, January 2004

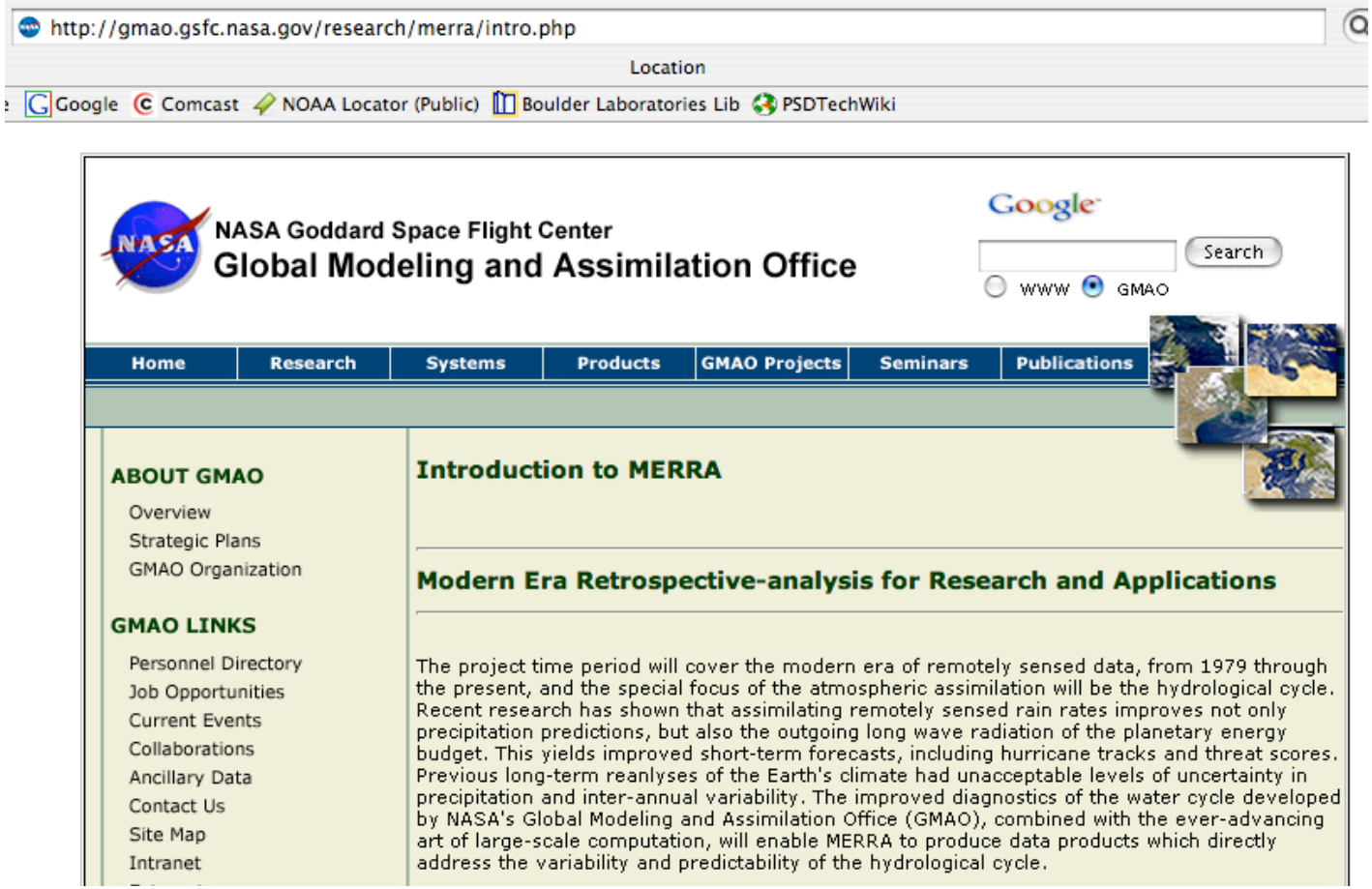
Very large differences, due to land-surface treatment and terrain differences in models with different resolutions. Both will be effects that will have to be dealt with if CDAS reanalyses are used with newer version of NCEP GFS.

This may explain NCEP's test results showing better performance using short training data set (since your short training data set used same analysis system as operational)



Would reanalyses be *that* difficult?

- NASA's doing MERRA, presumably saving QC'ed obs, which is a large part of the effort. Can NCEP use MERRA QC'ed obs for quick reanalysis?



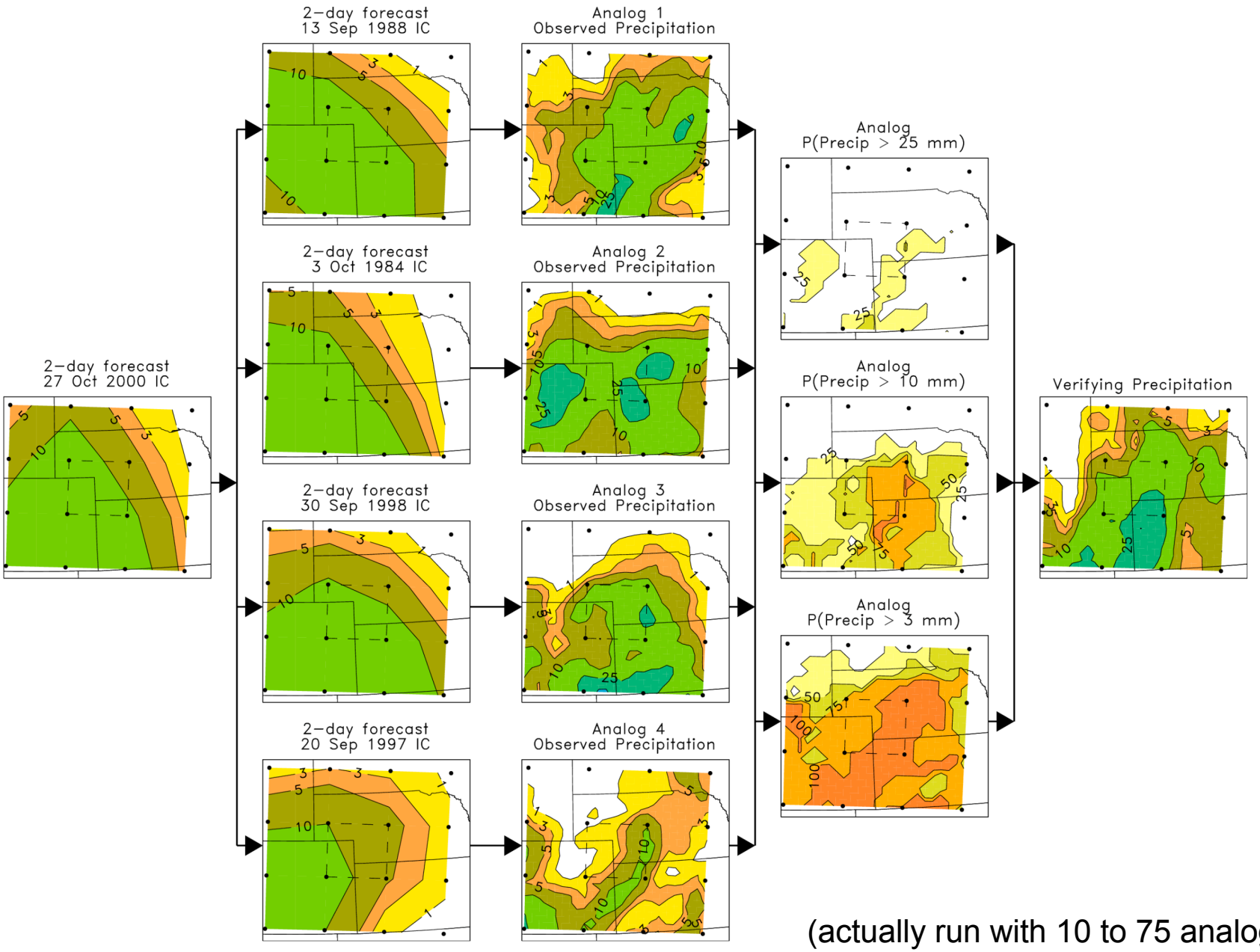
Our strong recommendation

- Do reanalyses as companion to reforecasts right from the start.
- Cannot assume that a reforecast system without reanalysis will give you large calibration benefits
  - strongly encourage your own testing (current GFS w. CDAS reanalyses) before committing to specific plan.

Issue: How many members are needed in reforecast?

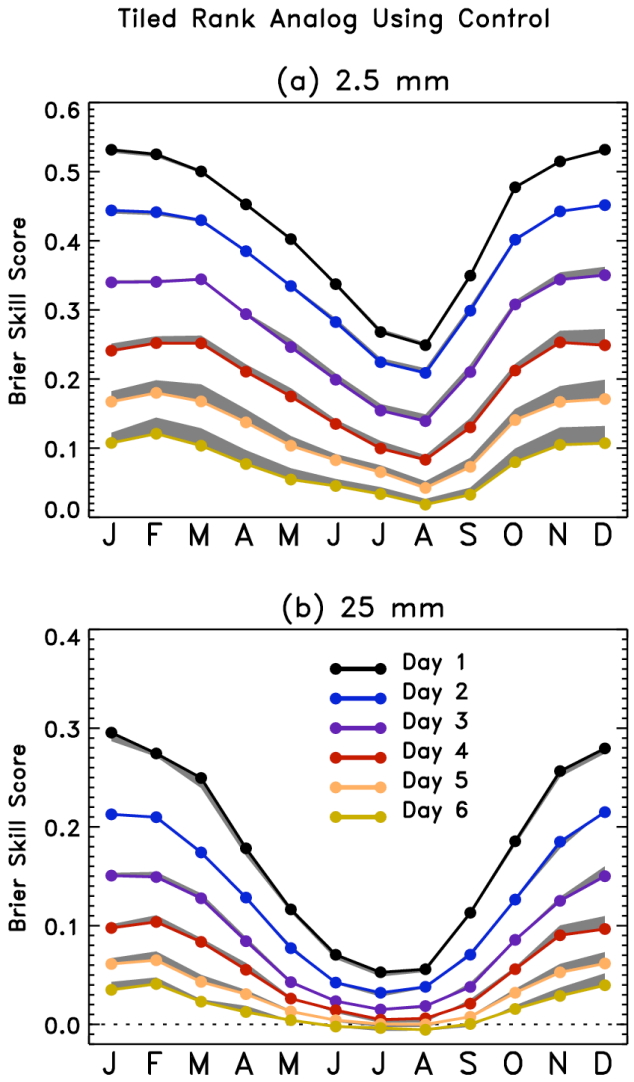
Ideally, reforecast ensemble is as large as the real-time ensemble. But can we develop calibration techniques that are effective with fewer members?

Analog high-resolution precipitation forecast calibration technique



(actually run with 10 to 75 analogs)

Precipitation forecasts: how many members are needed?

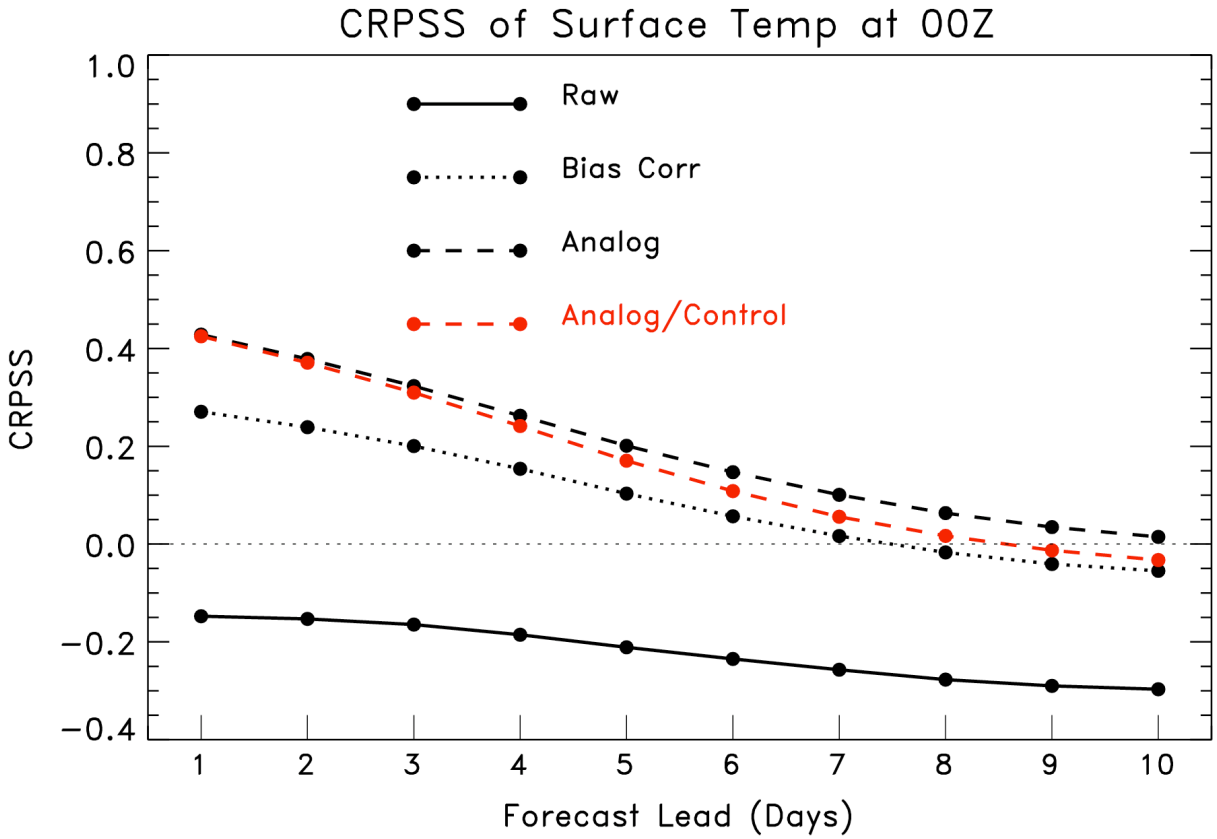


Analog reforecast process repeated, as in prior cartoon. But now rather than matching ensemble-mean pattern, match today's control forecast to past control forecast.

Grey area measures degradation relative to baseline using ensemble mean.

Not much degradation in skill, esp. at short leads! (and you don't even have to run an ensemble to get a probabilistic forecast).

Temperature forecasts: how many members?



Statistical correction of control forecast using control reforecast is nearly as skillful at short leads, less skillful at longer leads.

(May not have fully exploited value of control reforecast, however.)

Moving toward operational implementation at NCEP: issues

- Companion reanalysis possible? (we'll keep harping on this…)
- Is NCDC ready for reforecast storage? Does NOMADS have bandwidth to receive/transmit reforecasts?
- Will reforecast access from ESRL, MDL be simple and convenient?
- Are all reforecasts, all levels guaranteed to be saved, or just selected subsets? What subsets?
- Would NCEP like any ESRL library routines, e.g., for analog calibration?

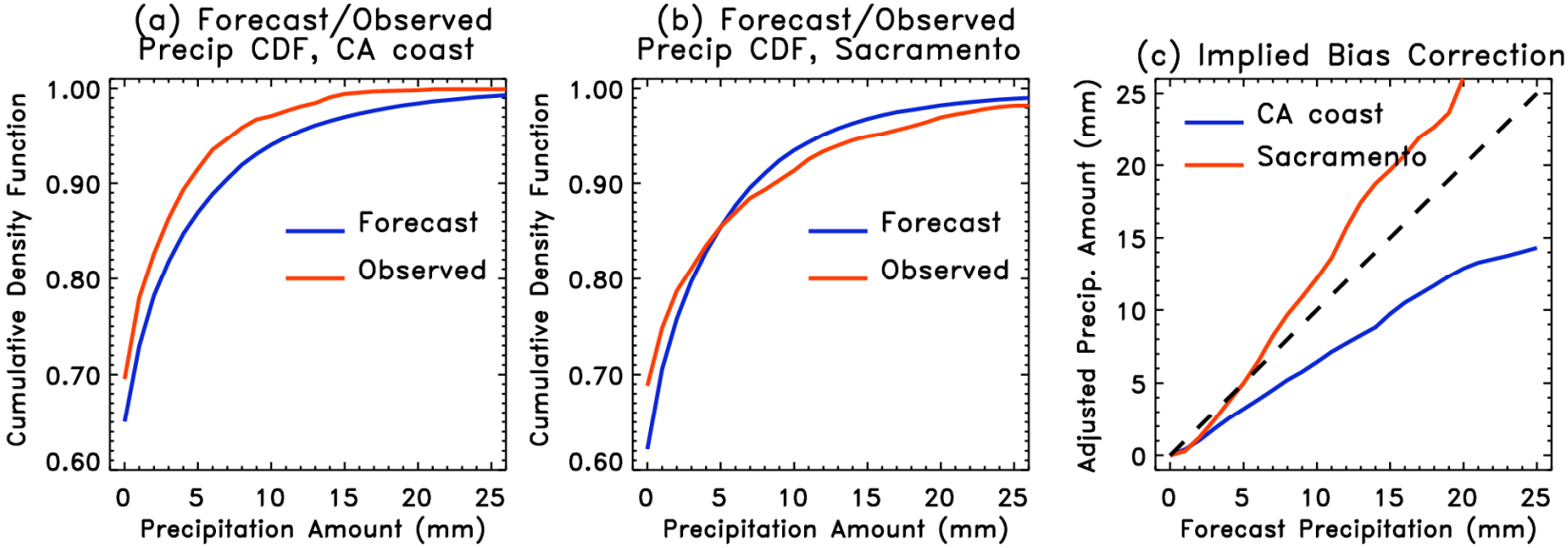
The bigger picture

- NOAA needs coordinated probabilistic forecast program. What is the role of reforecast-based ensemble products?
  - Bias correction of NAEFS ensemble products? [simple, NCEP can handle]
  - Suite of multiple, statistically adjusted probabilistic products pumped out to WFOs via NDGD? [complex, cross-NOAA, need coordinated plan involving NWS/OST, NCEP, MDL, ESRL, etc.]. Canada?

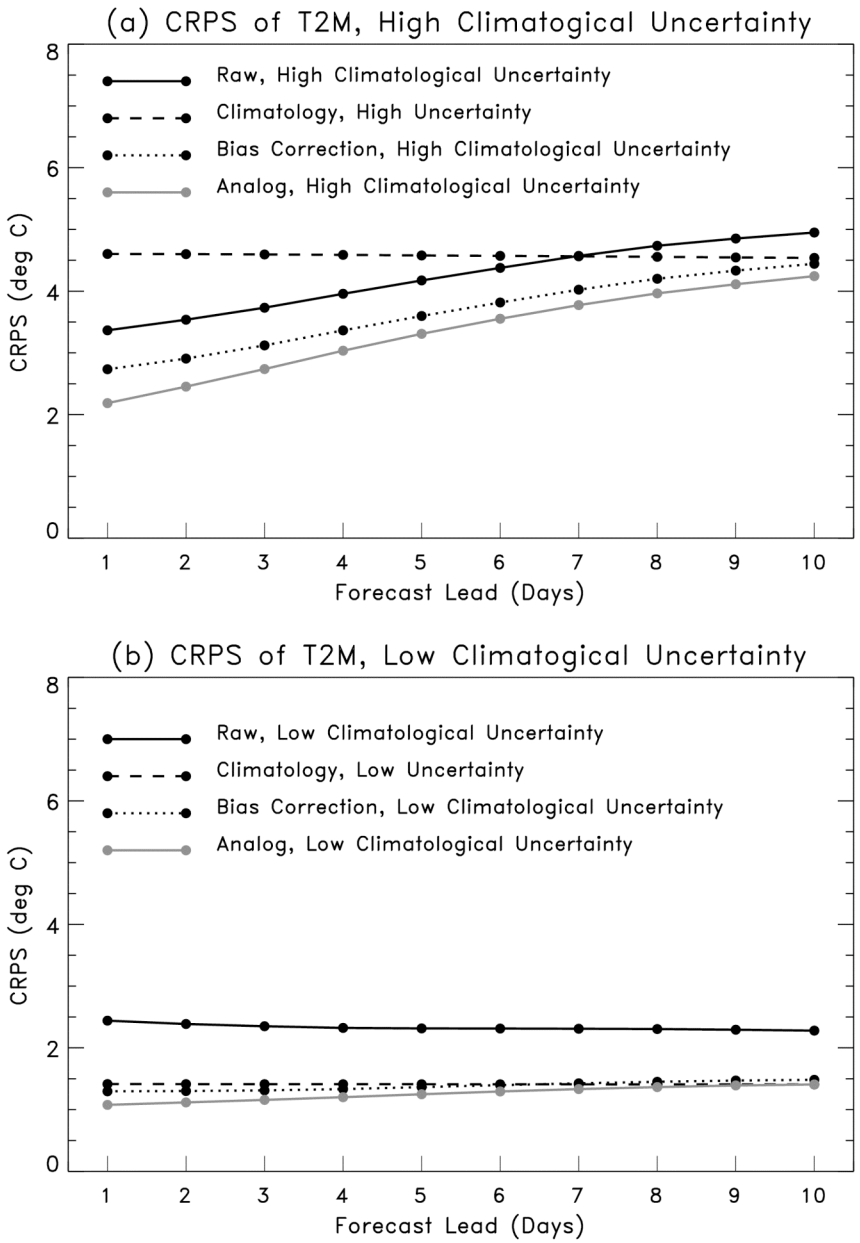
Conclusions

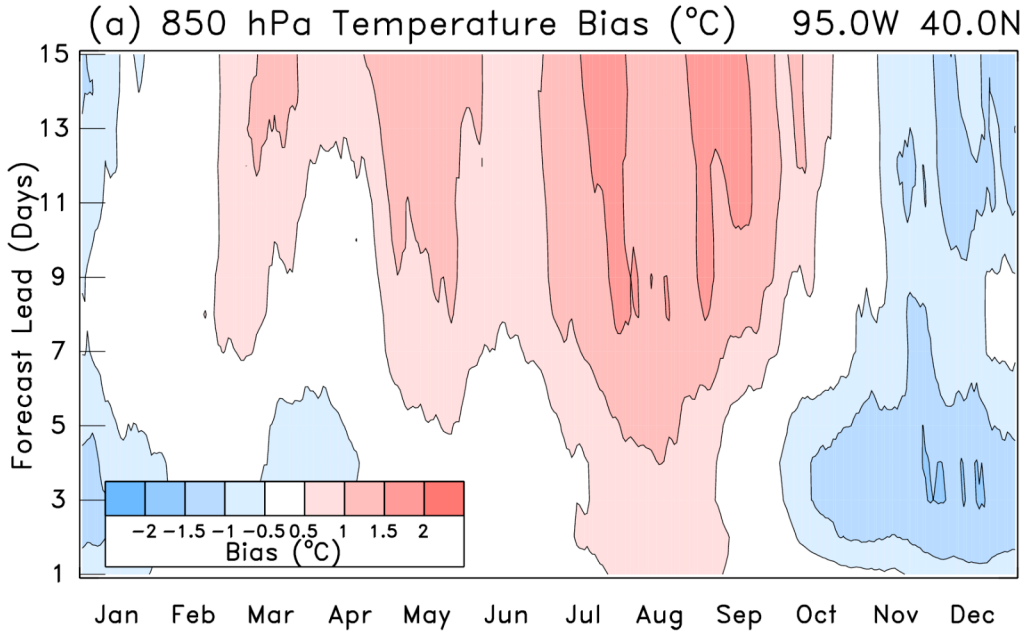
- Reforecasts will dramatically help NWS meet NRC guidelines for calibrated, skillful probabilistic forecast products.
- Reanalyses -- a necessary part of reforecast procedure.
- And end-to-end procedure for widely disseminating probabilistic products is a big endeavor, and all parts of NOAA should participate in a coordinated plan.

Bias correction using forecast and observed CDFs?

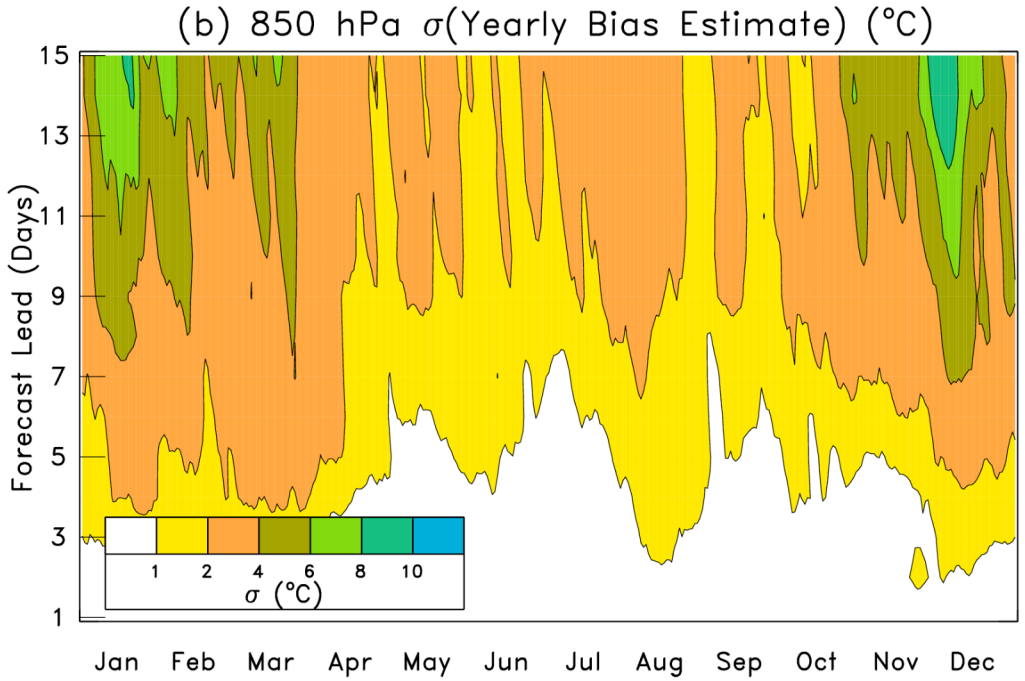


T_{2m} CRPSS, low and high climatological spread





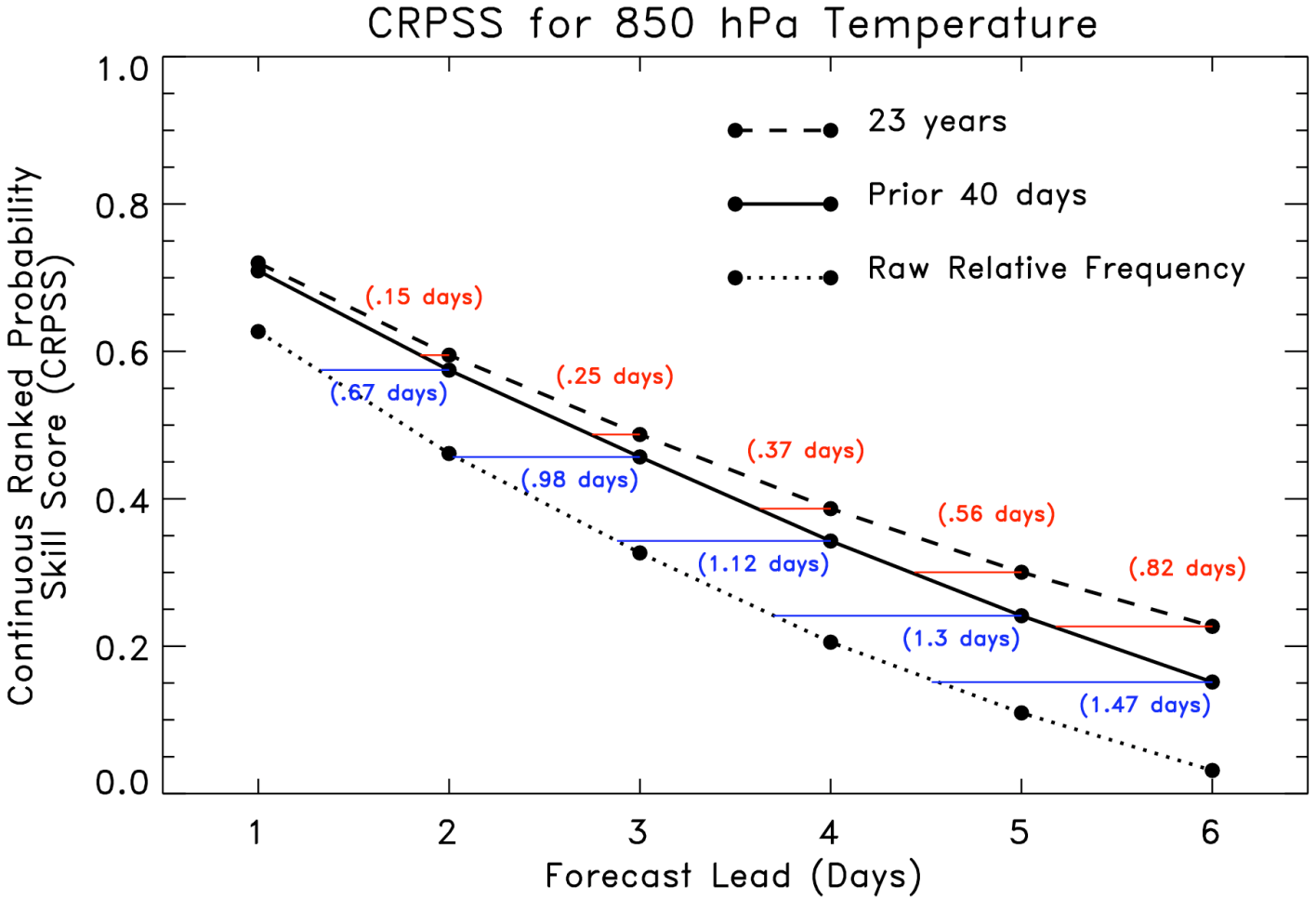
850 hPa temperature bias for a grid point in the central U.S.



Spread of yearly bias estimates from 31-day running mean F-O

Note the spread is often larger than the bias, especially for long leads.

Effects of short / long training, T850



This was a quick-and-dirty study using T850 data over CONUS, NCEP-NCAR reanal. for verification. Not as much benefit from many years of data as seen with Tsfc.