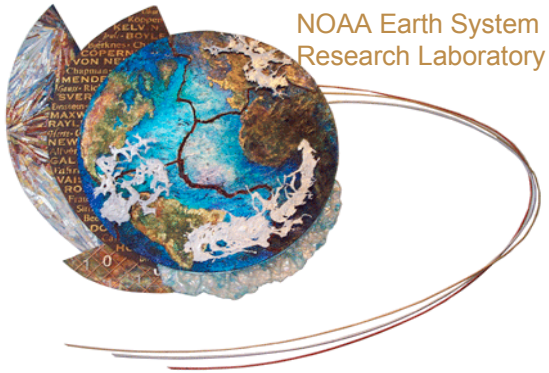


Ensemble Weather Forecasting and PQPF: Using Reforecasts

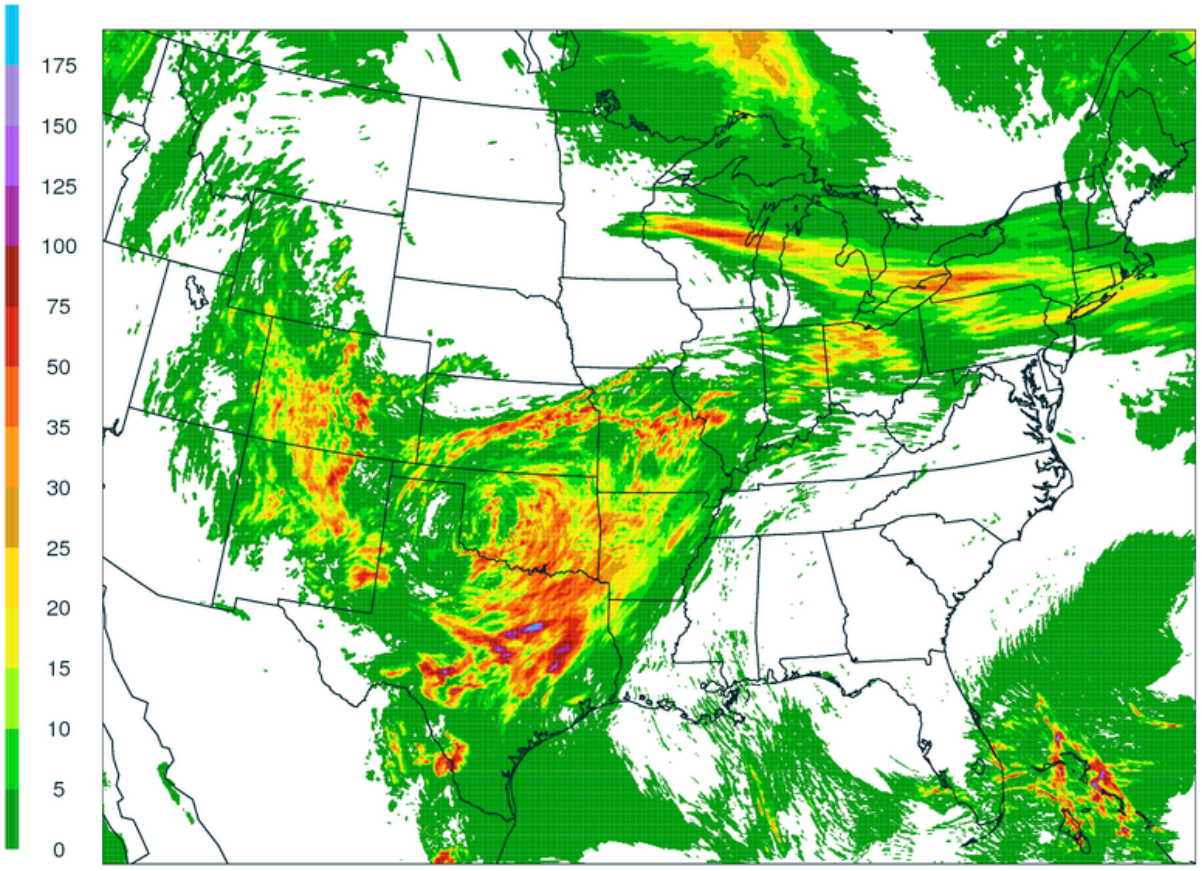
Tom Hamill

NOAA Earth System Research Lab

Boulder, CO

tom.hamill@noaa.gov

Two grand successes of NWP:
(1) Improved, high-resolution forecast models



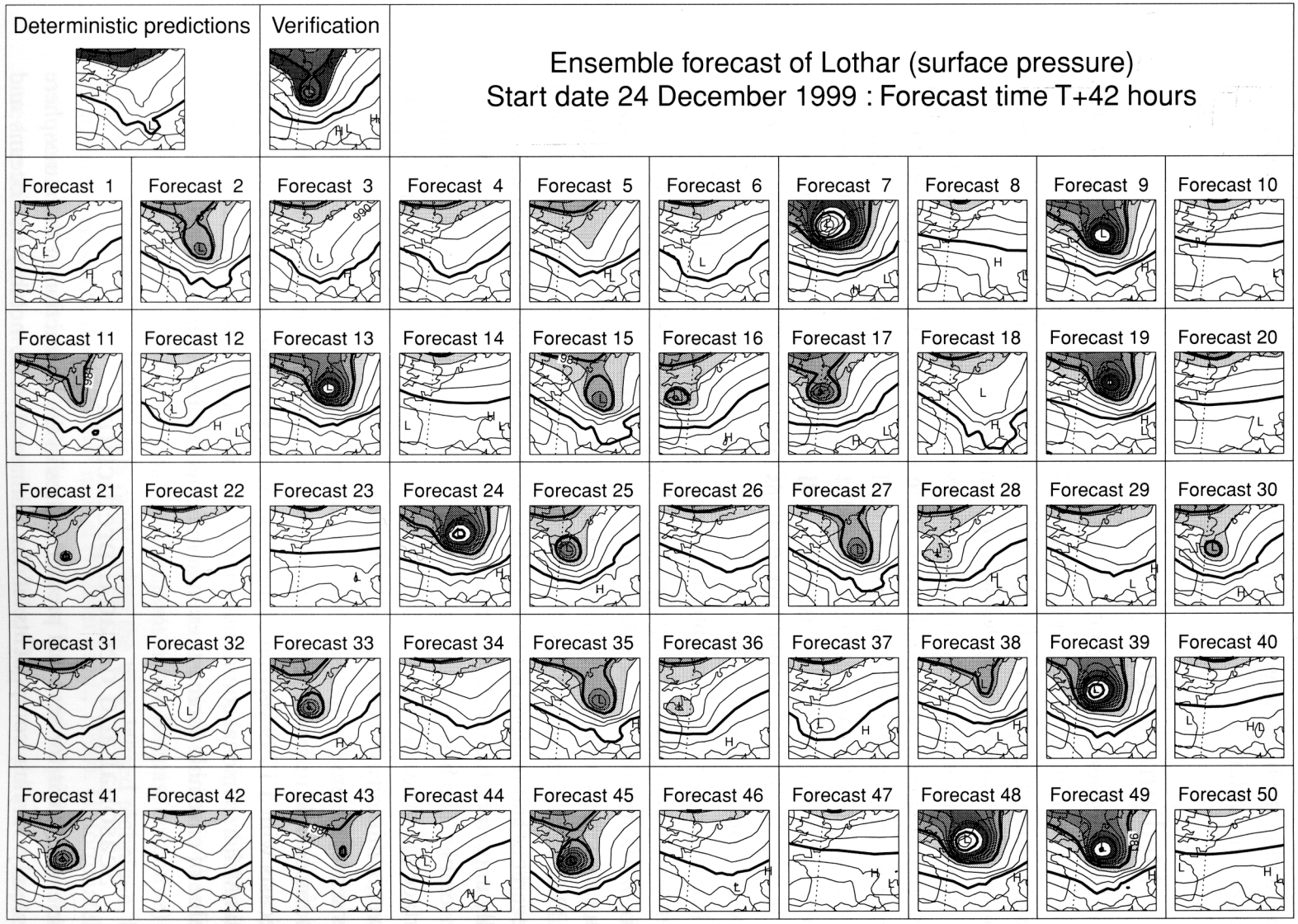
We now have models with explicit convection that produce forecasts that look, for the first time, like radar images of precipitation. This is probably not a coincidence.

Grand success (2): ensemble forecasts

Multiple simulations of the weather from slightly different initial conditions, perhaps different forecast models

Deterministic forecast totally misses damaging storm over France; some ensemble members forecast it well.

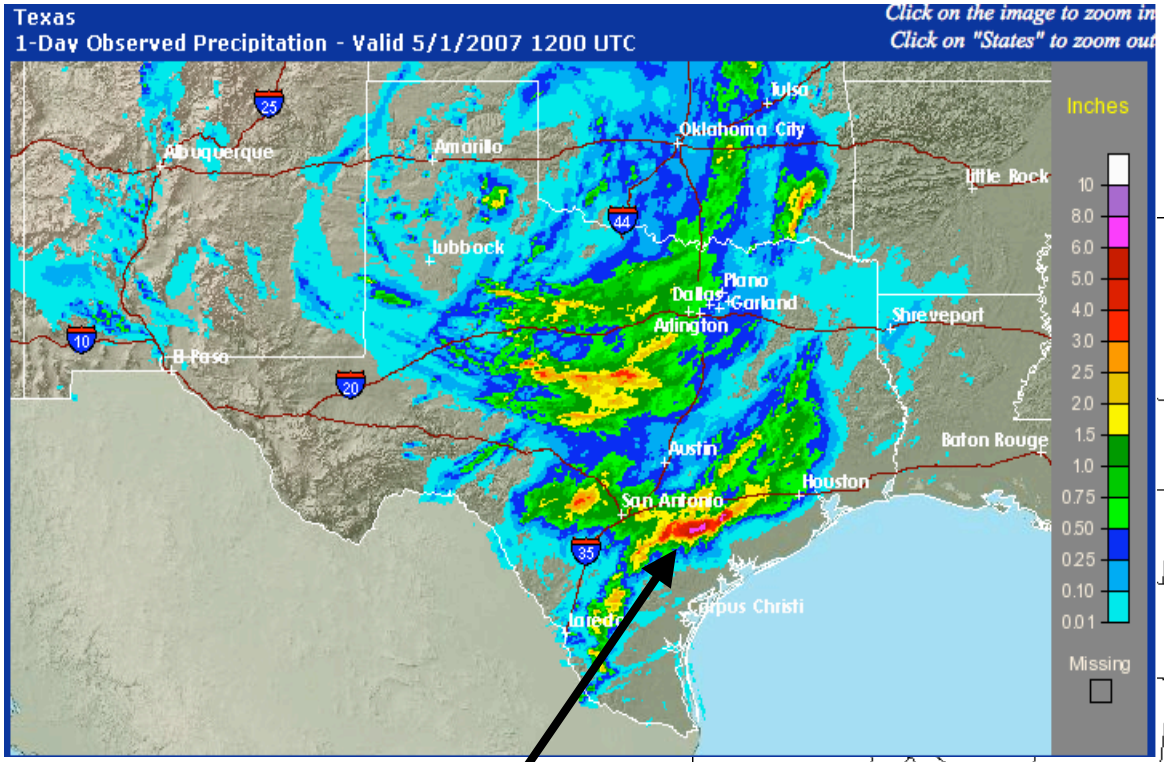
Probabilities commonly estimated from frequency of event in the ensemble.

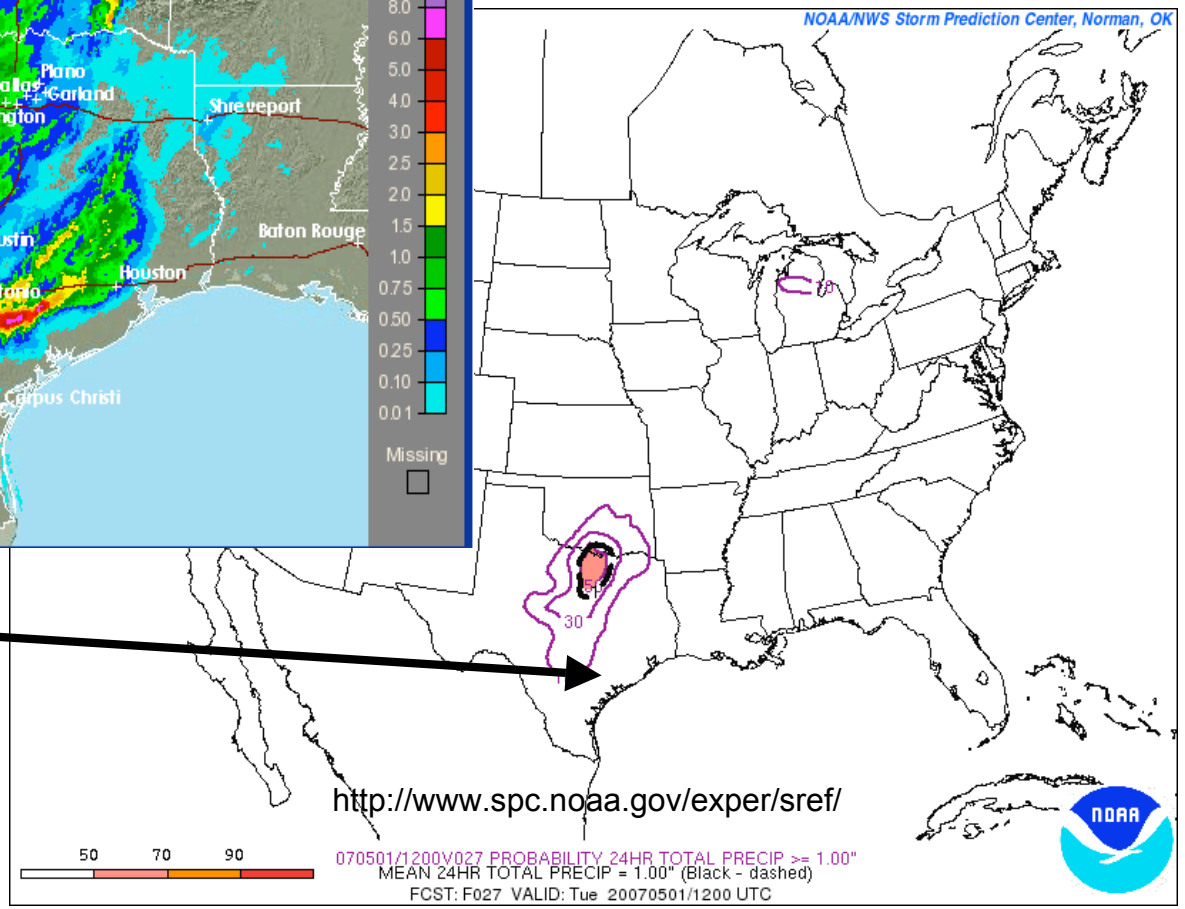


from Tim Palmer's book chapter, 2006.

Problem with current ensemble forecast systems

Forecasts may be biased and/or deficient in spread, so that probabilities are mis-estimated. More so for surface temp & precip. than Z500





Heavy rain in an area where none of the ensemble members predicted it.

http://www.spc.noaa.gov/exper/sref/

What I think hydrologists want

- An ensemble of data to feed into ensemble streamflow models, rather than just probability forecasts.
- Reliability (when the frequency of this ensemble says P=90%, it happens 90 % of the time).
- Sharpness (more 0 and 100%, less of climatological probability, if still reliable).
- Geographic specificity, to the extent it's predictable (e.g., more snow in west Boulder than east Boulder).
- Correct spatial and temporal correlations.

Possible paths forward

(1) Use CPU resources to rapidly develop higher-resolution ensembles with improved physical veracity. Improve methods of generating initial conditions, generate ways of dealing with uncertainty of the forecast model itself. [What we've been doing]

(2) Use those CPU cycles to run a fixed model and data assimilation system, albeit an older, low-resolution one. Run real-time, plus many past forecast cases. Diagnose the forecast error characteristics and generate statistically adjusted forecasts ("*reforecasting*")

(3) Compromise between the two.

Approach 1:
(Building and continually improving a highest-resolution ensemble)

- ADVANTAGES :
  - (1) CPU cycles dedicated to forecasts at highest resolution, with best physics.
  - (2) Small-scale features may actually be *resolved* by the model, rather than inferred from larger-scale conditions and "statistical voodoo."
  - (3) As soon as improved model is developed, it can be implemented.

- DISADVANTAGES :
  - (1) Raw probability forecasts biased. And don't expect bias $\rightarrow$ 0 with the next implementation.
  - (2) Correction of model problems difficult for human (or computer) to estimate without a long, careful look.
  - (3) Rapid changes $\rightarrow$ little experience with model before next version.
  - (4) Resolving a feature $\neq$ successfully predicting a feature. You may be led into a sense of overconfidence by high-resolution model.

Approach 2:
Reforecasting (correcting our mistakes)

- ADVANTAGES:
  - (1) Preliminary results show that the equivalent of > 10 yrs. of NWP model development can be obtained through judicious forecast calibration with a large set of reforecasts.
  - (2) Can nearly eliminate bias & spread deficiencies, downscale.
  - (3) End users like a stable, known product, and the forecast characteristics of reforecast-based products won't change often.

- DISADVANTAGES:
  - (1) Major improvements may not be able to be implemented quickly. If new model, must take the time to run reforecasts (expensive).
  - (2) Processes that form precipitation, like thunderstorms, can't be resolved, and must be parameterized.
  - (3) You learn much about the error characteristics of an old model, not a new one.

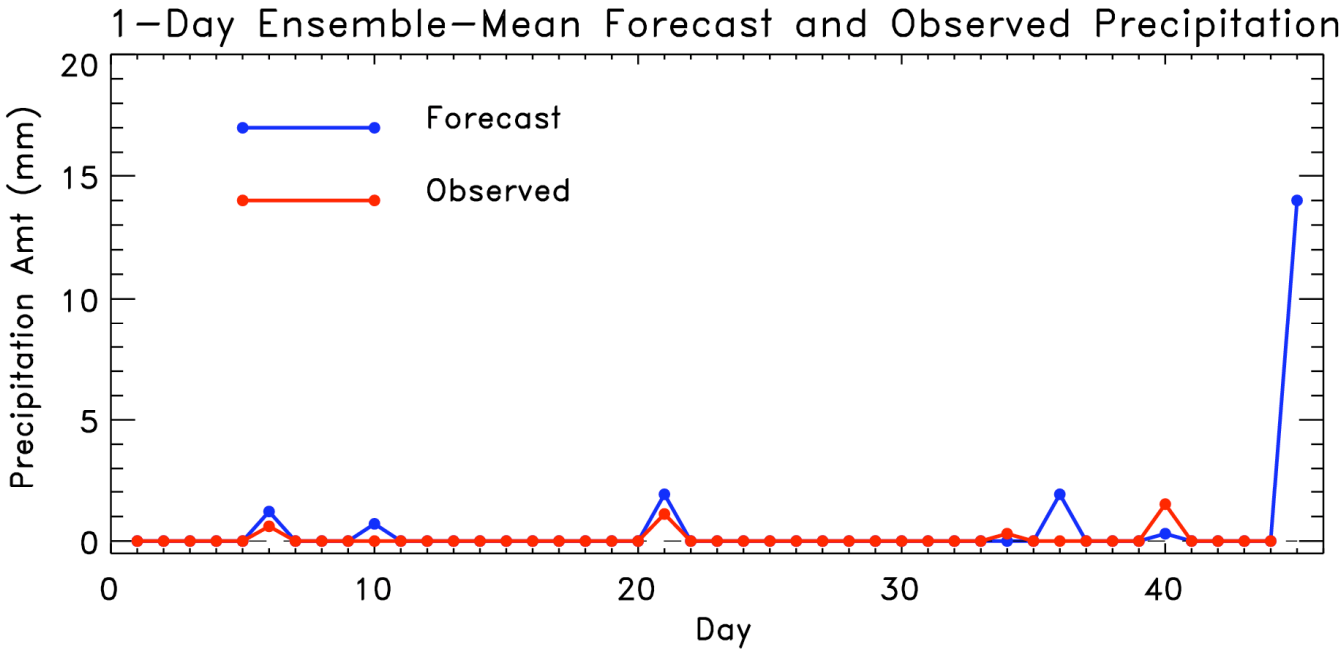
Topics today

- Won't talk about:
  - Approach 1, developing and improving hi-res. models. You're probably well-educated already.
  - Climate forecasting and reforecasting. Marginal skill, not "low-hanging fruit."

- Will talk about
  - Reforecasting for shorter-range forecasts, 1 day to several weeks. Here is where there is a large gain from statistical post-processing.
  - How reforecasting may fit into NWS plans.

Do we really need reforecasts extending over years or decades?

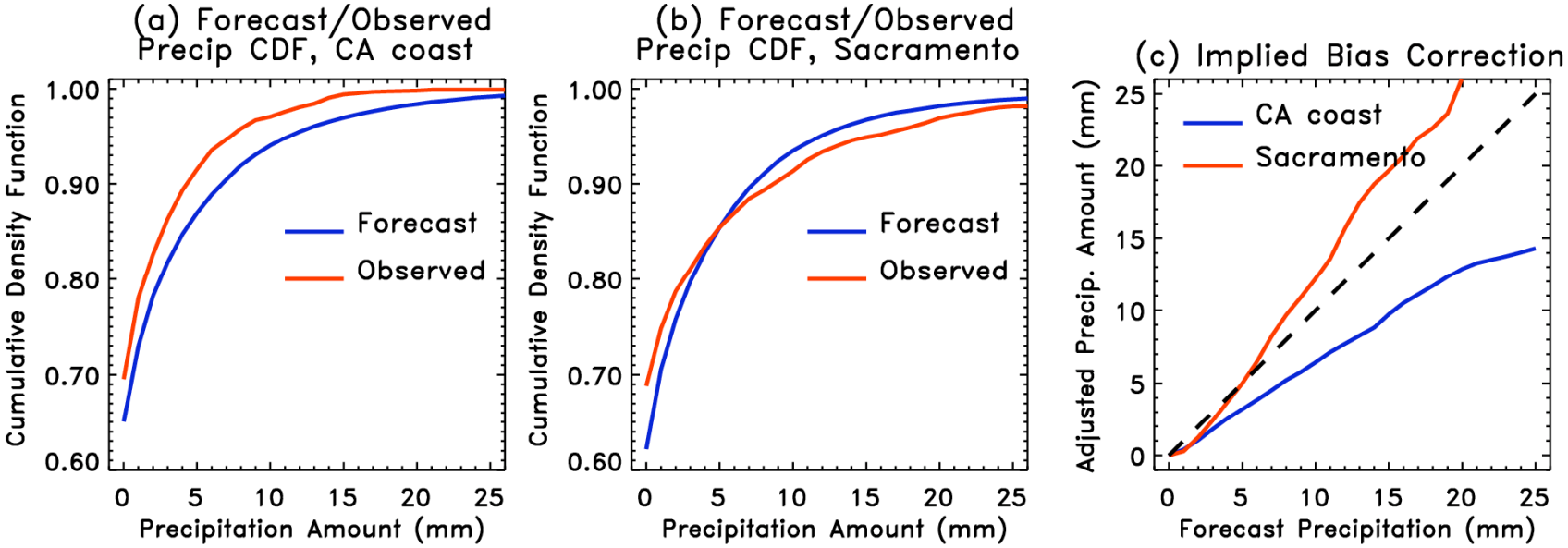
Consider training with a short sample in a climatologically dry region. How could you calibrate this latest forecast?



you'd like enough training data to have some similar events at a similar time of year to this one.

Why not boost sample size by compositing statistics over different locations?

Probably a good idea, if done with care.
However, even nearby grid points may have different forecast errors.



Panels (a) and (b) provide the cumulative density function (CDF) of 1-day forecasts of precipitation for 1 January (CDFs determined from reforecast data and observations in Dec-Jan). Panel (a) is for a location on the CA coast, just north of San Francisco, and panel (b) is for Sacramento, CA. Panel (c) provides the implied function for a bias correction from the forecast amount to a presumed observed amount. Note the very different corrections implied at two nearby locations.

Framing the calibration problem

- Suppose the climate were “stationary” (unchanging from decade to decade).
- Suppose that we had quality weather observations going back many millennia
- Suppose we had an ensemble of reforecasts available back over those many millennia.

Then how might we utilize the reforecasts to improve today’s forecast?

Theory underlying analog calibration technique

$f\left(\mathbf{x}^T \middle| \mathbf{x}^f\right)$ the probability distribution of the true state given today's forecast, where

$$\mathbf{x}^T = \left(x_1^T, \ldots, x_p^T\right)$$

$$\mathbf{x}^f = \left(x_1^f(1), \ldots, x_1^f(m), \ldots, x_n^f(1), \ldots, x_n^f(m)\right) = \left(\mathbf{x}_1^f, \ldots, \mathbf{x}_n^f\right)$$

Here, before simplification, $\mathbf{x}^T$ refers to the true state vector (presumably high-resolution), and $\mathbf{x}^f$ refers to the (lower-resolution) ensemble-forecast state vector.

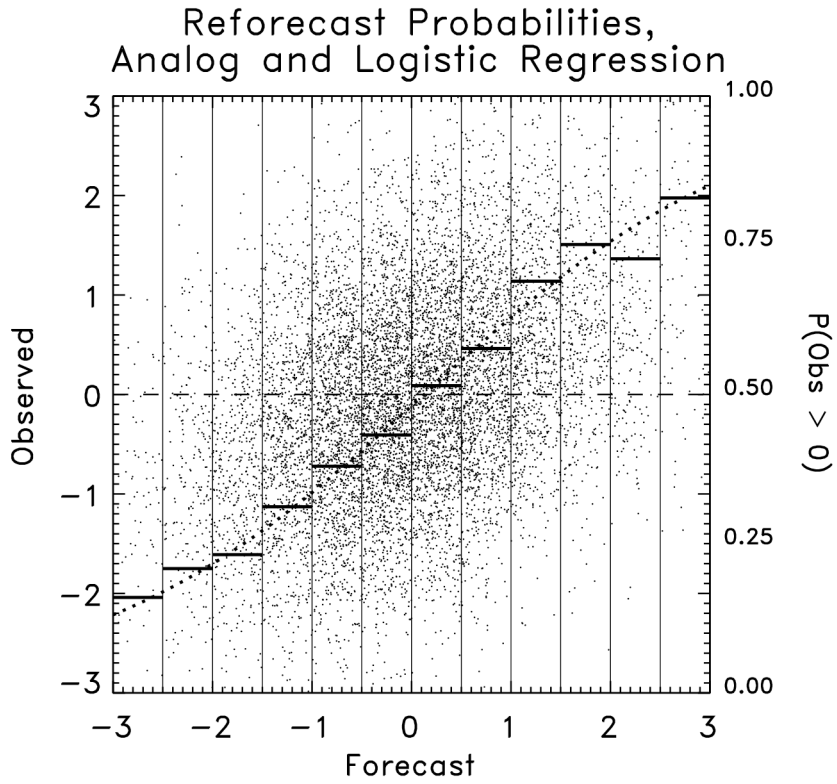
Estimating the conditional distribution with analogs

Suppose we have old forecasts that are ~ identical to today's:

$$\mathbf{x}_{t=t1}^{fOLD}, \mathbf{x}_{t=t2}^{fOLD}, \ldots, \mathbf{x}_{t=tN}^{fOLD} \approx \mathbf{x}^{f}$$

Then $f\left(\mathbf{x}^{T} \middle| \mathbf{x}^{f}\right)$ estimated from

$$\left[\mathbf{x}_{t=t1}^{TOLD}, \mathbf{x}_{t=t2}^{TOLD}, \ldots, \mathbf{x}_{t=tN}^{TOLD}\right]$$



NOAA's reforecast data set

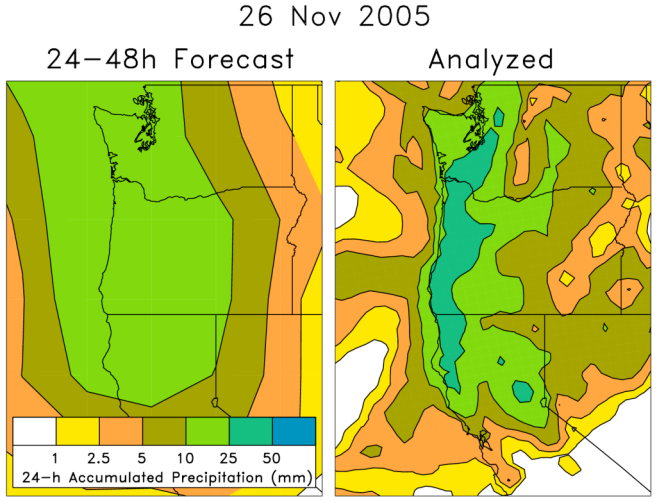
- **Model**: T62L28 NCEP GFS, circa 1998
- **Initial Conditions**: NCEP-NCAR Reanalysis II plus 7 +/- bred modes.
- **Duration**: 15 days runs every day at 00Z from 19781101 to now. (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst/week2*).
- **Data**: Selected fields (winds, hgt, temp on 5 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*).
- **Real-time** probabilistic precipitation forecasts: http://www.cdc.noaa.gov/reforecast/narr

A simplified construct for calibration using reforecasts

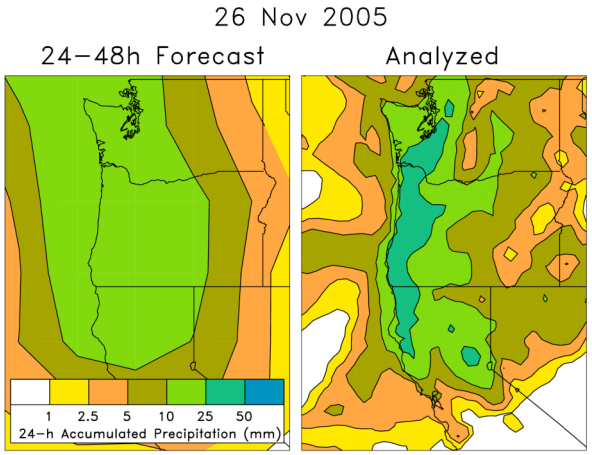
(1) Most of the information in our GFS reforecast ensemble contained in the ensemble mean, so…

$$f\left(\mathbf{x}^T \left| \overline{\mathbf{x}}^f \right.\right) \leftarrow f\left(\mathbf{x}^T \left| \mathbf{x}^f \right.\right)$$

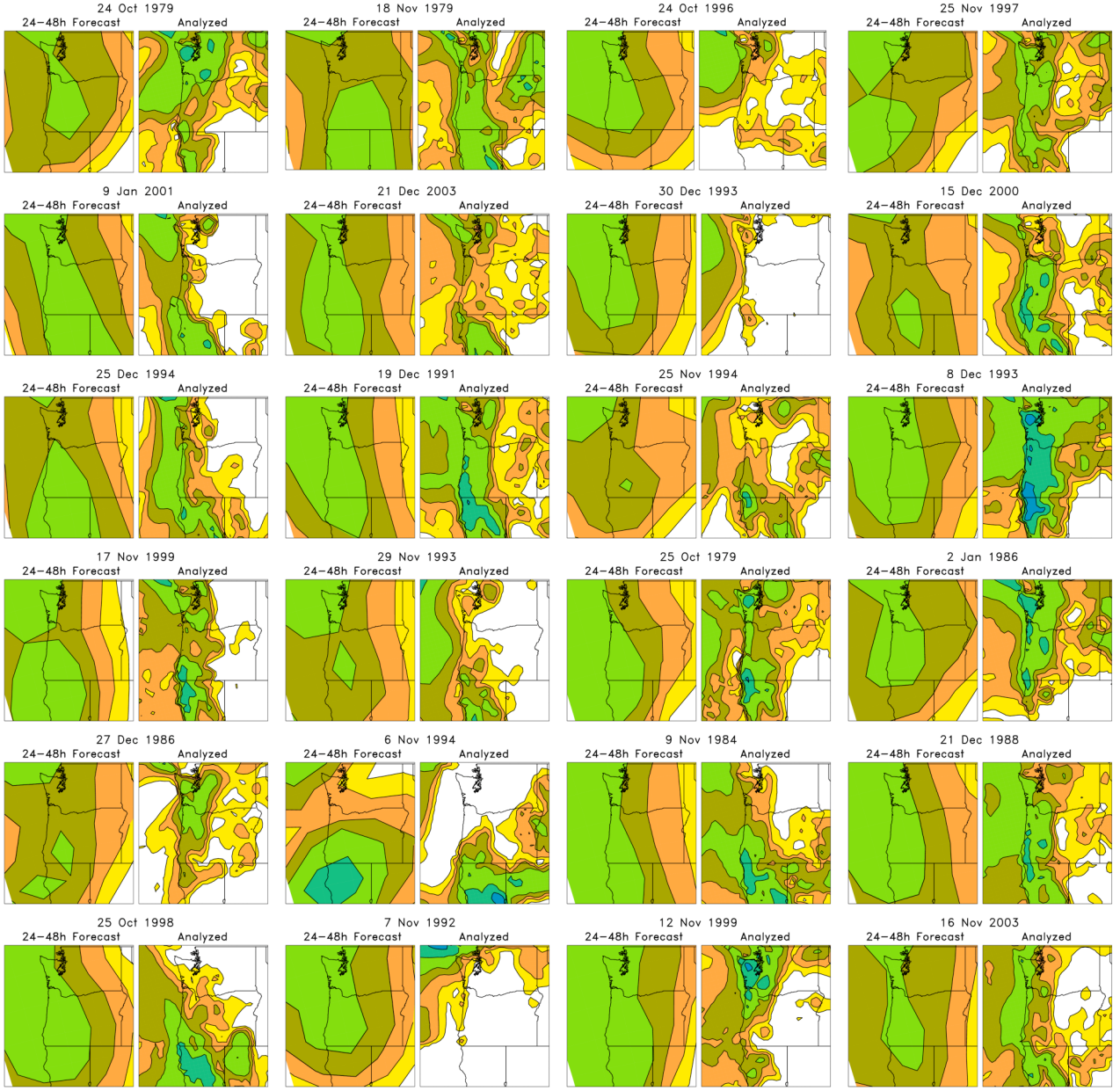
(2) Let's find the distribution of the observed conditional upon the part of the forecast state that's nearby; i.e., don't worry about Washington, DC when making a forecast for Washington, State.



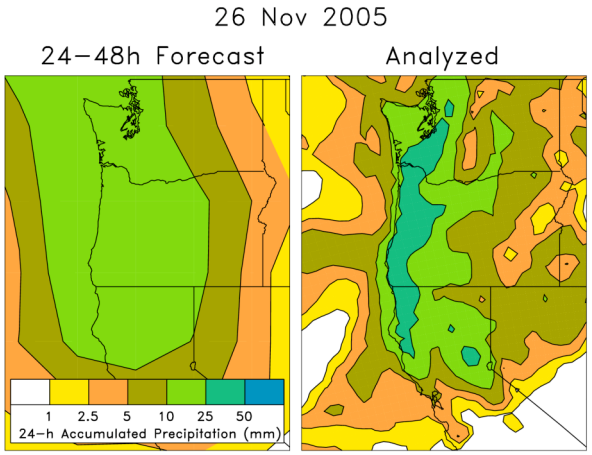
Producing a distribution of observed given forecast using analogs



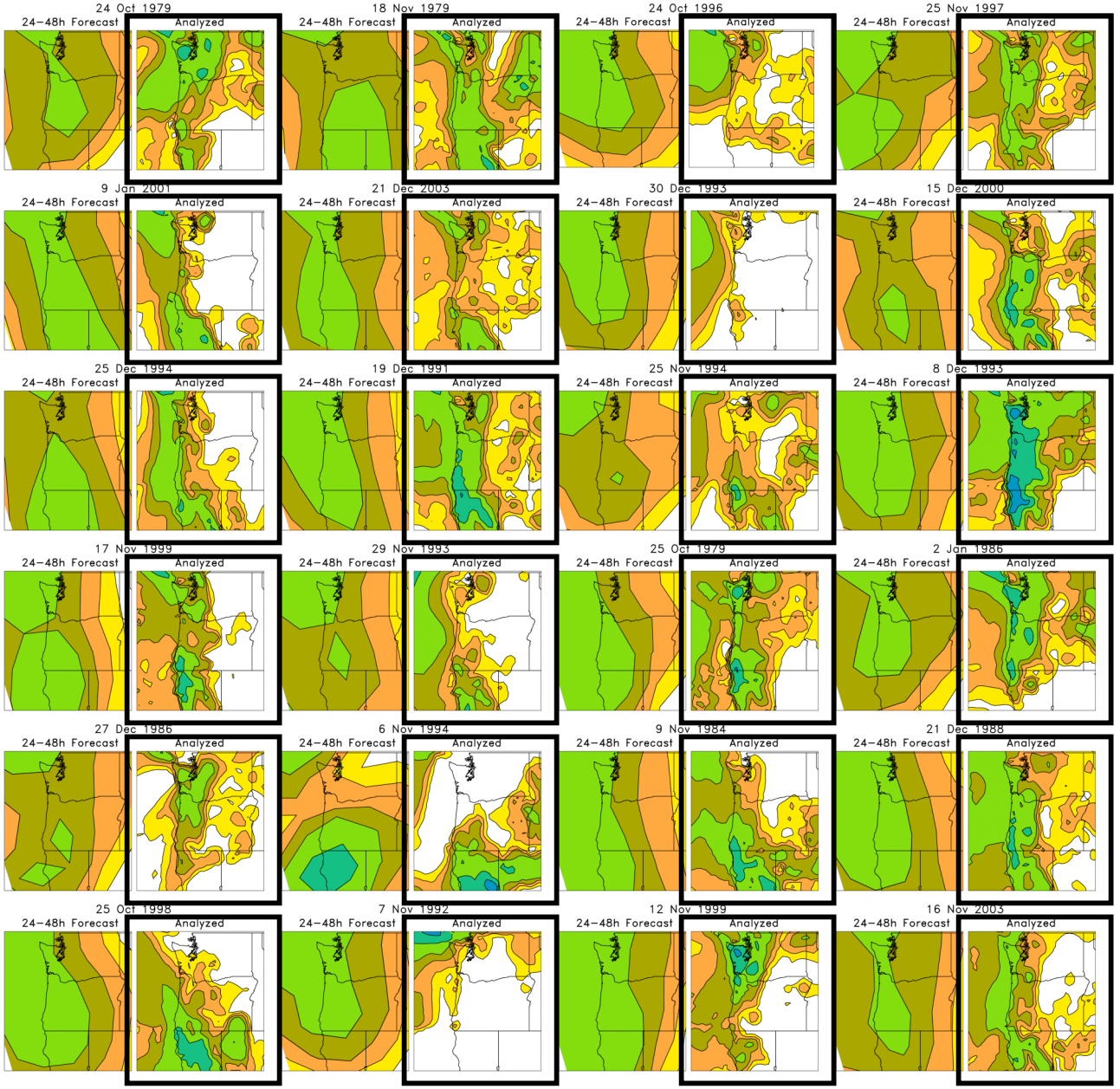
On the left are old forecasts similar to today's ensemble-mean forecast. For making probabilistic forecasts, form an ensemble from the accompanying analyzed weather on the right-hand side.



Producing a distribution of observed given forecast using analogs



On the left are old forecasts similar to today's ensemble-mean forecast. For making probabilistic forecasts, form an ensemble from the accompanying analyzed weather on the right-hand side.

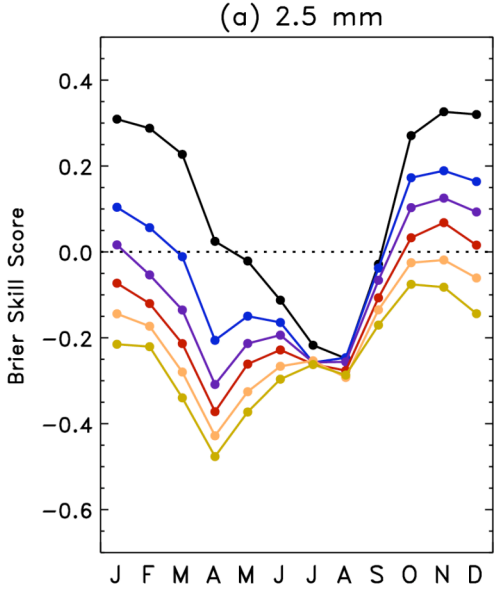


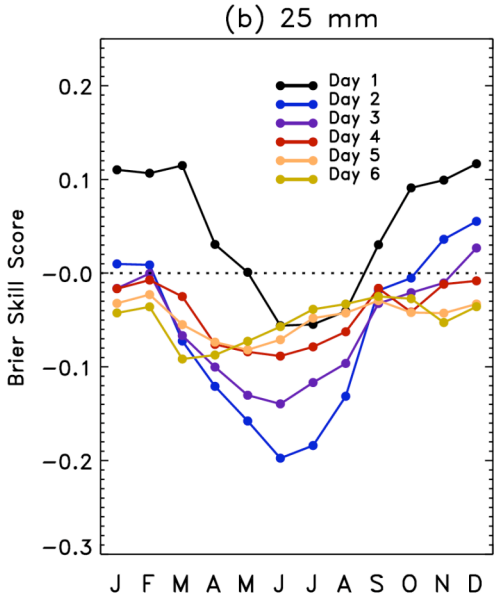
Asymptotic behavior of analog technique

- Q: What happens as corr(F,O) $\rightarrow$ 0 ? A: Ensemble of observed analogs becomes random draw from climatology. 

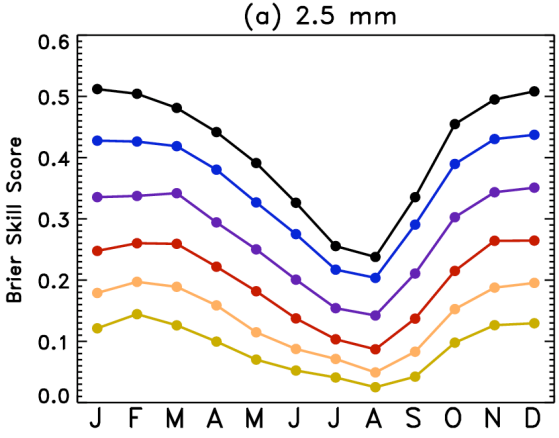
- Q: What happens as corr(F,O) $\rightarrow$ 1 ? A: Ensemble of observed analogs looks just like today's forecast. Sharp, skillful forecasts. 

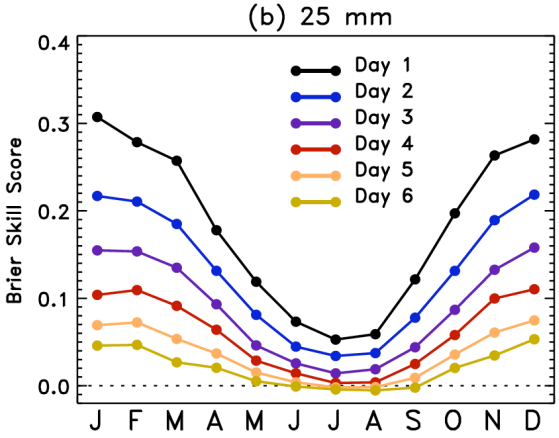
Ensemble Relative Frequency





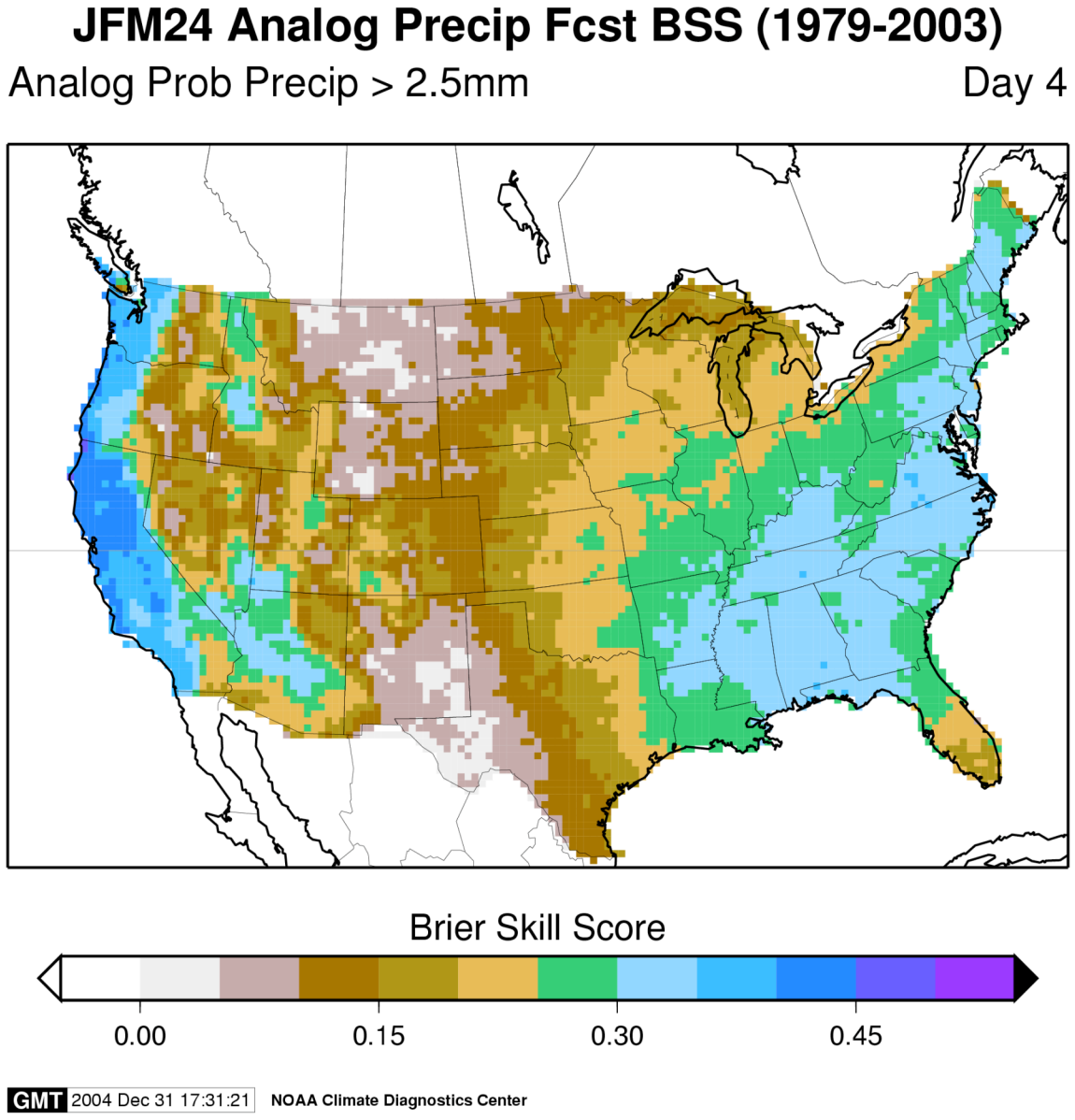
Basic Analog Technique



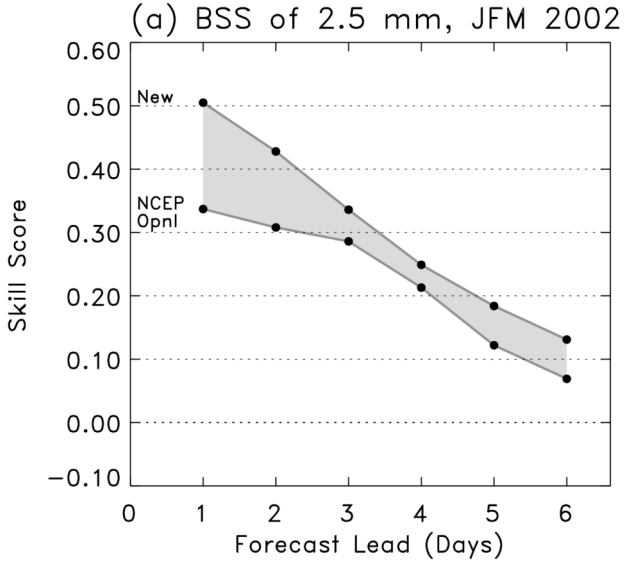


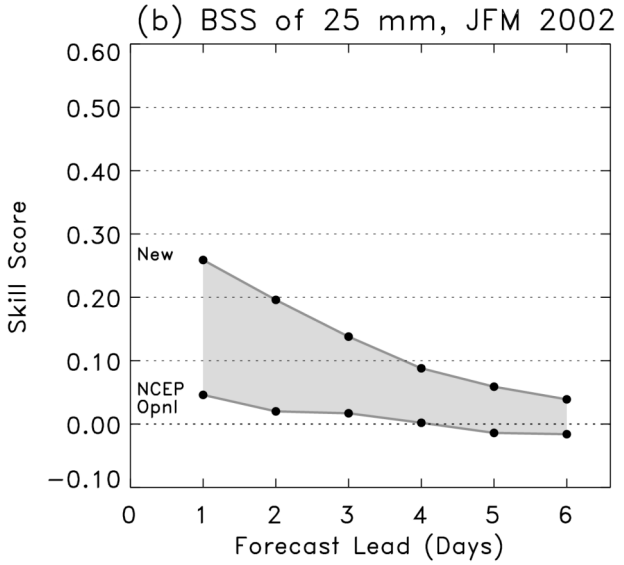
Verified over 25 years of forecasts; skill scores use conventional method of calculation which may overestimate skill
(Hamill and Juras 2006, QJRMS, Oct).

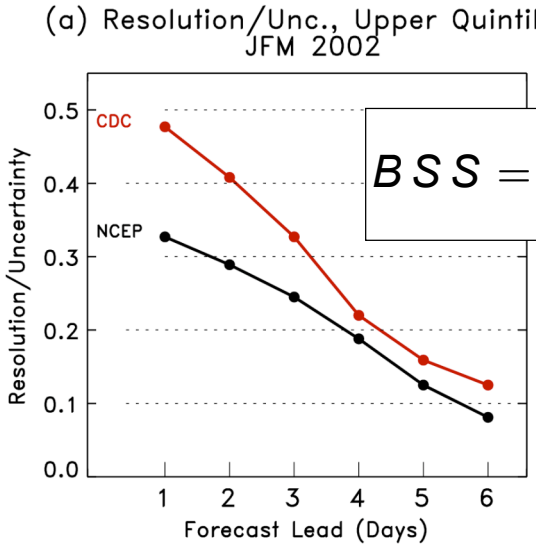
Skill as function of location

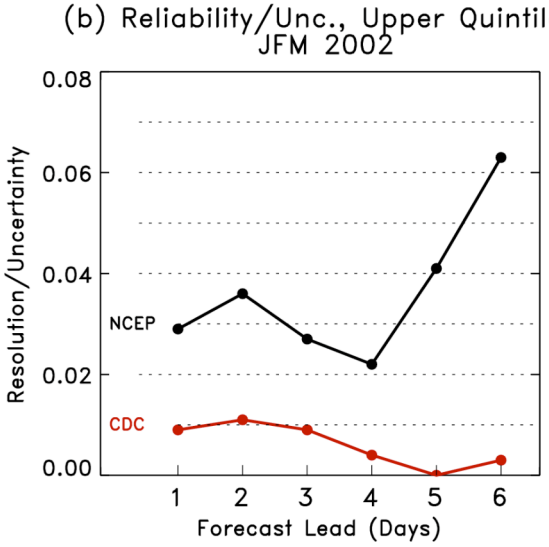


Comparison against NCEP medium-range T126 ensemble, ca. 2002



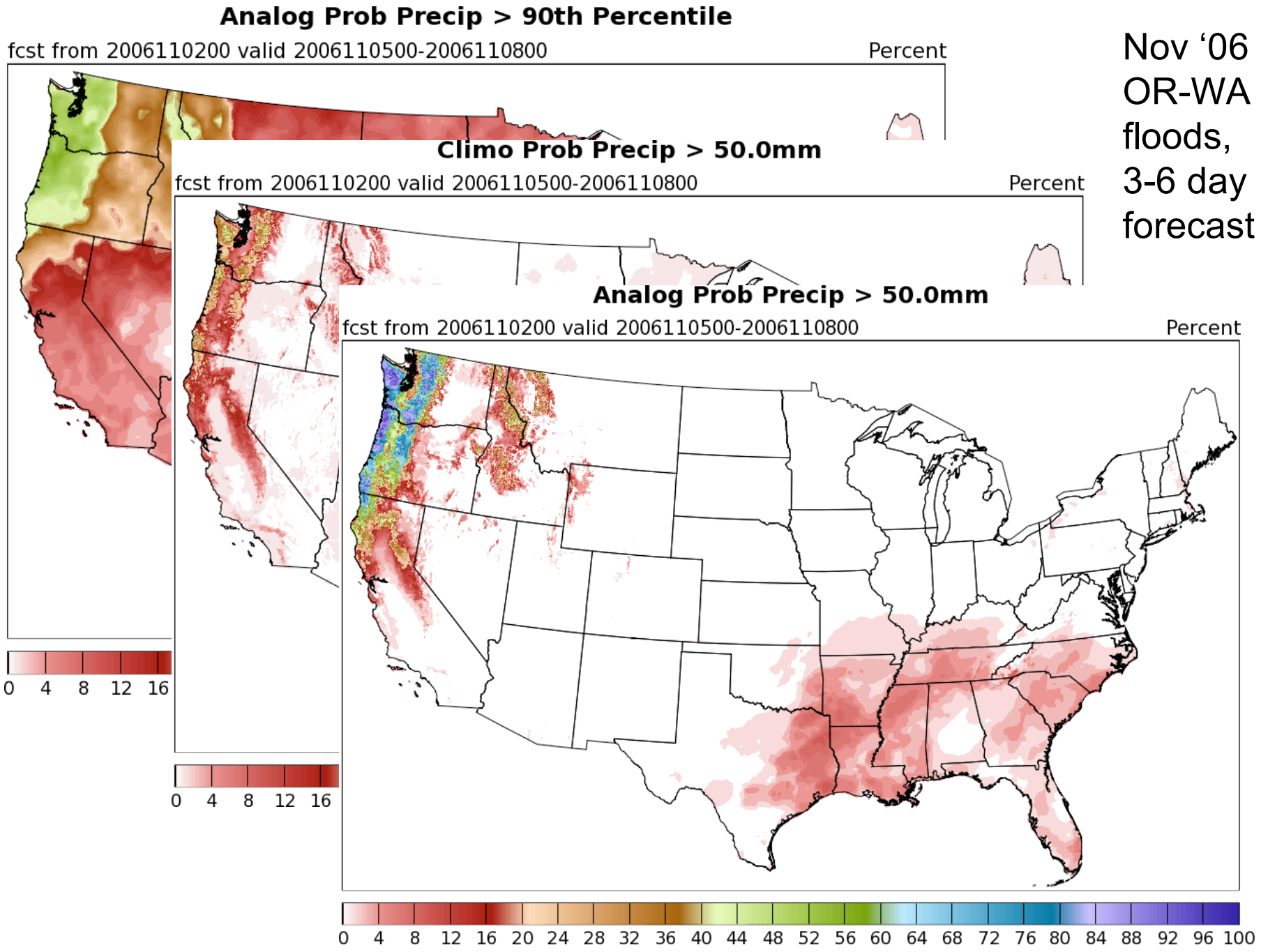






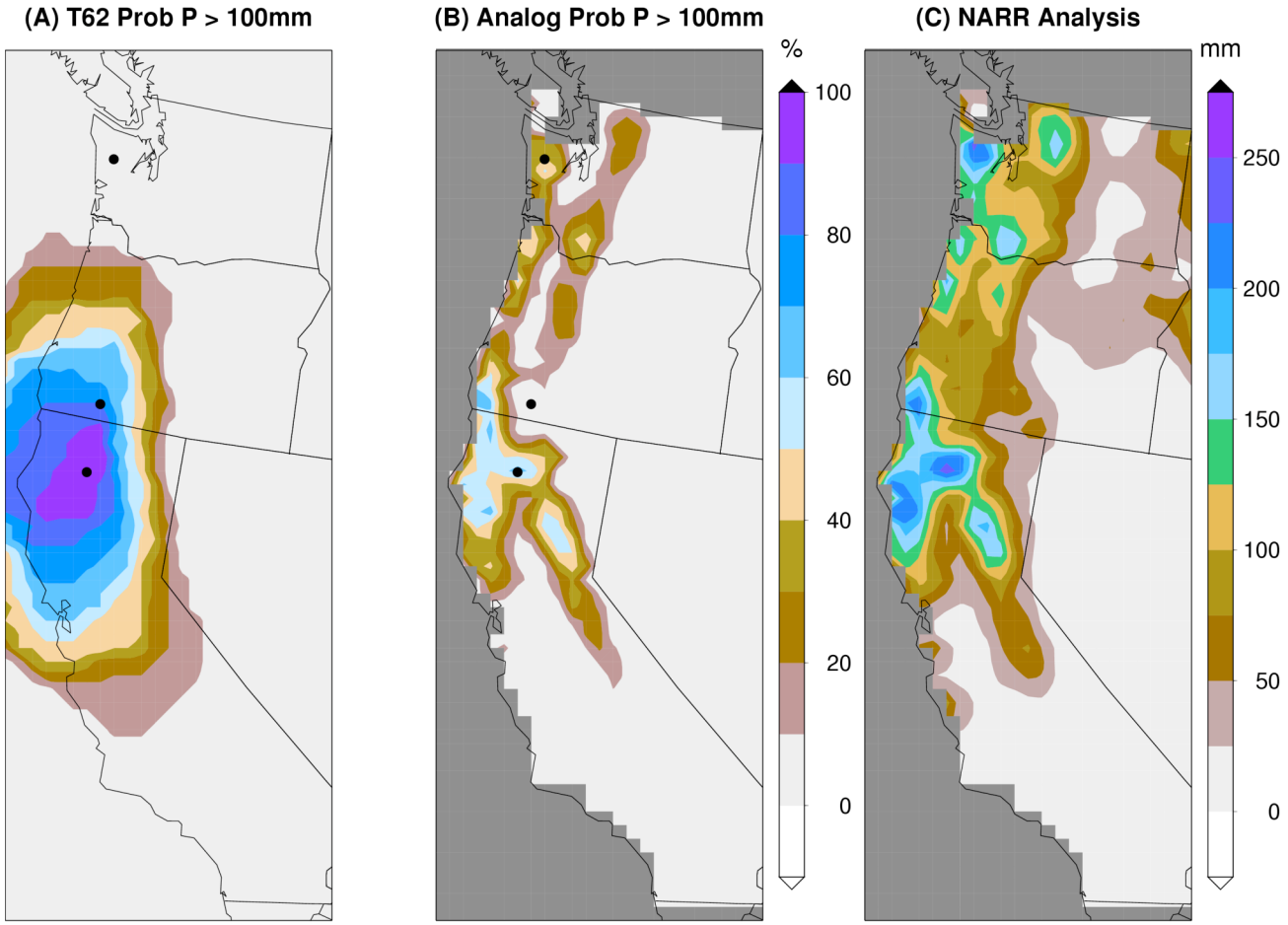
$$BSS = \frac{resolution - reliability}{uncertainty}$$

the improvement is a little bit of increased reliability, a lot of increased resolution.

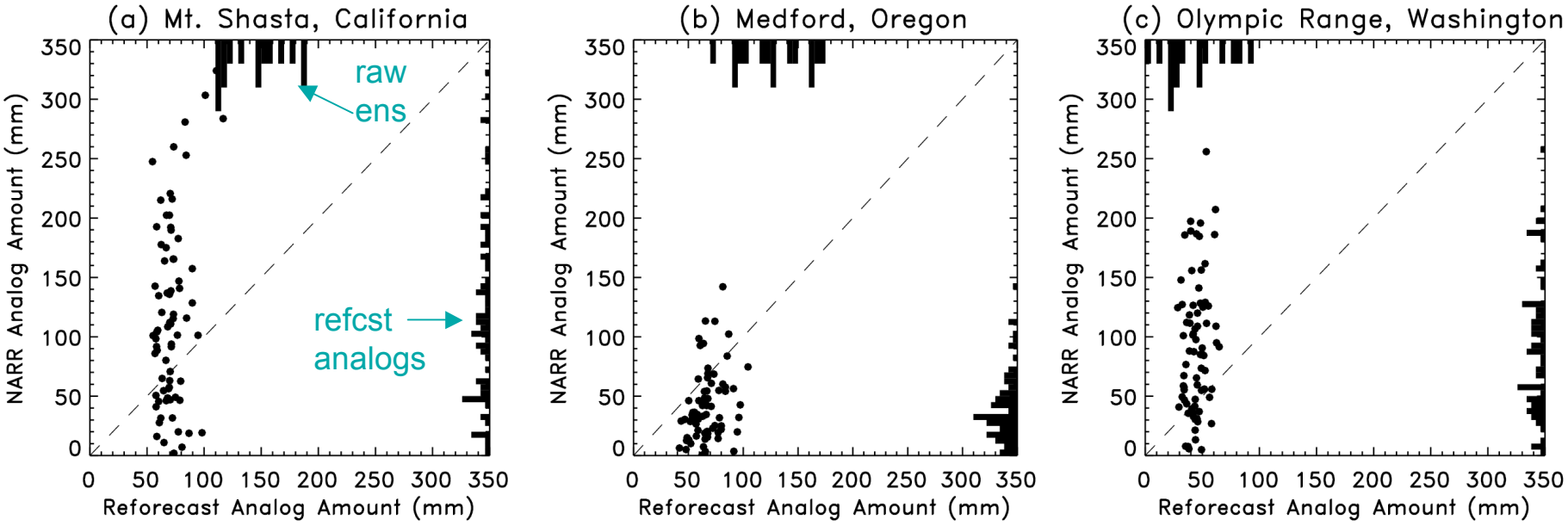


Nov '06 OR-WA floods, 3-6 day forecast

Analog example:
Day 4-6 heavy precipitation in California, 0000 UTC 29 December 1996 - 0000 UTC 1 January 1997



Bias, spread, and downscaling corrections in analog technique

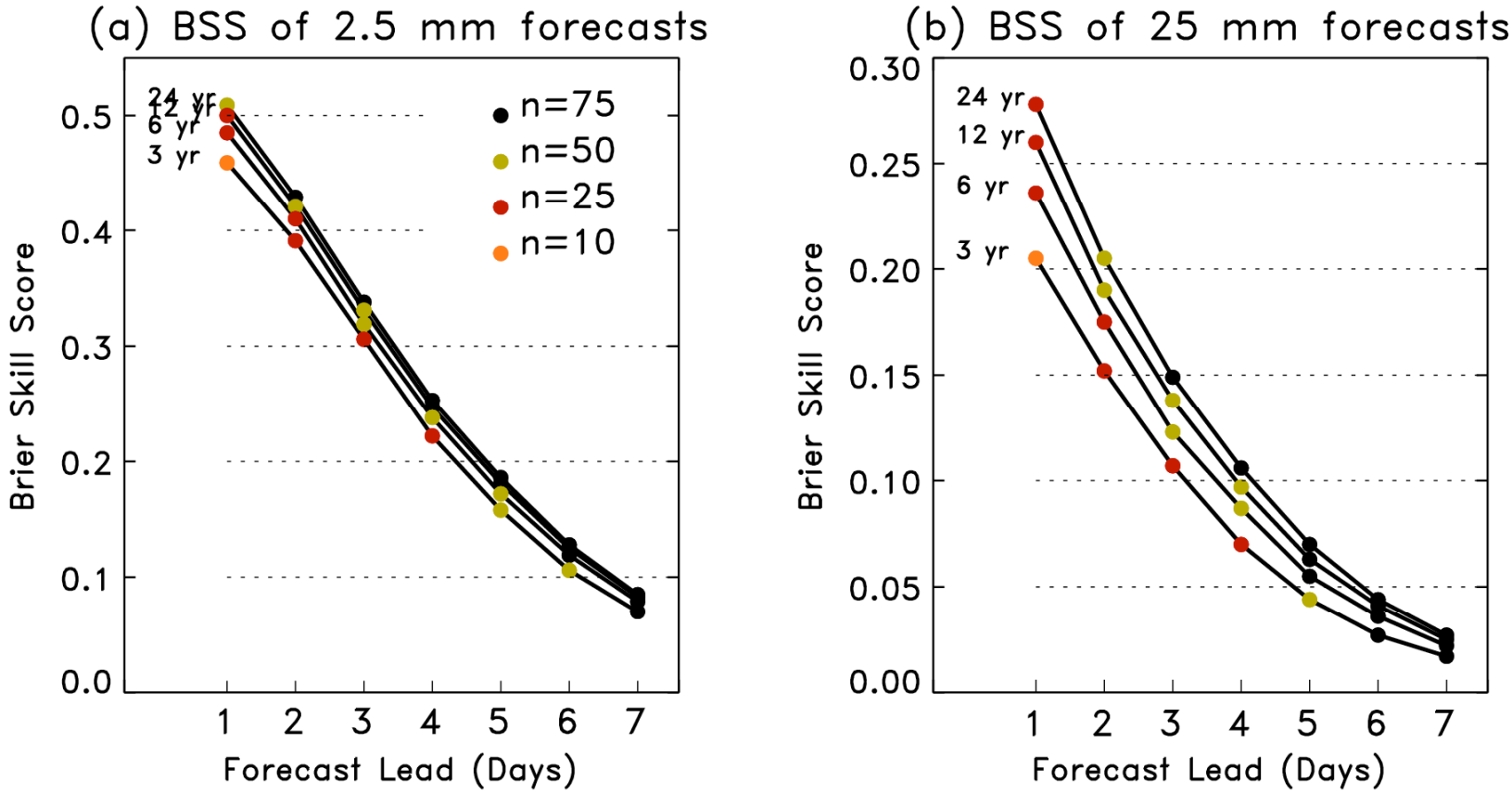


Can't find any other reforecast analogs with precip as heavy. But introduce large scatter by taking associated observed analogs.

Again, few close reforecast analogs. But observed data recognizes overforecast bias.

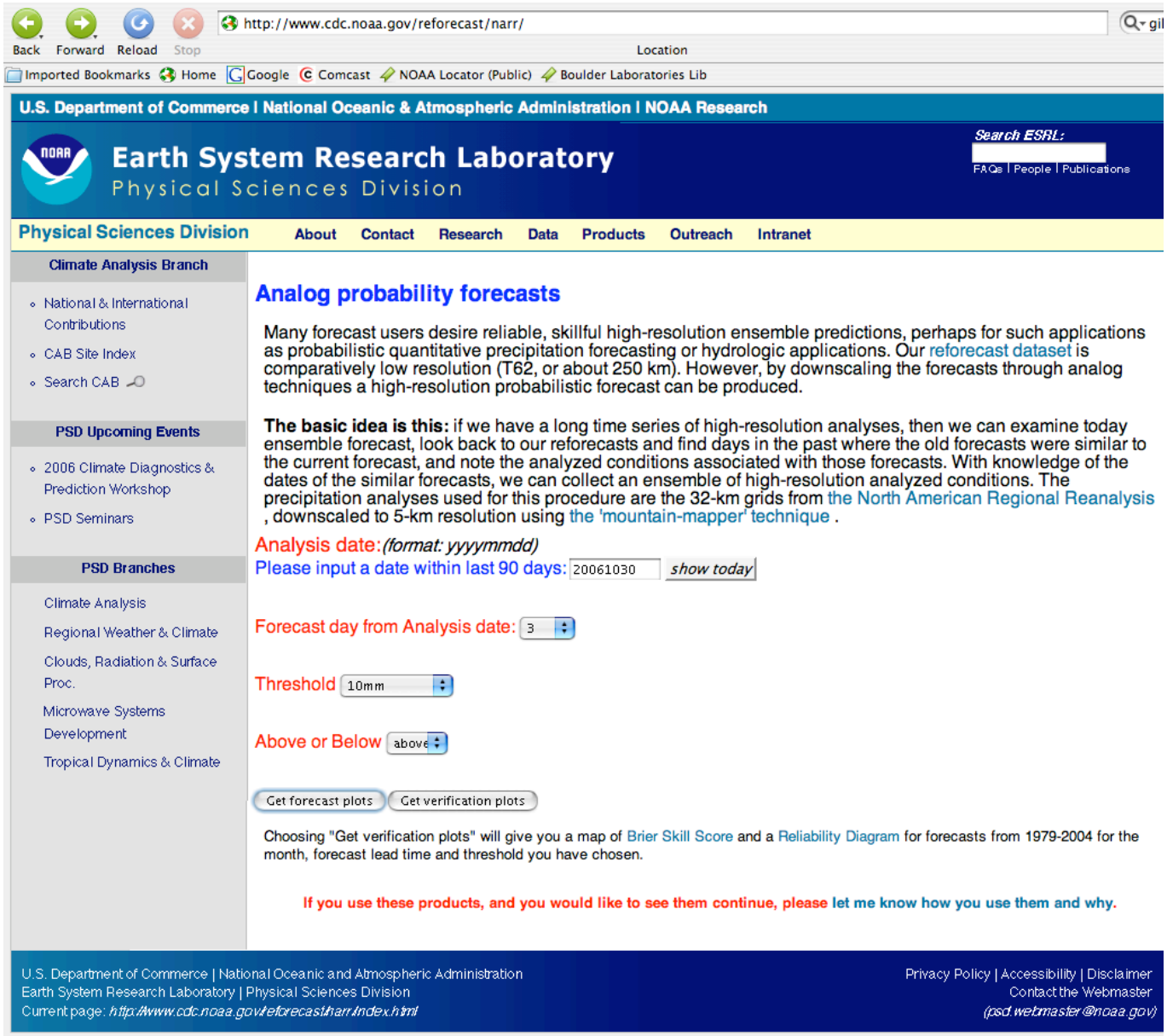
Here there are close reforecast analogs. Observed data introduces spread, increases amount.

Effect of training sample size

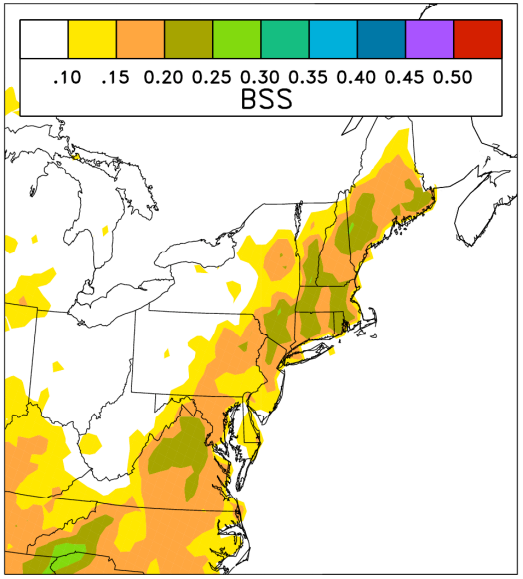


colors of dots indicate which size analog ensemble provided the largest amount of skill.

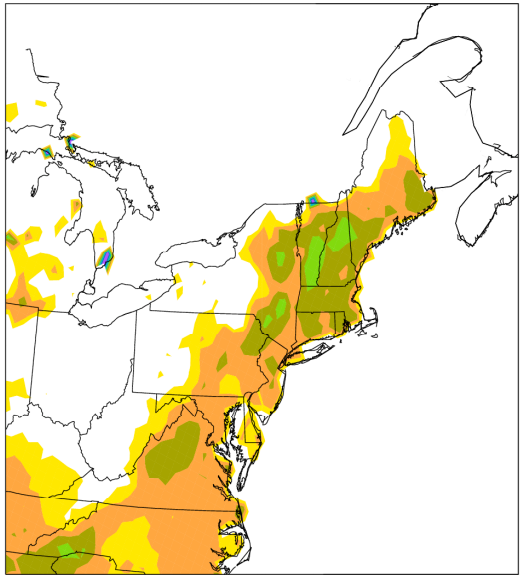
Real-time products



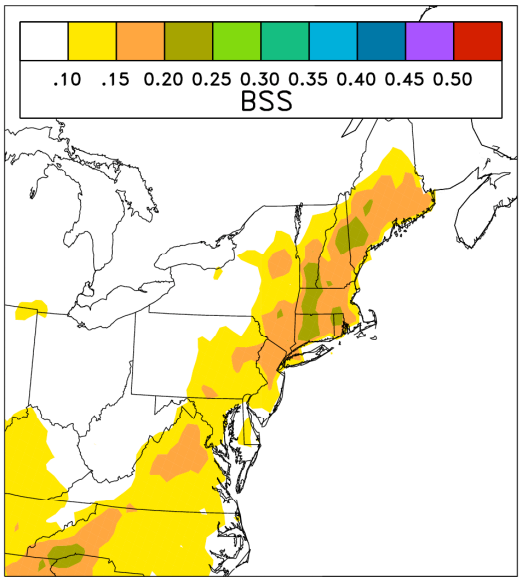
(a) Smoothed Rank Analog JFM 25mm
1−Day Forecast 10 mbrs



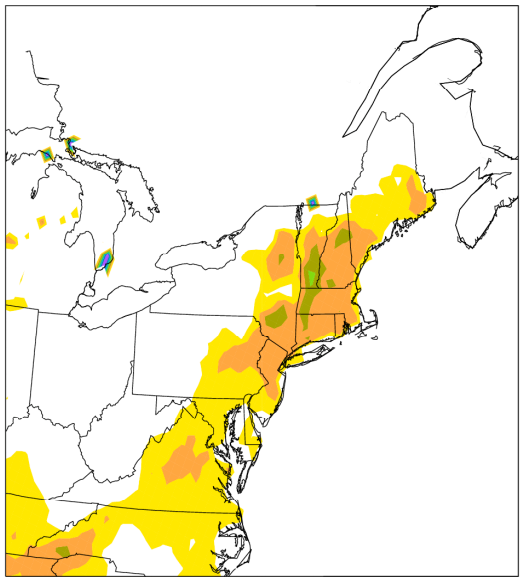
(b) Logistic Regression JFM 25mm
1−Day Forecast



(a) Smoothed Rank Analog JFM 25mm
2−Day Forecast 25 mbrs



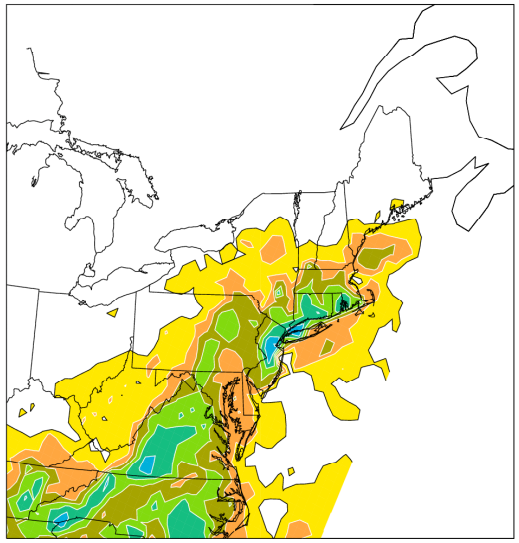
(b) Logistic Regression JFM 25mm
2−Day Forecast



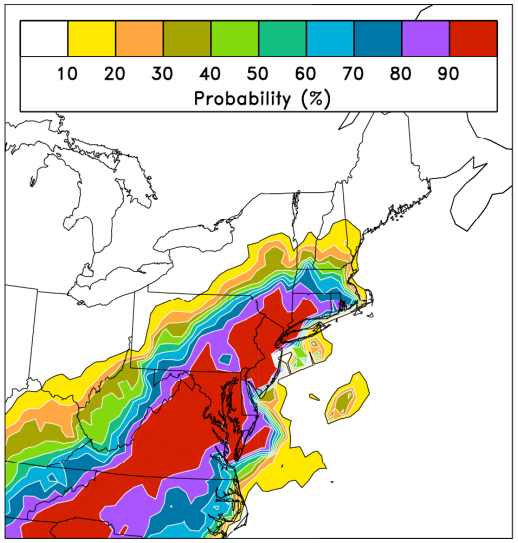
We compare here the smoothed rank analog approach to the logistic regression approach for wintertime (JFM) data over the northeast USA. The focus is specifically on the 25-mm threshold, i.e., the quality of forecasting heavy-precipitation events.

First, notice that maps of the overall precipitation forecast skill are relatively similar, here for day-1 and day-2 forecasts. The logistic regression appears to be slightly more skillful over New England on day 1.

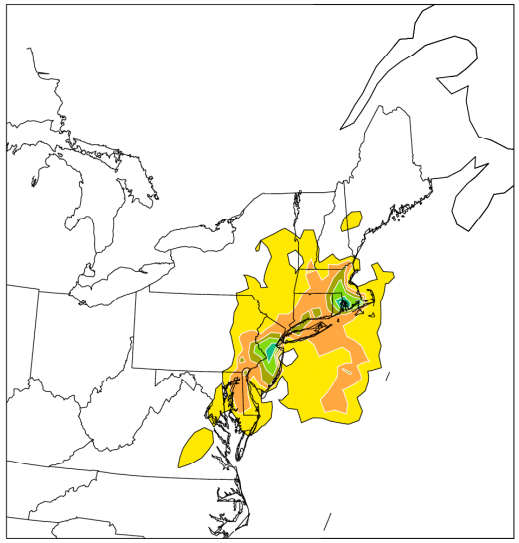
(a) Smoothed Rank Analog
Pr(Precip > 25 mm), 1–day fcst,
0000 UTC 1993 03 13



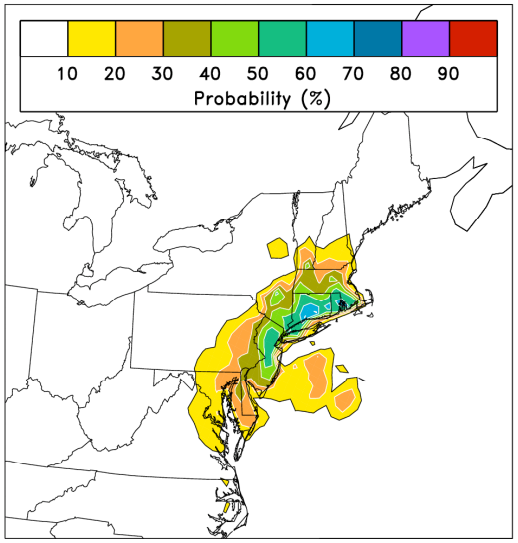
(b) Logistic Regression
Pr(Precip > 25 mm), 1–day fcst,
0000 UTC 1993 03 13



(a) Smoothed Rank Analog
Pr(Precip > 25 mm), 1–day fcst,
0000 UTC 1995 02 04



(b) Logistic Regression
Pr(Precip > 25 mm), 1–day fcst,
0000 UTC 1995 02 04



Next, consider some individual storms and their forecasts. For record-setting events like 1993's "Storm of the Century", logistic regression "extrapolates the regression" and produces much higher probabilities. 10-member rank analog techniques produced much lower probabilities, since most if not all reforecast analogs that were selected inevitably had lower forecast (and presumably analyzed) precipitation amounts.

Possible paths forward

(1) Use CPU resources to rapidly develop higher-resolution ensembles with improved physical veracity. Improve methods of generating initial conditions, generate ways of dealing with uncertainty of the forecast model itself. [What we've been doing]

(2) Use those CPU cycles to run a fixed model and data assimilation system, albeit an older, low-resolution one. Run real-time, plus many past forecast cases. Diagnose the forecast error characteristics and generate statistically adjusted forecasts ("*reforecasting*")

(3) Compromise between the two.

Can we do both hi-res model development and reforecasting, or a compromise?

- **Alternative 1**: Continue development of high-res. models. Do reforecasting with inexpensive, low-res model, so operations are impacted minimally.
  - Suppose operational T300, 60-layer, 50-member ensemble forecast system.
  - Reforecast T150, 40 layer, 5-member ensemble :
    - Operational cost: 120x less
    - 120 days of reforecasts for one day of operational forecast, so a 20-year reforecast for the cost of 60 days of operational model forecasts.
    - If new reforecast model implemented once, say, every 4 years, minimal impact to operations integrated over time.

- **Alternative 2**: Continue development of high-res. models. Do reforecasting offline, on non-operational computer system.
  - ~ $700K would buy a computer system that could do a T170L42, 5-member reforecast out to 10 days in ~ 1 year wall time.

What's next for reforecasting?

- Growing interest from NWP centers worldwide
  - ECMWF exploring once-weekly ensemble reforecasts (with my participation)
  - Canadians planning 5-year ensemble reforecasts
  - NCEP envisioning 1-member, real-time reforecast for bias correction.
- Possibility that NOAA/ESRL may get money to do a more complete, 2nd-generation reforecast data set for NOAA.
- Being discussed in NOAA's strategic planning.

Research questions

- Given computational expense of reforecasts, how do we best:
  - Limit the number of reforecasts that we need to do (fewer ensemble members, not every day, etc.)
  - Can we do things like composite the data across different locations to boost sample size?
  - Do we need a new reanalysis every time we do a new reforecast?
  - Do the benefits of reforecasts propagate down to users like hydrological forecasters?
- We welcome your thoughts and requirements for next-generation reforecast system.

References

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

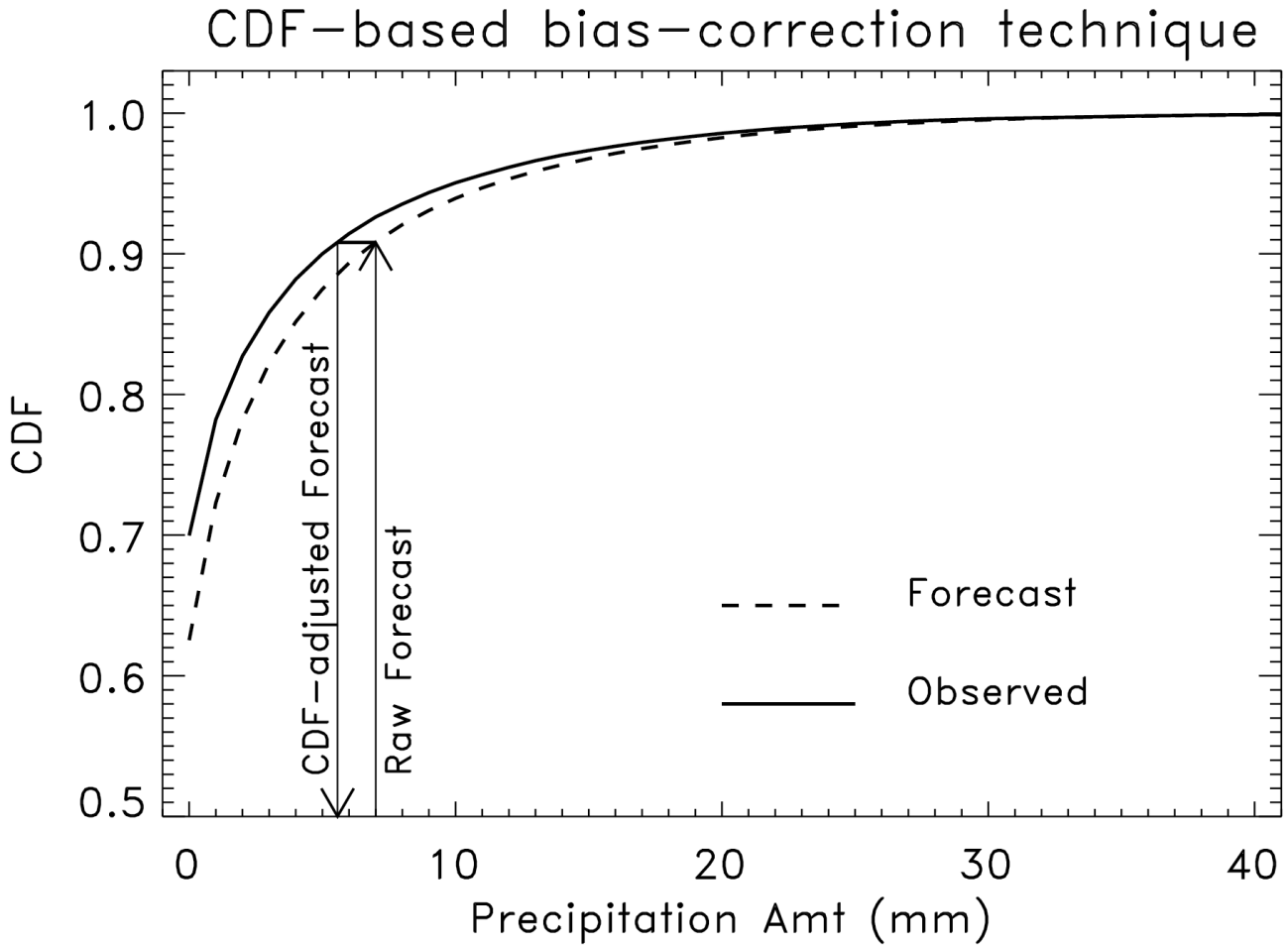
Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., in press.
http://www.cdc.noaa.gov/people/tom.hamill/reforecast_analog_v2.pdf

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

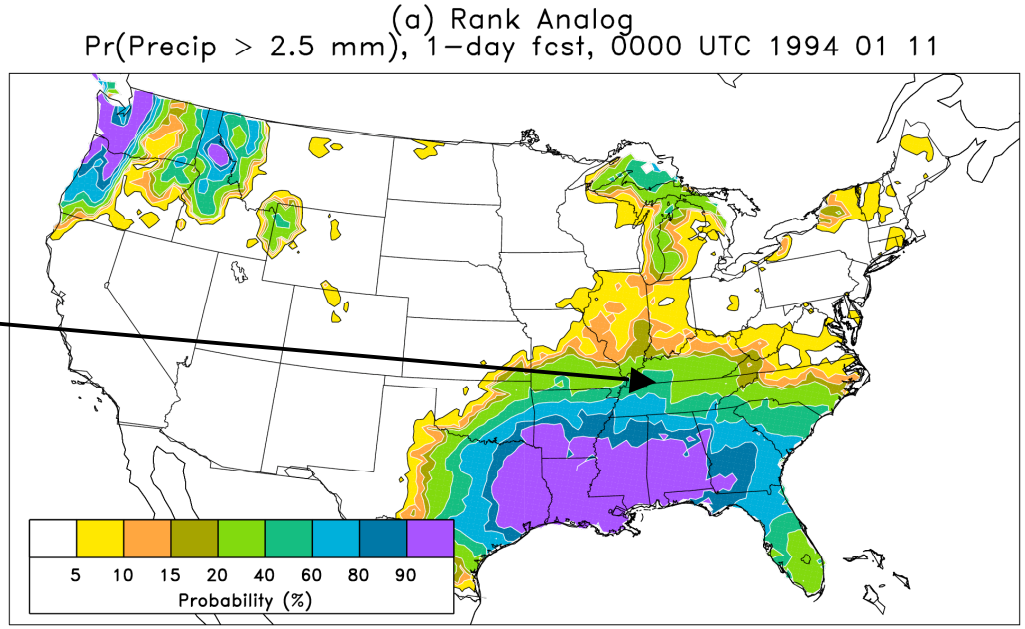
Hamill, T. M. and J. S. Whitaker, 2006: White Paper. "Producing high-skill probabilistic forecasts using reforecasts: implementing the National Research Council vision." Available at
http://www.cdc.noaa.gov/people/tom.hamill/whitepaper_reforecast.pdf .

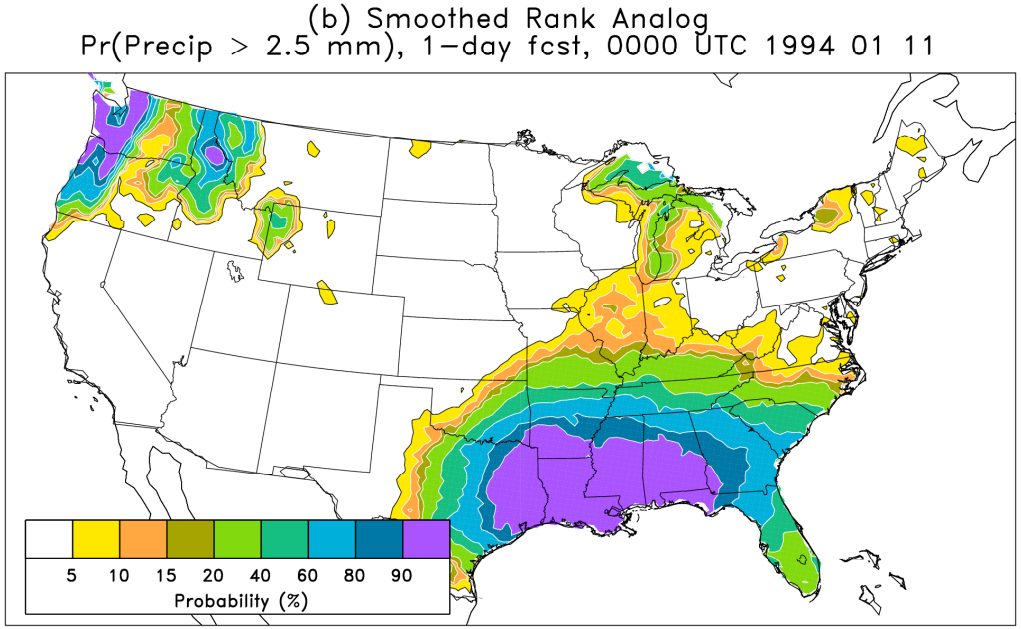
CDF-based bias correction



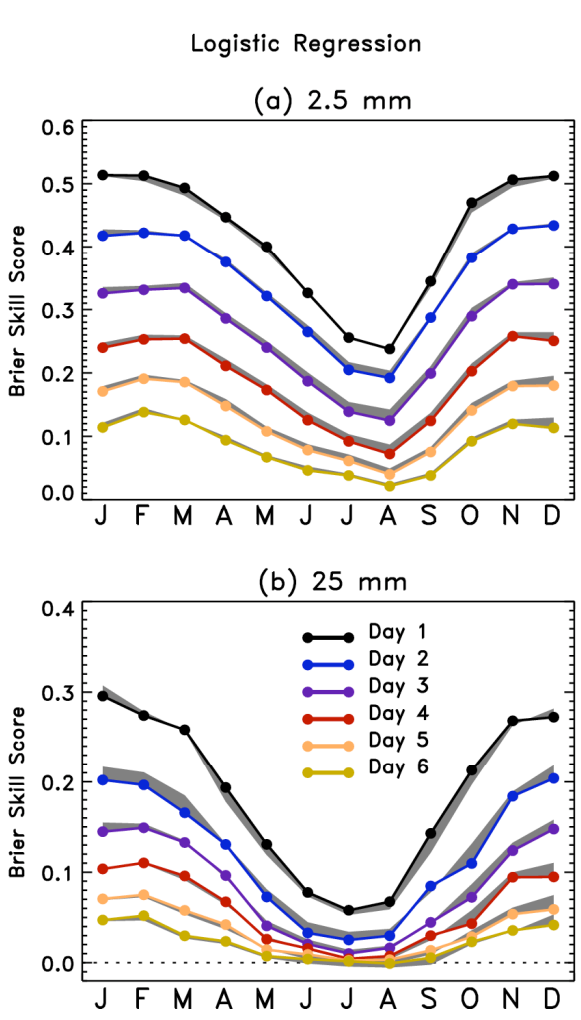
Problem: Different sets of analogs for adjacent regions may sometimes lead to discontinuities in probabilities.

However, it is possible to smooth.

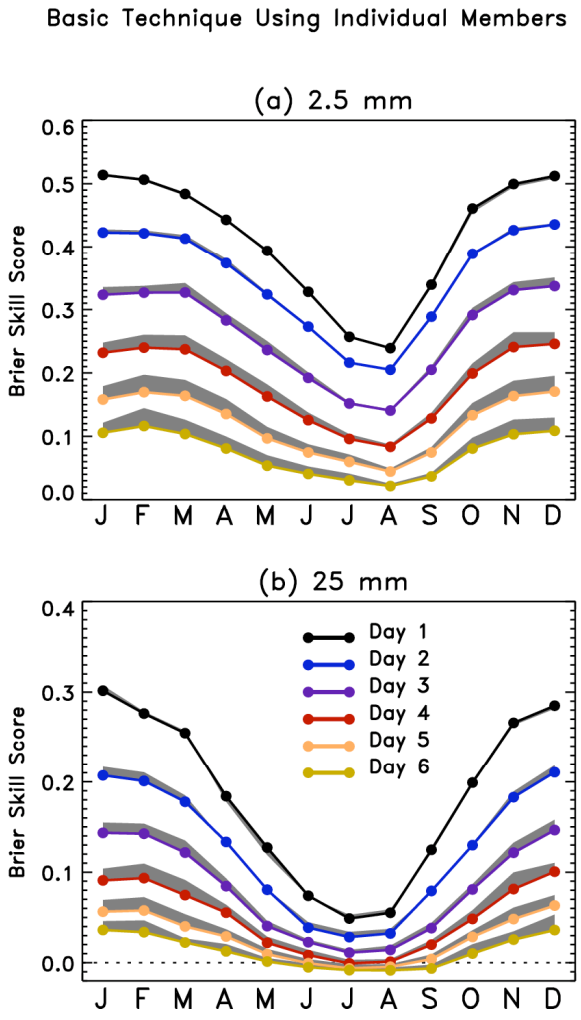




Some other tests

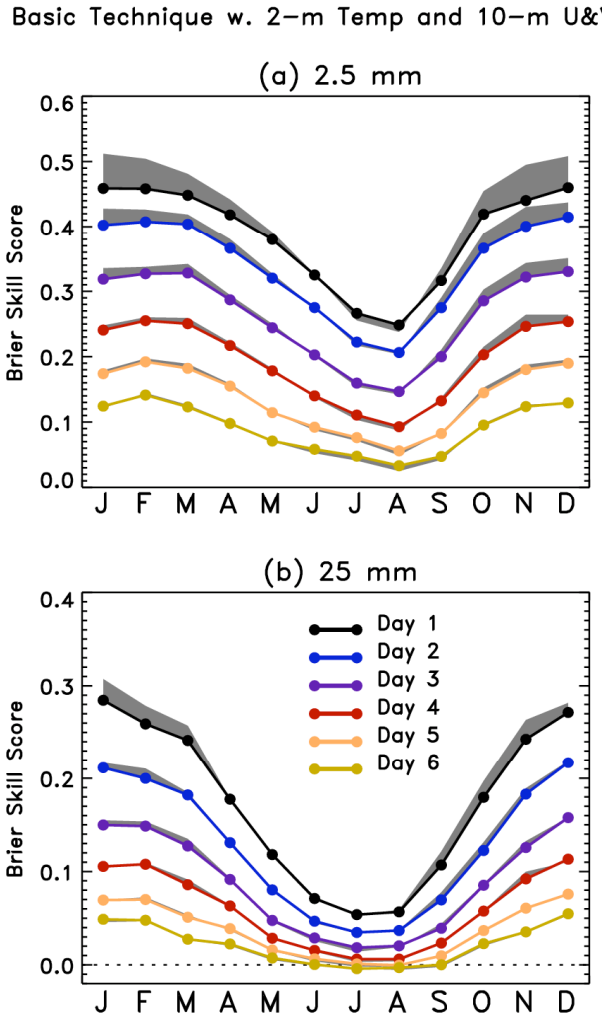


Mixed results when probabilistic forecasts generated using logistic regression approach.

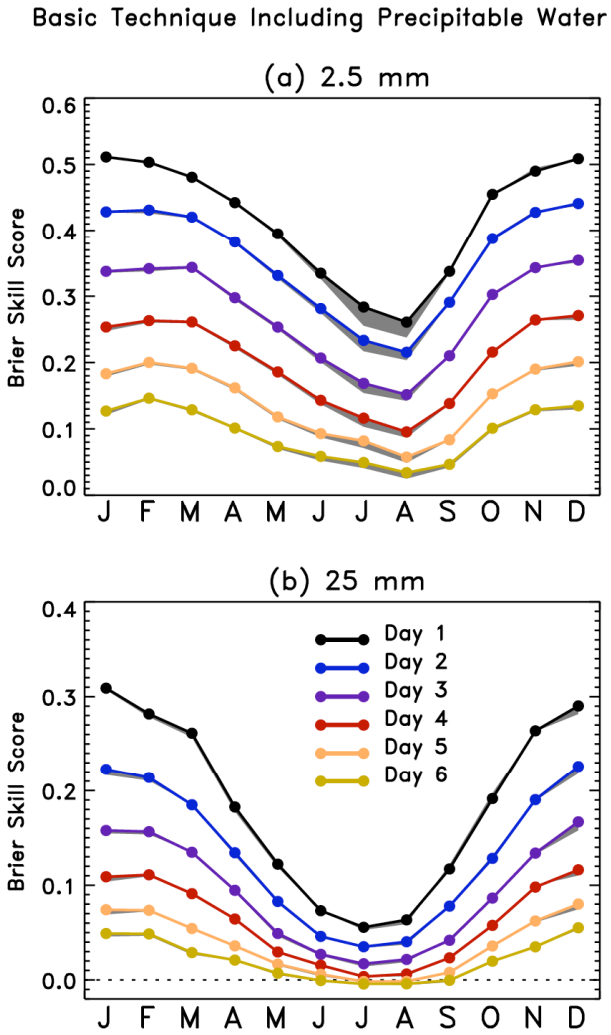


Worse skill when attempting to fit individual members.

Other tests, continued



Worse skill when basing analogs on precip/U/V/T fit.



Some skill improvement in the summer when adding precipitable water as predictor.