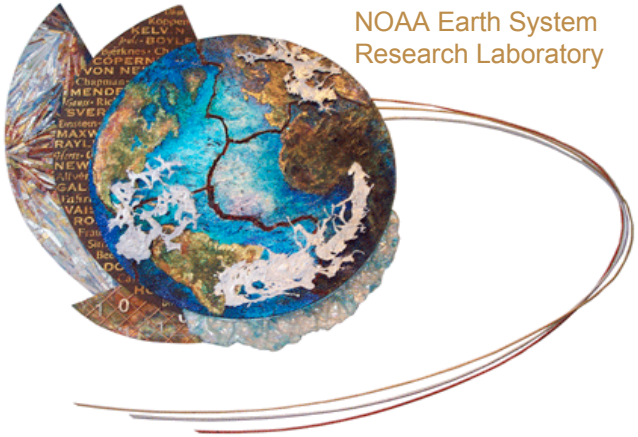


Krzysztofowicz and Evans' "Bayesian Processing of Forecasts" - evaluation with GFS reforecasts

Tom Hamill

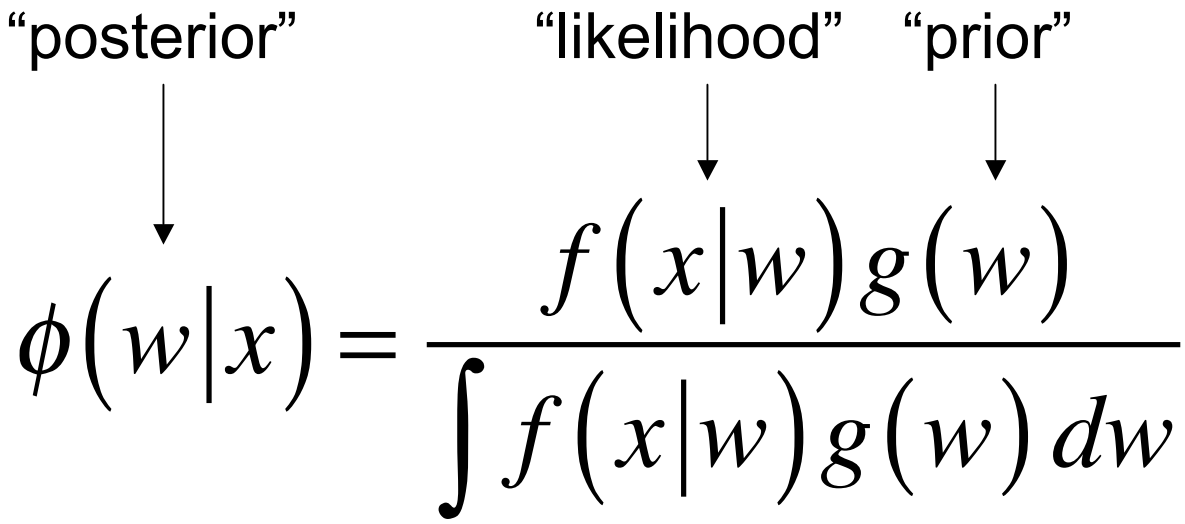
NOAA Earth System Research Lab

tom.hamill@noaa.gov

General problem

- Ensemble forecast skill degraded by deficiencies in initialization method, model error; generally can't estimate pdf directly from ensemble very well.
- Calibration: want pdf for observed | forecast.
- General strategy: Use past (f,o) pairs to train how to adjust current forecast.

Bayes Rule



x is forecast, w is observed.

Would like to leverage **large information content in $g(w)$ that commonly will be available**, even if few (w,x) pairs available for training.

Probabilistic Forecasts from the National Digital Forecast Database

ROMAN KRZYSZTOFOWICZ AND W. BRITT EVANS
University of Virginia, Charlottesville, Virginia

(Manuscript received 22 March 2007, in final form 15 June 2007)

ABSTRACT

The Bayesian processor of forecast (BPF) is developed for a continuous predictand. Its purpose is to process a deterministic forecast (a point estimate of the predictand) into a probabilistic forecast (a distribution function, a density function, and a quantile function). The quantification of uncertainty is accomplished via Bayes theorem by extracting and fusing two kinds of information from two different sources: (i) a long sample of the predictand from the National Climatic Data Center, and (ii) a short sample of the official National Weather Service forecast from the National Digital Forecast Database. The official forecast is deterministic and hence deficient: it contains no information about uncertainty. The BPF remedies this deficiency by outputting the complete and well-calibrated characterization of uncertainty needed by decision makers and information providers. The BPF comes furnished with (i) the meta-Gaussian model, which fits meteorological data well as it allows all forms of marginal distribution functions, and nonlinear and heteroscedastic dependence structures, and (ii) the statistical procedures for estimation of parameters from asymmetric samples and for coping with nonstationarities in the predictand and the forecast due to the annual cycle and the lead time. A comprehensive illustration of the BPF is reported for forecasts of the daily maximum temperature issued with lead times of 1, 4, and 7 days for three stations in two seasons (cool and warm).

1. Introduction

a. The uncertainty quantification problem

The National Digital Forecast Database (NDFD) was designed by the National Weather Service (NWS) to store the official forecasts of the sensible weather elements produced by the NWS field offices throughout the United States (Glahn and Ruth 2003). The official forecasts are subjective in that they are made judgmentally by human forecasters with the support of software systems and are based on information from multiple sources, including output from numerical weather prediction models and guidance from the national centers. With the exception of the occurrence of precipitation, which is forecasted in terms of probability, all other weather elements are forecasted deterministically. Hence the deficiency of the NDFD: it contains no information about forecast uncertainty (Ryan 2003).

To remedy this deficiency, the Meteorological Development Laboratory of the NWS began developing statistical techniques for assessing the uncertainty in forecasts disseminated through the NDFD (Peroutka et al. 2005). This article presents a solution to the same problem, but via a different technique and in a different format.

b. Bayesian processor of forecast

The Bayesian processor of forecast (BPF) for the NDFD is a specialized application of the Bayesian theory of probabilistic forecasting formulated and tested in various settings over the past two decades (e.g., Krzysztofowicz 1983; Alexandridis and Krzysztofowicz 1985; Krzysztofowicz and Watada 1986; Krzysztofowicz and Reese 1991; Krzysztofowicz 1999; Krzysztofowicz and Kelly 2000b).

The BPF developed and illustrated herein quantifies the uncertainty in a deterministic forecast of the daily maximum temperature—one of the predictands se-

Recent *WAF* Apr. 2008 paper proposing a new calibration method, "BPF," or "Bayesian Processor of Forecasts." Hypothesis is that it may be appealing for calibration because it may leverage long-term climatological information, lessening the need for long training data sets.

Actively being tested at NCEP/EMC and ESRL/PSD, focus on precipitation.

Starting from basics

- Working front to back through Krzysztofowicz & Evans *WAF* article, it's pretty dense and tough at first to see the forest for the trees. Lots of transformation of variables.

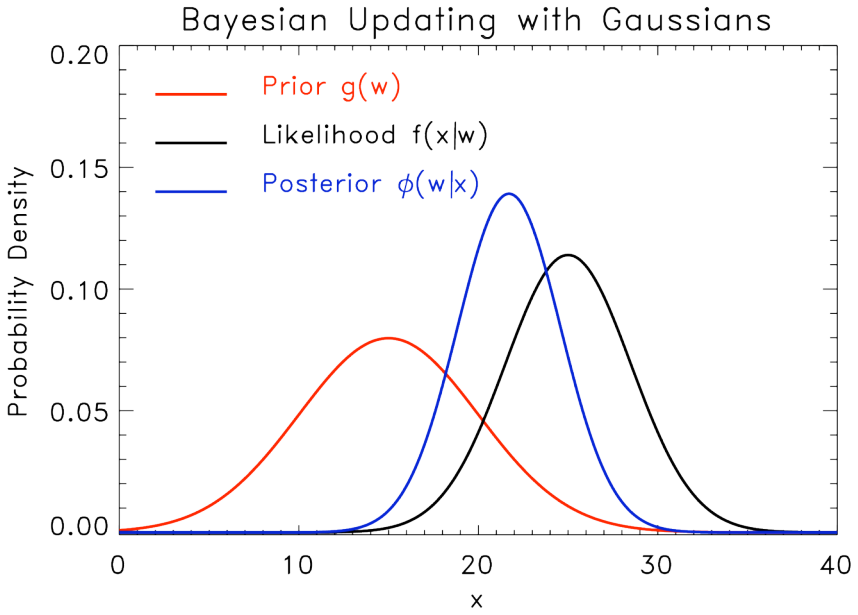
- After careful reading, the essence of the technique, once data is transformed to be Gaussian, is thankfully rather simple. Let's review this first.

Key simplifying assumption 1: products of prior & likelihood functions are easy to evaluate when distributions are Gaussian

- Let $f(x|w) \sim N(\mu_a, \sigma_a^2)$, and let $g(w) \sim N(\mu_b, \sigma_b^2)$.

Then

$$f(x|w)g(w) \sim N\left(\mu_a \frac{\sigma_b^2}{\sigma_a^2+\sigma_b^2} + \mu_b \frac{\sigma_a^2}{\sigma_a^2+\sigma_b^2}, \left(\sigma_a^{-2}+\sigma_b^{-2}\right)^{-1}\right)$$



- Normality of posterior preserved.
- Parameters of posterior are simple functions of prior, likelihood parameters

Somewhat more realistic assumptions

- Let $f\left(x|w\right) \sim N\left(a_x w + b_x, \sigma_x^2\right)$ (i.e., regress sample x on w).
  Let $g\left(w\right) \sim N\left(\mu_w, \sigma_w^2\right)$. Then

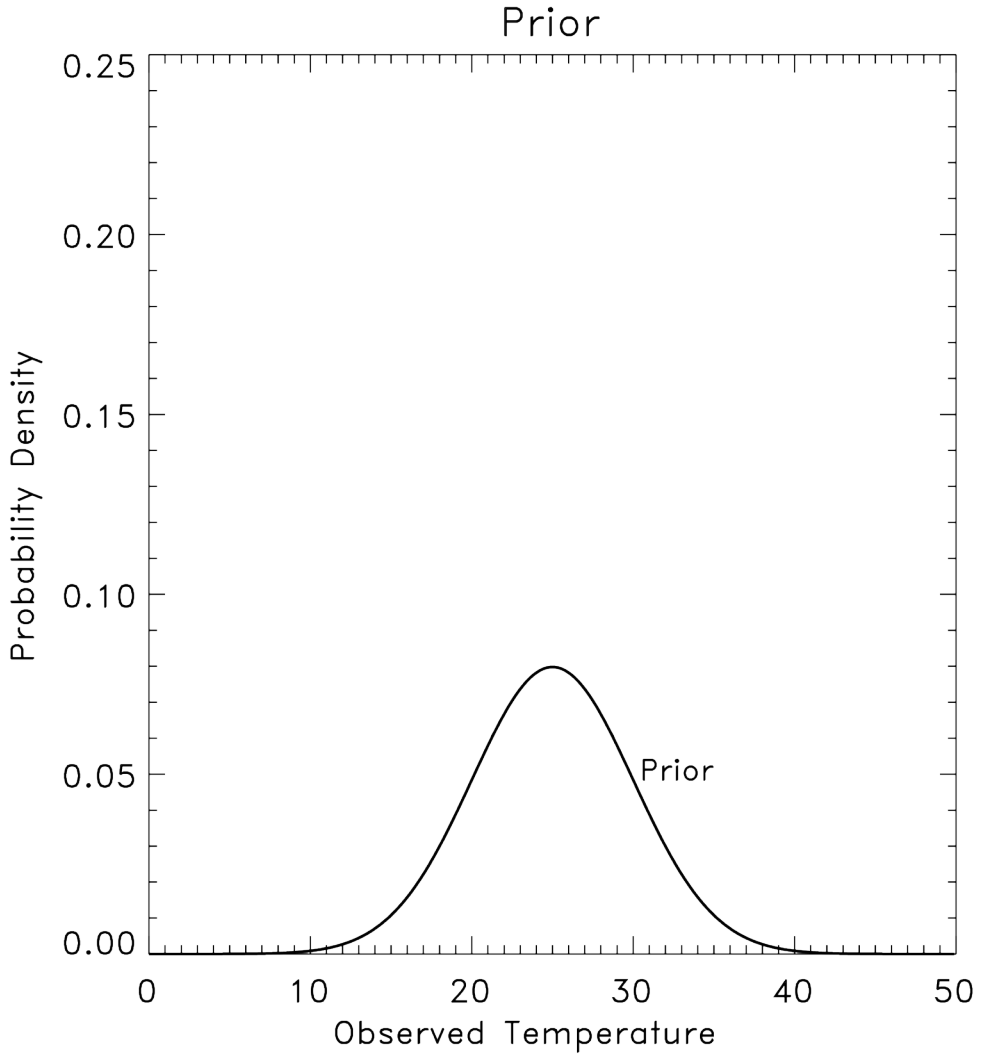
$$\phi\left(w|x\right) = \frac{f\left(x|w\right)g\left(w\right)}{\int f\left(x|w\right)g\left(w\right)dw} \quad (Bayes\ Rule)$$

$$\phi\left(w|x\right) = N\left(Ax + B, s_x^2\right)\ where$$

$$A = \frac{a_x\sigma_w^2}{\sigma_x^2 + a_x^2\sigma_w^2}, \quad B = \frac{\mu_w\sigma_x^2 - a_x b_x \sigma_w^2}{\sigma_x^2 + a_x^2\sigma_w^2}, \quad s_x^2 = \frac{\sigma_x^2\sigma_w^2}{\sigma_x^2 + a_x^2\sigma_w^2}$$

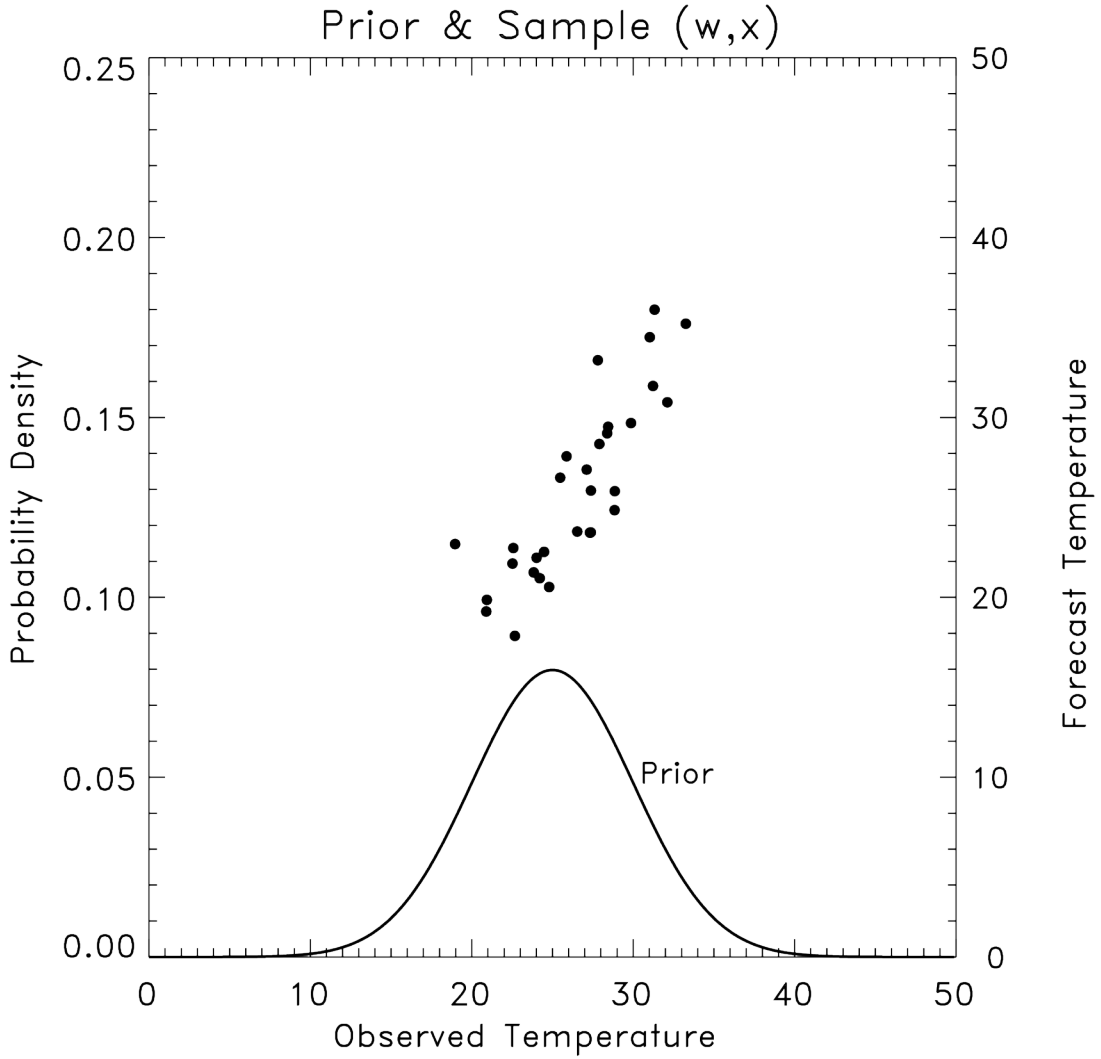
This from Krzysztofowicz 1987 JASA, employs theory of conjugate family of distributions (see also Degroot 1970 therein). These equations are basically eq (24) from K.&Evans, MWR, 2008, but there g(w) is standard normal with mean 0.0 and variance 1.0.

Example



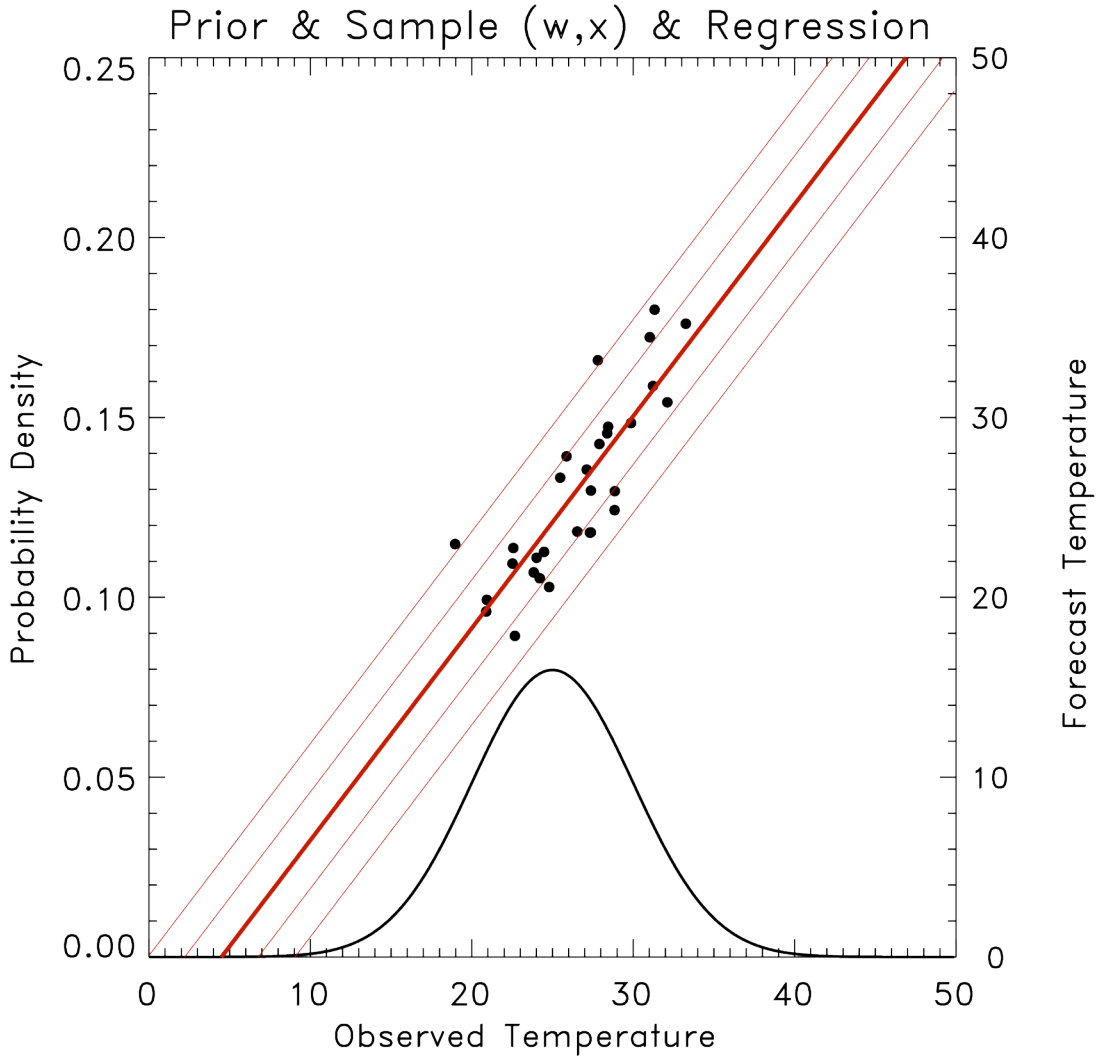
prior distribution estimated from observed climatology

Example



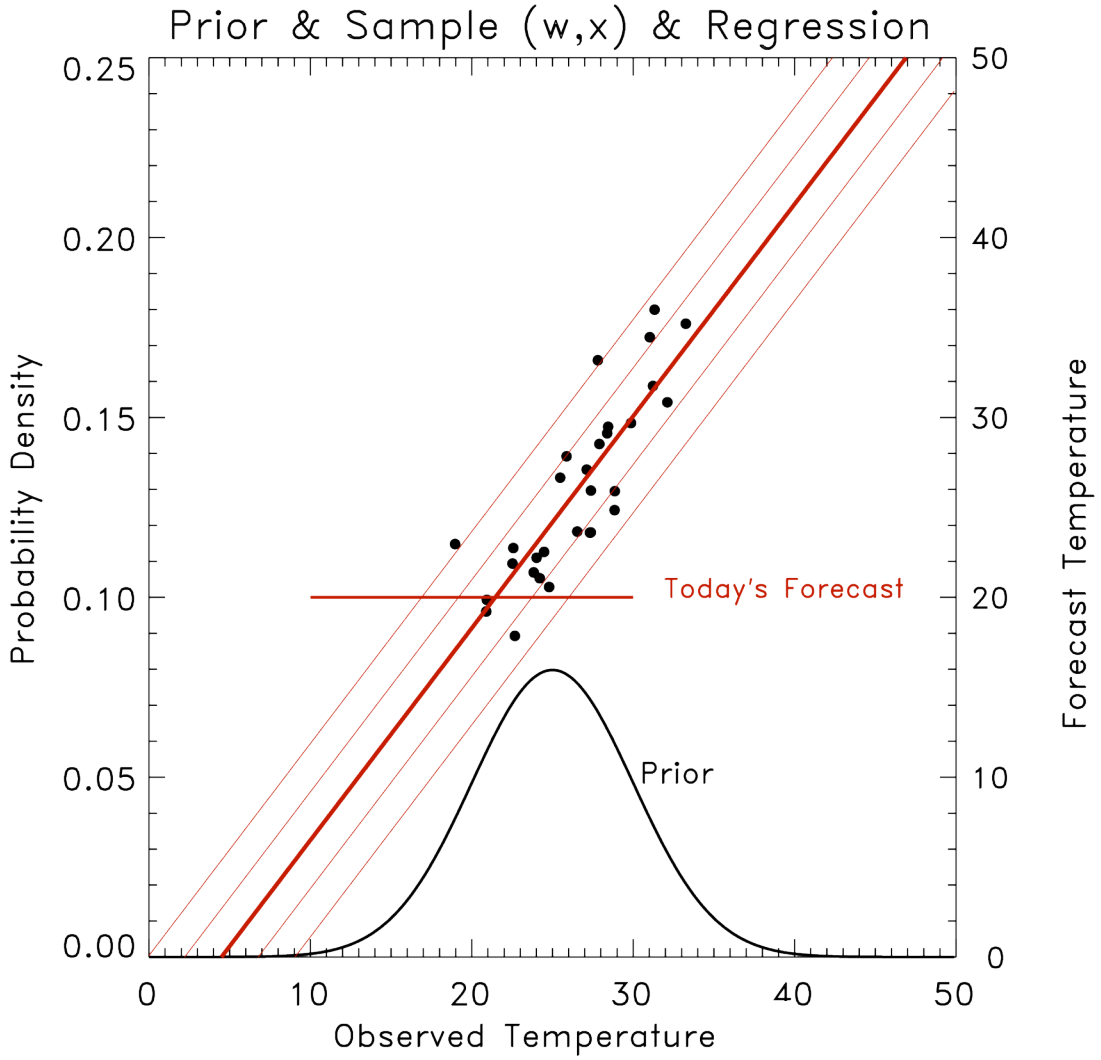
recent sample of observed (w; abscissa) and forecast (x; ordinate)

Example



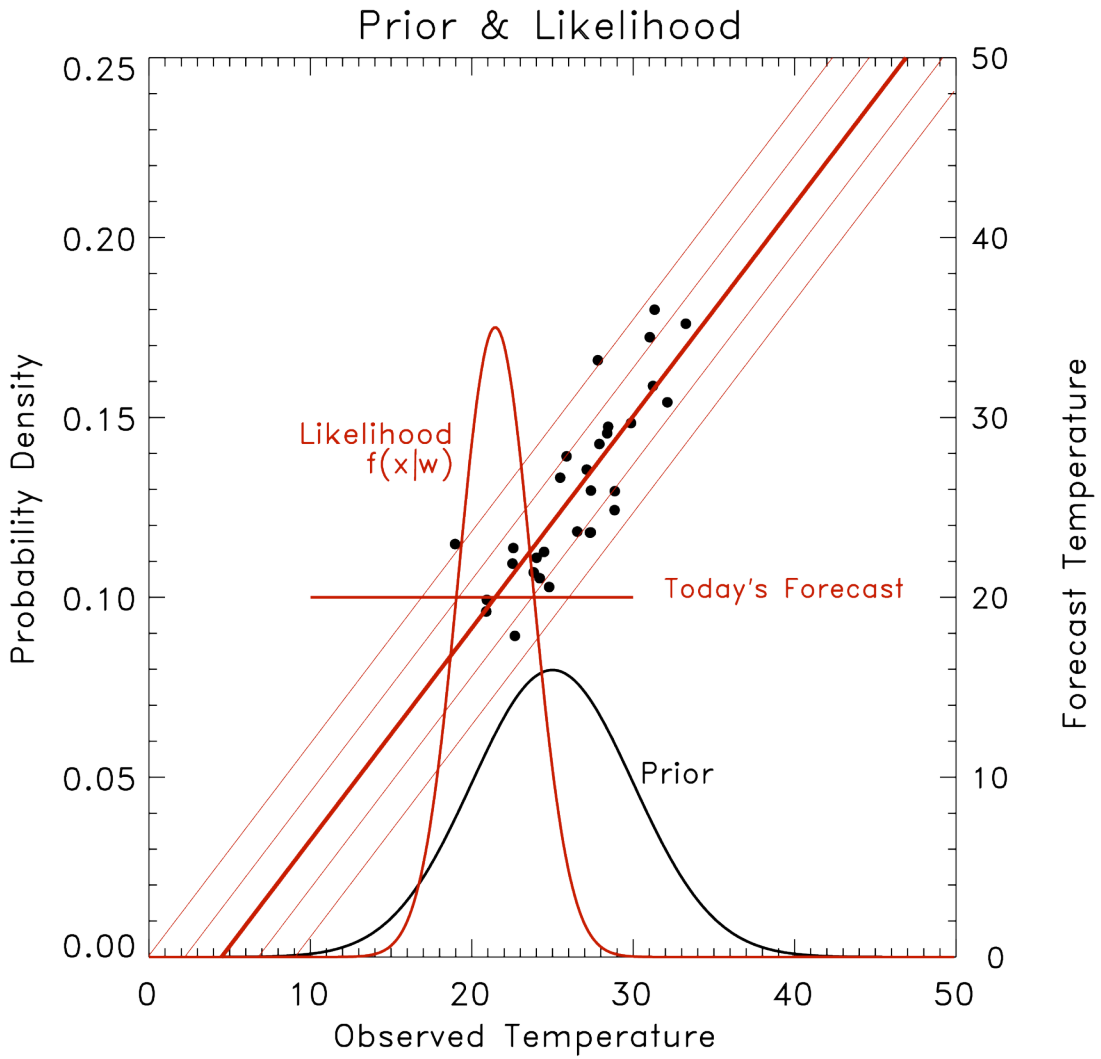
linear regression relationship f(x|w) with 1- and 2-σ confidence intervals

Example



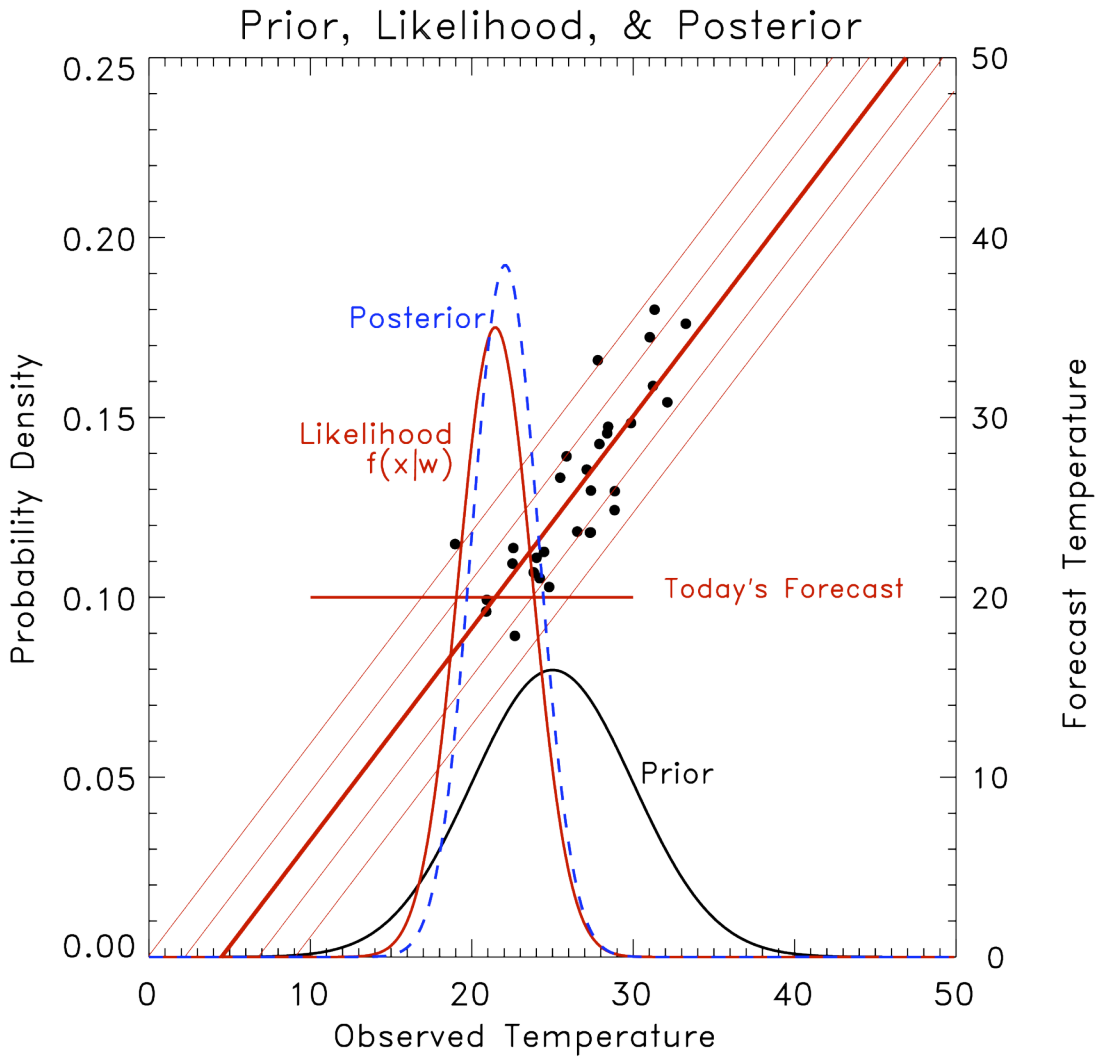
now suppose today's forecast is + 20C

Example



estimate the likelihood function based on regression relationship

Example



posterior obtained by application of Bayes' rule; product multiplication and normalization of prior & likelihood, or equivalently, application of equations on slide 6.

Essence of how it works in Krzysztofowicz & Evans

- Determine a parametric best fit distribution (Weibull) to climatology, and a mapping from Weibull to standard normal distribution.
- Get smaller training data set of obs & forecast (w,x).
  - Transform w with previously determined Weibull for climatology
  - Determine a new, separate parametric best fit distribution for x; map x to standard normal.
  - Perform regression analysis to predict x|w in standard normal space.
- Given today's forecast x, determine likelihood function, (conditional distribution of standard normalized x|w) and apply Bayes rule to predict posterior distribution in standard normal space.
- Remap this distribution back to its original coordinates.

In equations

Z is random variable representing transformed forecast X.
V is random variable representing transformed obs W.
ν is specific quantity of V.

$$E\left(Z|V=\nu\right)=a\nu+b$$
$$Var\left(Z|V=\nu\right)=\sigma^2$$

regression of transformed w, x to determine a, b, σ^2

$$A=\frac{a}{a^2+\sigma^2}$$
$$B=\frac{-ab}{a^2+\sigma^2}$$
$$T^2=\frac{\sigma^2}{a^2+\sigma^2}$$

maps cumulative probability to standard normal deviate

maps forecast value x to cumulative probability of non-exceedance using distribution fitted from training data .

$$\Phi(w)=Q\left(\frac{1}{T}\left[Q^{-1}\left(G(w)\right)-AQ^{-1}\left(K(x)\right)-B\right]\right)$$

maps observed w to cumulative probability of non exceedance using distribution fitted from long-term climatological training data .

maps standard normal deviate to cumulative probability.

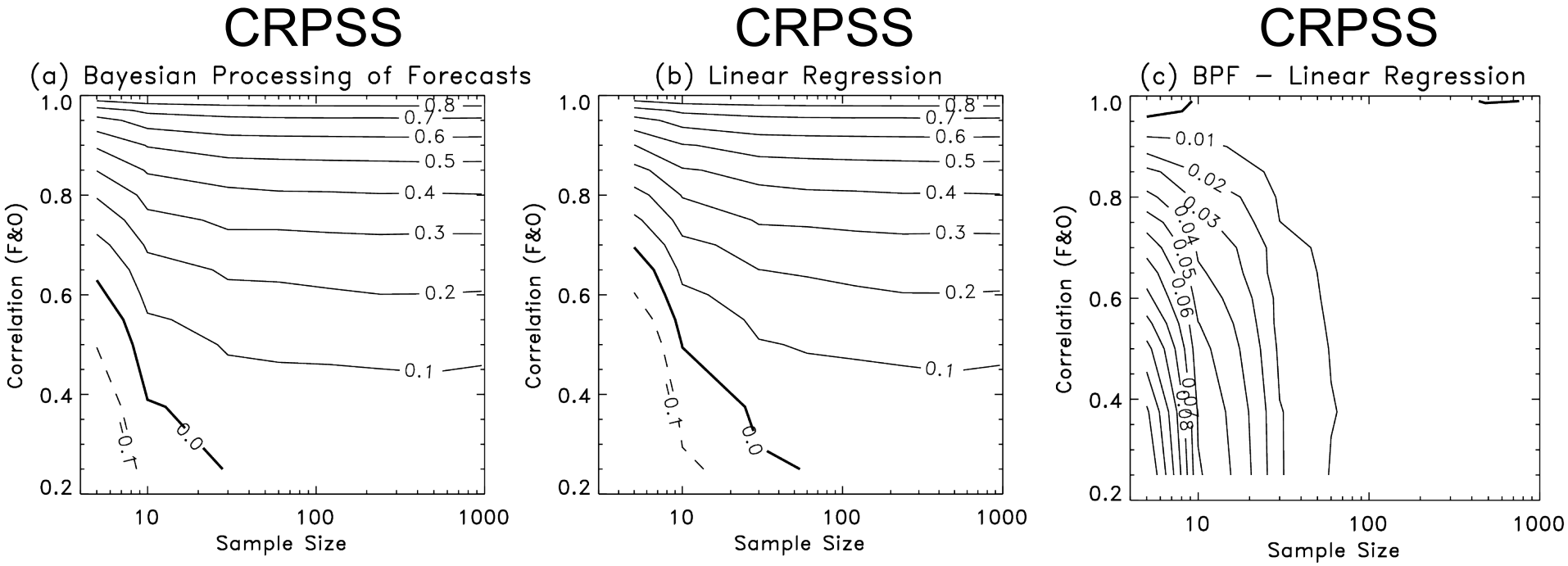
Before testing with real data…

- Let's verify that it works well for synthetic data
  - Everything is already standard normal, so we strip the remapping of distributions from the problem.
  - Can test against known standard, like linear regression algorithm used in MOS.

Test case setup: ~N(0,1), no autocorrelation

- Climatology estimated from 10,000 *iid* samples drawn from ~N(0,1)
- Forecast, observed drawn from ~N(0,1); autocorrelation=0.0; Test correlations of forecast and observed from 0.25 to 0.99. Test sample sizes of 5,10, 30, 60, 120, 240, 480, and 960.
- Replicate process 40,000 times, calculate Continuous Ranked Probability Skill Score (CRPSS) in standard manner.

Results: ~N(0,1), no autocorrelation



Only for small samples sizes (<30) and low forecast skill (measured in correlation of forecast and observed) is there much difference in skill. Then BPF the winner.

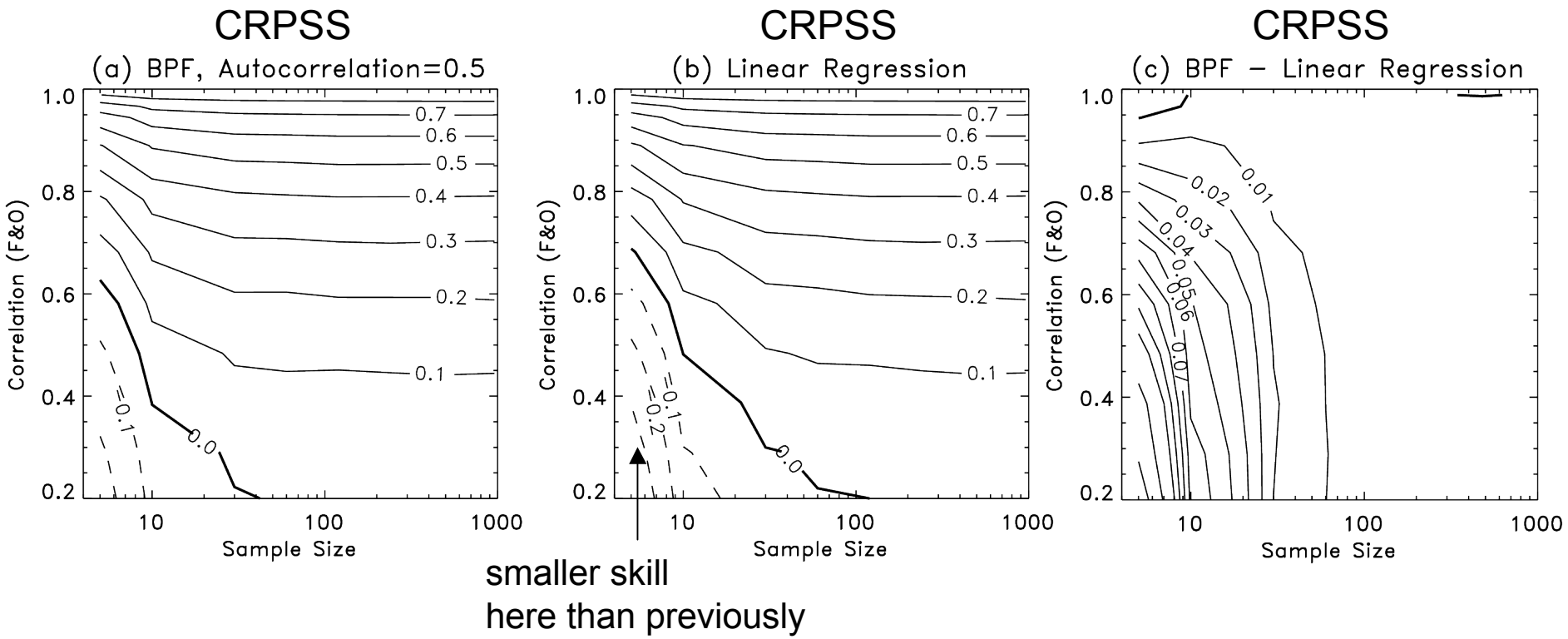
Test case setup: ~N(0,1), 0.5 lag-1 autocorrelation

- Climatology estimated from 10,000 iid samples drawn from ~N(0,1), autocorrelation = 0.5 (typical of surface temperature data)[1]

$$x_{t+1} - \mu = 0.5\left(x_t - \mu\right) + \varepsilon_{t+1}, \quad \varepsilon_{t+1} \sim N(0,1)$$

- Forecast, observed drawn from ~N(0,1); autocorrelation = 0.5; Test correlations of forecast and observed from 0.25 to 0.99. Test sample sizes of 5,10, 30, 60, 120, 240.
- Replicate process 40,000 times, calculate CRPSS in standard manner.

Results: ~N(0,1), 0.5 autocorrelation

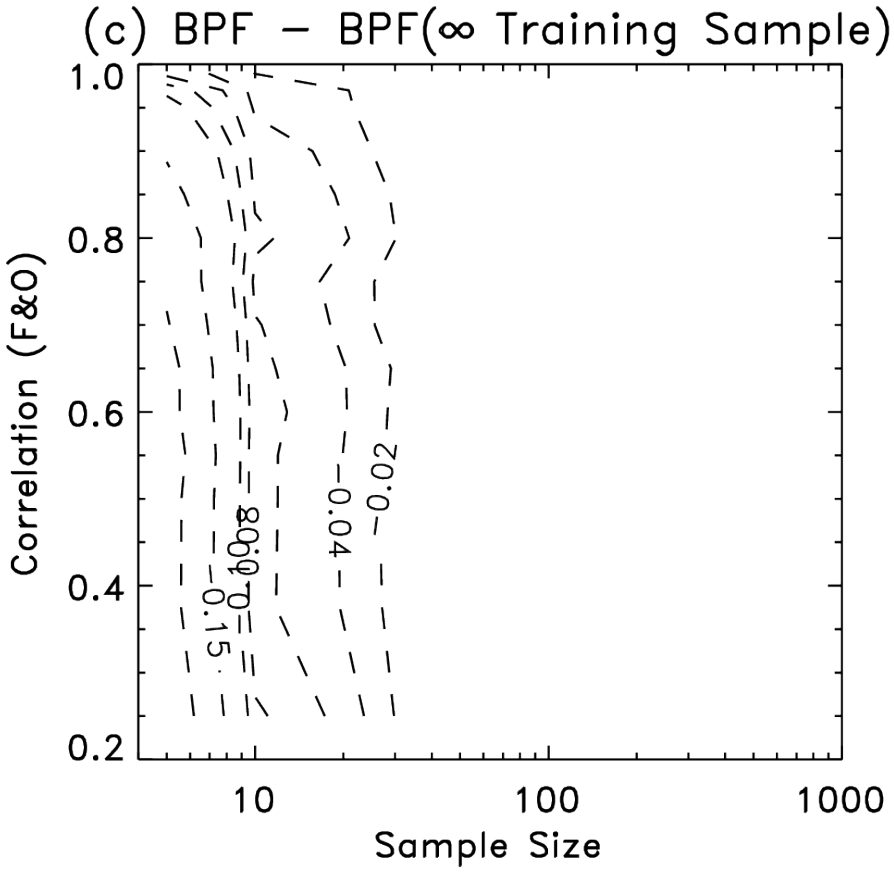


smaller skill here than previously

Qualitatively, not much difference relative to 0.0 autocorrelation, though skill at smallest sample size and lowest correlations is somewhat smaller, as expected. BPF still outperforms linear regression at low (F,O) correlation, small sample size. Sample size of ~60 adequate, little improvement from more samples.

BPF CRPSS, finite-infinite sample size

- Here, the skill of a finite sample is subtracted from the skill of an effectively infinite sample.
- By ~50 samples, most of the benefit of infinite sample achieved.



Comments / questions / issues

- BPF technique may not be as easily extended to multiple predictors as linear regression. Has conjugate family math been worked out for multiple predictors as with single predictor?

$$\phi\left(w\middle|x_1,x_2\right)=\frac{f\left(x_1,x_2\middle|w\right)g\left(w\right)}{\int f\left(x_1,x_2\middle|w\right)g\left(w\right)dw} \quad (Bayes\ Rule)$$

$$\phi\left(w\middle|x_1,x_2\right)=N\left(A_1x_1+A_2x_2+B,s_x^2\right)\ where$$

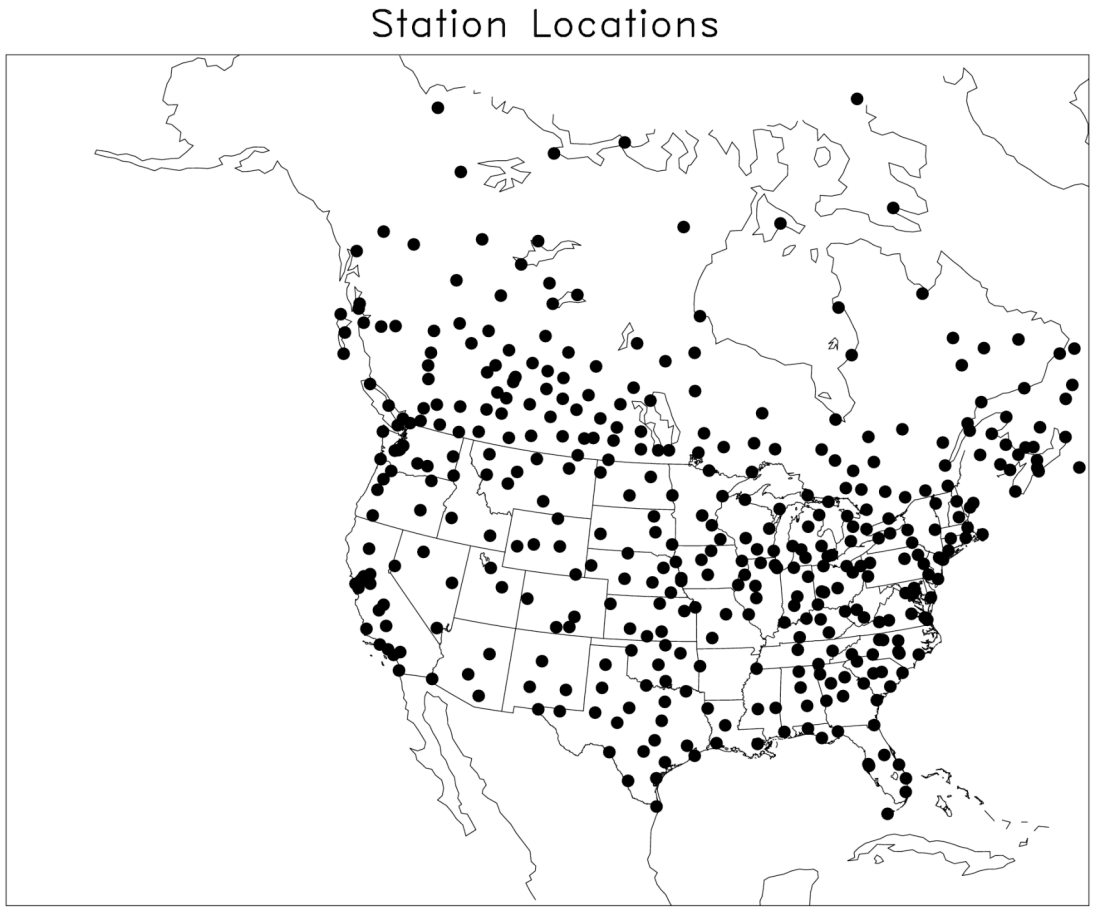
$$A_1=?,\quad A_2=?,\quad B=?,\quad s_x^2=?$$

  - If multiple linear regression better than linear regression, relative improvement of BPF over regression techniques may be exaggerated.
- Similarly, what about BPF using ensembles of forecasts?
- Experimental setup did not include state-dependent bias, which is common. Such bias may increase required sample size.
- *Haven't included any of the mechanics of K/E paper for dealing with non-normal distributions.*

On to real surface temperature data…

- Recently published an article with Renate Hagedorn on temp, precip reforecast skill with GFS/ECMWF reforecast data sets.
- Conclusion: for 2-meter temperature, short training data set adequate. Used “non-homogeneous Gaussian regression.” (NGR)
- More skill yet to be obtained if BPF used instead of NGR?

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.

Forecast data used

- Fall 2005 GFS 2-meter ensemble forecast temperature data from reforecast data set.
- Forecasts computed 1 Sep - 1 Dec.; examine leads of 1/2 day to 10 days.
- Training data sets:
  - Prior 30 days of (w,x); don't evaluate if < 20 available.
  - Reforecasts: 1982-2001 = 26 years*31 samples/year (+/- 15 days) of (w,x). Don't evaluate < 75% of reforecast (w,x) available

Calibration procedure 1: "NGR"
"Non-homogeneous Gaussian Regression"

- **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.
- **Predictors**: ensemble mean and ensemble spread
- **Output**: mean, spread of calibrated normal distribution

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

- **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.
- **Another disadvantage**: doesn't leverage long-term climatology like BPF?

Calibration procedure 2:
"Bias correction"

- Calculate bias *B* from training data set; for n days of samples, simply

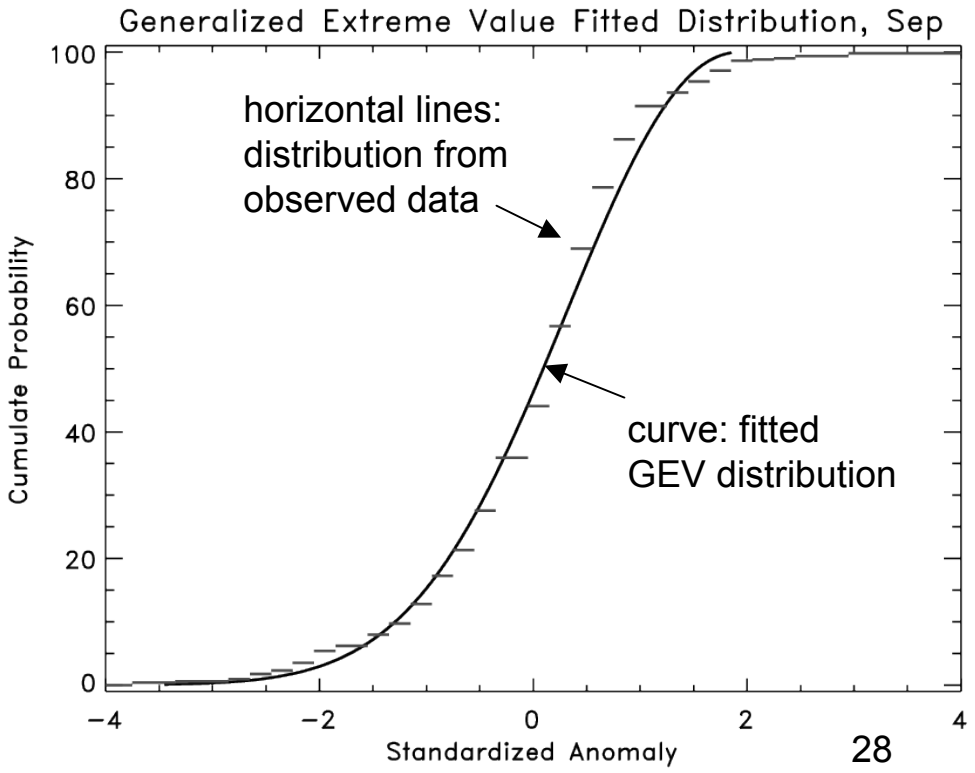
$$B = \frac{1}{n}\sum_{i=1}^{n}\left(\overline{x}_i - w_i\right)$$

- Subtract *B* from today's ensemble forecast

Problems with applying BPF using fitted Weibull / GEV?

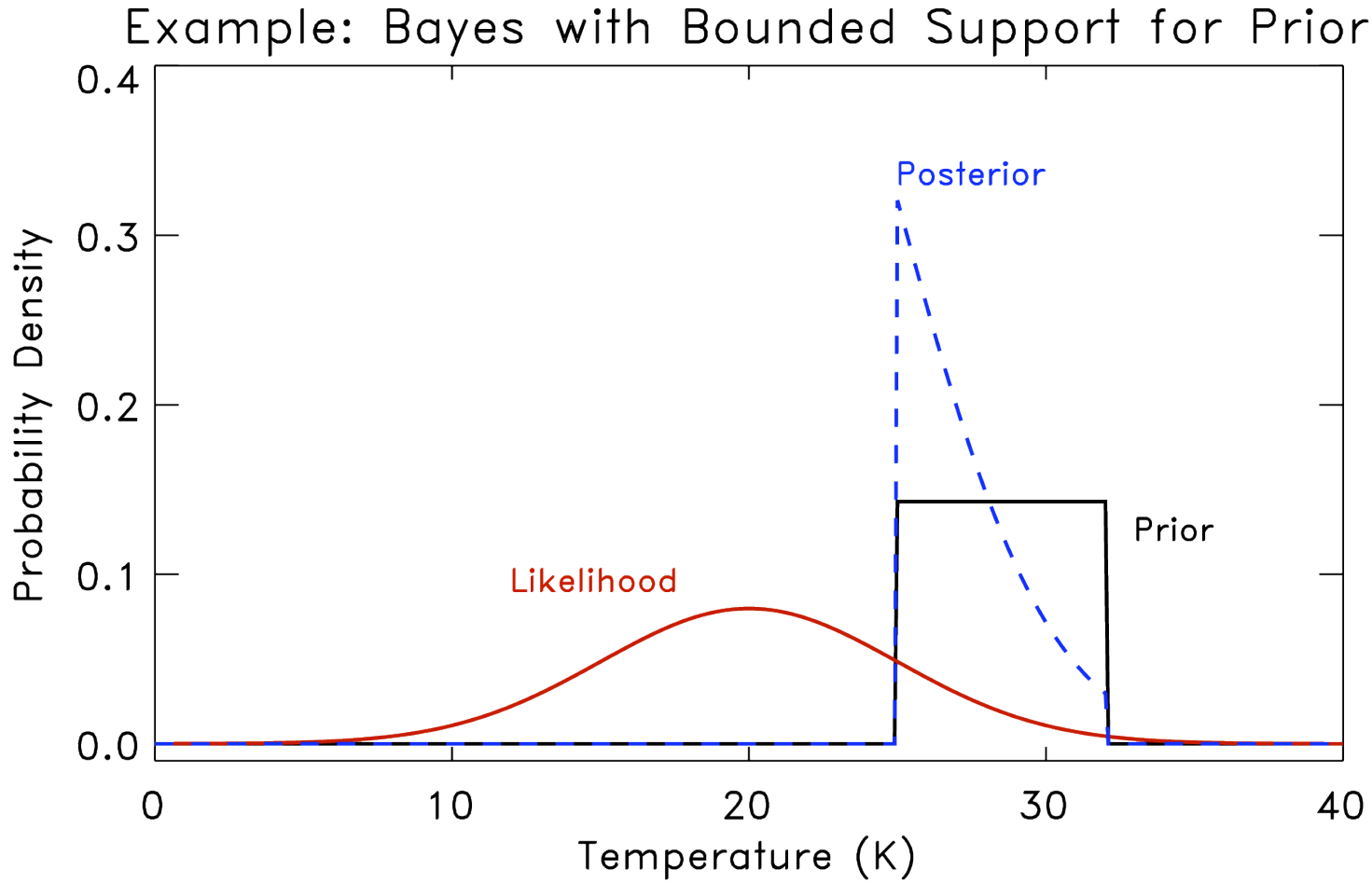
fitted prior has **zero** probability beyond this value, while 1-2 % of observed beyond this.

- BPF as proposed in K/E '08 paper fits a Weibull distribution to the prior and likelihood distributions, transforms them to a Gaussian.
- Need good parametric models for priors, likelihood. Weibull distribution (and related GEV) have "bounded support" and fits a distribution that has zero probability in tails.
- If prior has zero probability for a given temperature, posterior will have zero probability as well. In other words, **lousy forecasts of extreme events likely.**
- Other choices besides Weibull?

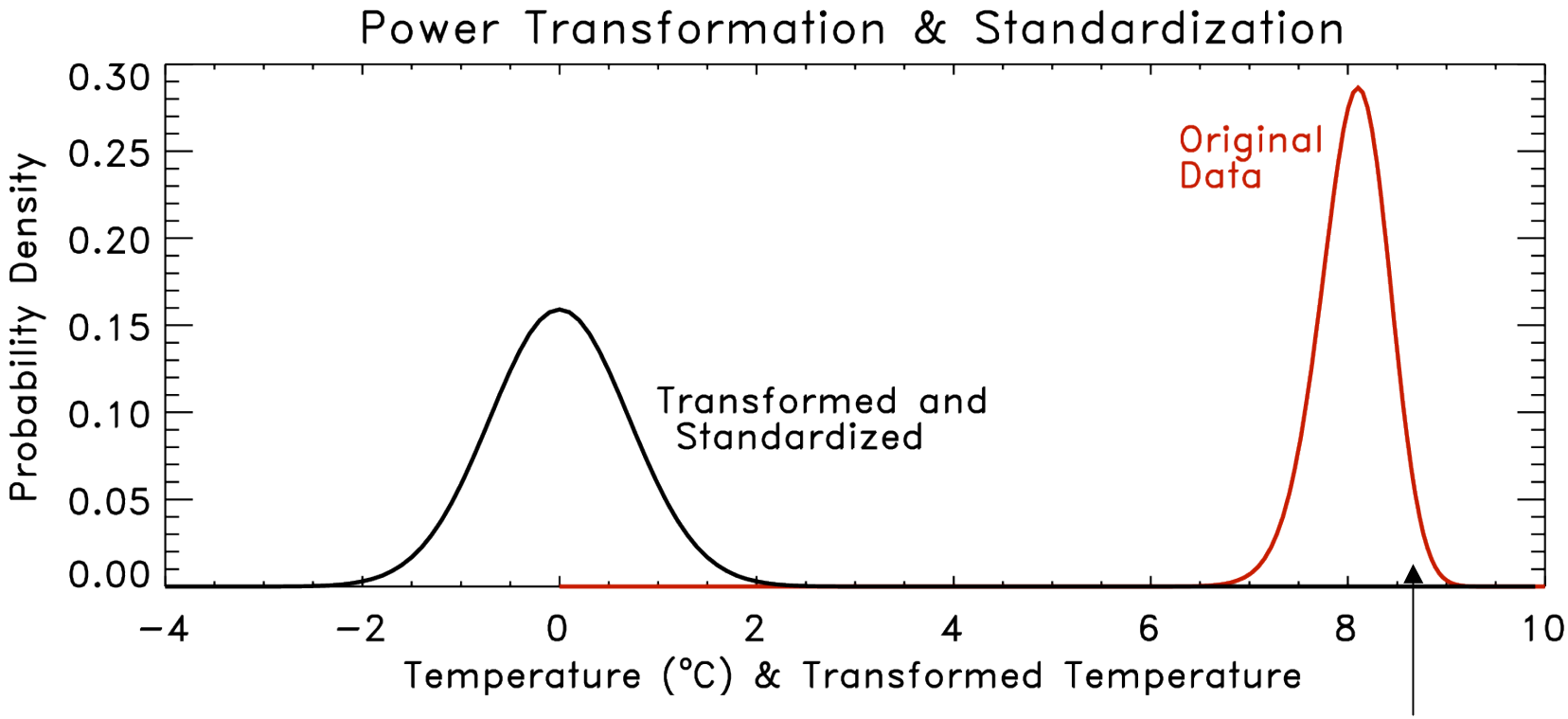


Note: GEV distribution fit with L-moments software from IBM web site.

Example: screwy posterior when prior has bounded support



Instead of Weibull/GEV, how about fitting distributions of power transformed variables, like $x_{new} = x_{old}^{\lambda}$?



see Wilks text, section 3.4.1 for more on power transformations.

Power transformations have trouble with negative data, e.g., $(-1)^{0.5}$

- Use new power transformation proposed by Yeo and Johnson, 2000, *Biometrika*, **87**, pp. 954-959. For variable x and possible exponent λ, the transformed variable ψ is

$$\psi(\lambda,x)=\begin{cases}\left\{(x+1)^{\lambda}-1\right\}/\lambda & (x\geq 0,\lambda\neq 0)\\ \log(x+1) & (x\geq 0,\lambda=0)\\ -\left\{(1-x)^{2-\lambda}-1\right\}/(2-\lambda) & (x<0,\lambda\neq 2)\\ -\log(x+1) & (x<0,\lambda=2)\end{cases}$$

Proposed method of using power transformations to convert distribution to standard normal

For a given sample of data (e.g., time series of observed temperature) :

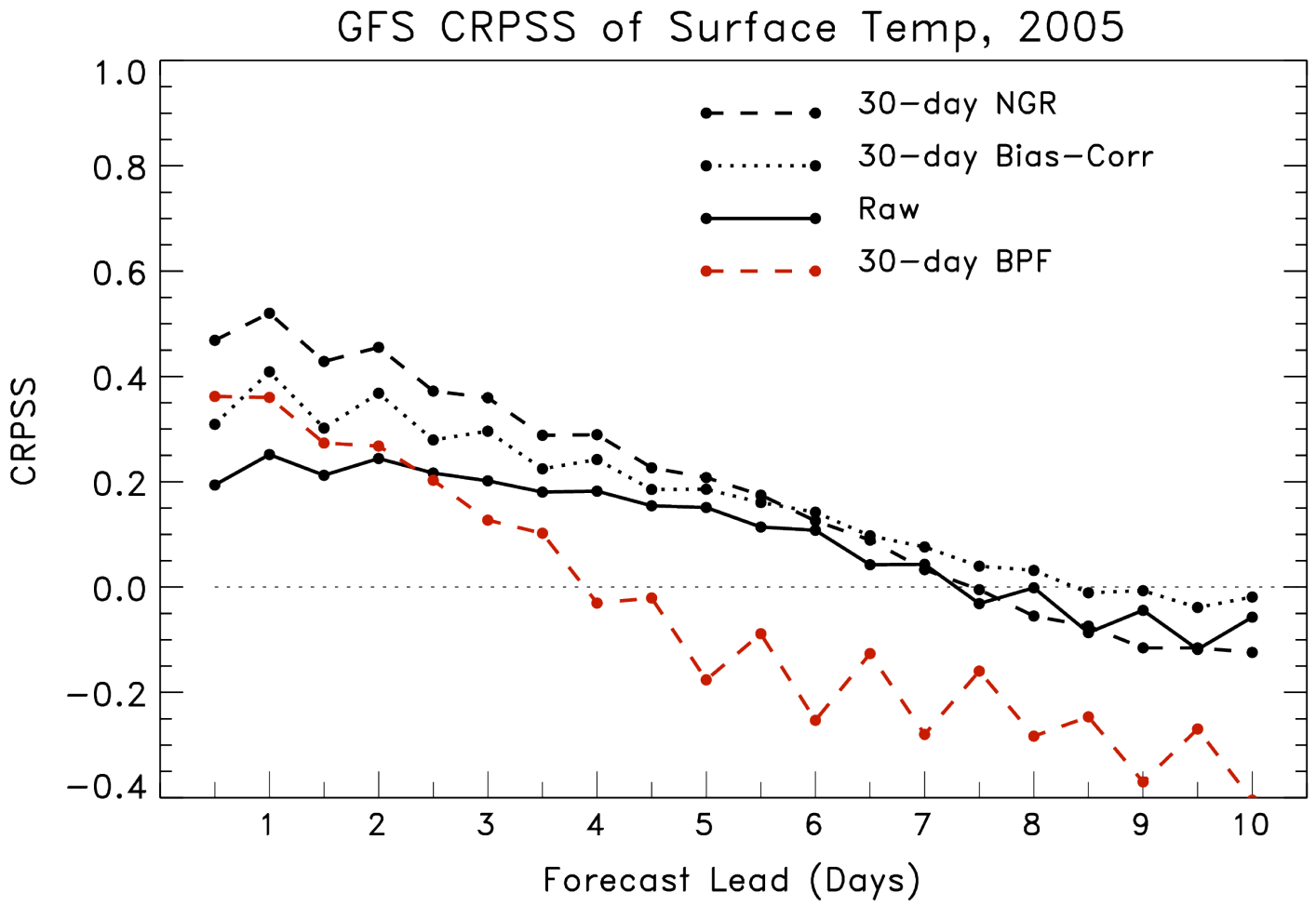
(1) Determine sample mean and standard deviation

(2) Normalize data, subtracting mean and dividing by standard deviation

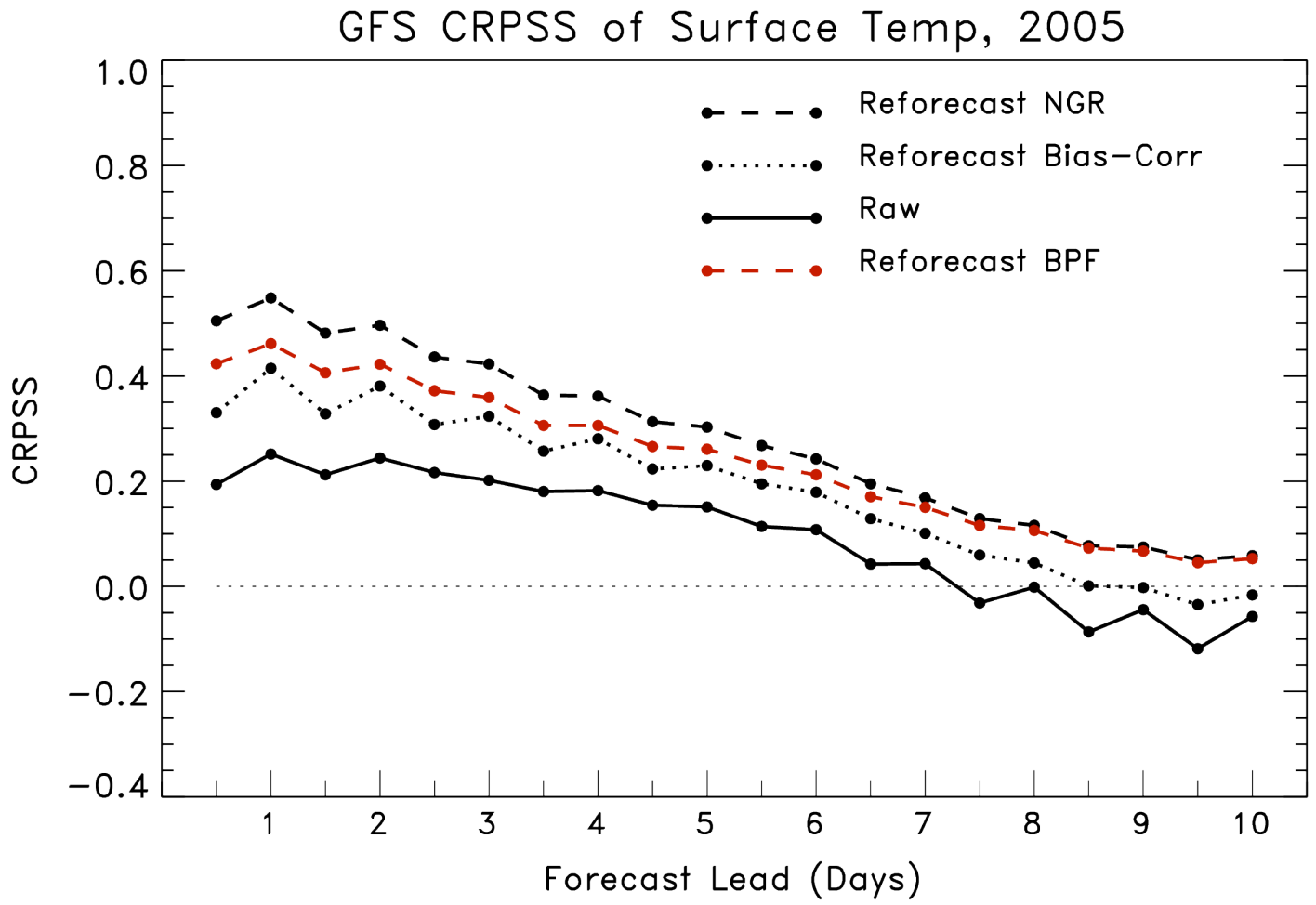
(3) Loop over a set of possible exponents for power transformations between 0.25 and 3.0
   - (a) Perform the power transformation of Yeo and Johnson (previous page)
   - (b) Determine sample mean and standard deviation
   - (c) Normalize data *again*, subtracting mean and dividing by standard deviation
   - (d) Compare CDF of transformed against standard normal CDF, and keep track of the fit for this exponent.

(4) Choose and use the exponent of the power transformation that gave a best fit. Note: (save 5 parameters: (1) original sample mean, (2) original sample standard deviation, (2) exponent of power transformation (4) transformed sample mean, and (5) transformed sample standard deviation. With these 5 parameters, can map from original coordinates to standard normal.

Results, 30-day training data



Results, reforecast training data

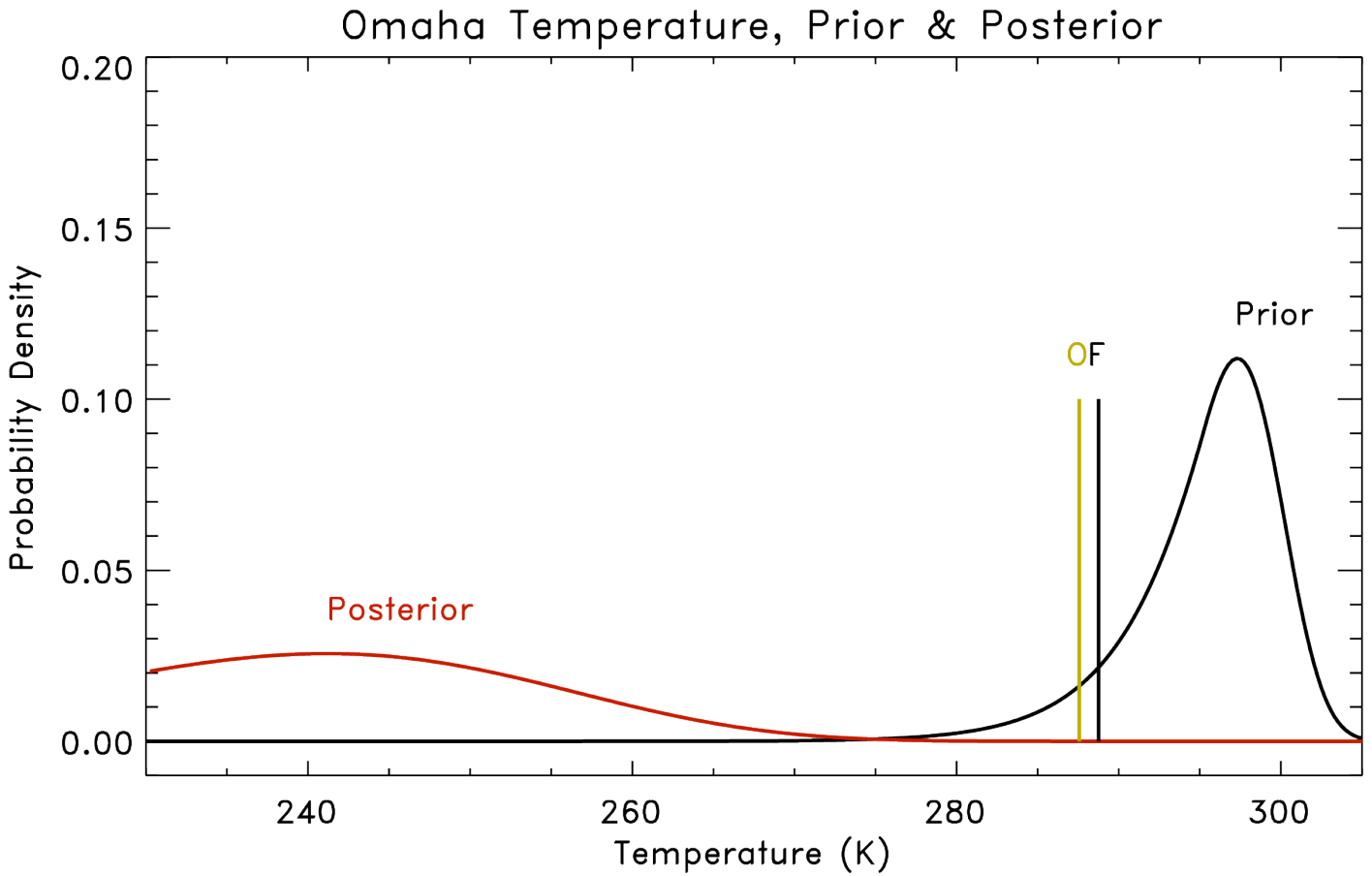


Note: skills are somewhat higher than in MWR papers; there I mistakenly used 2005 data in the computation of the climatology, so the climatological reference CRPS is too small.

Questions

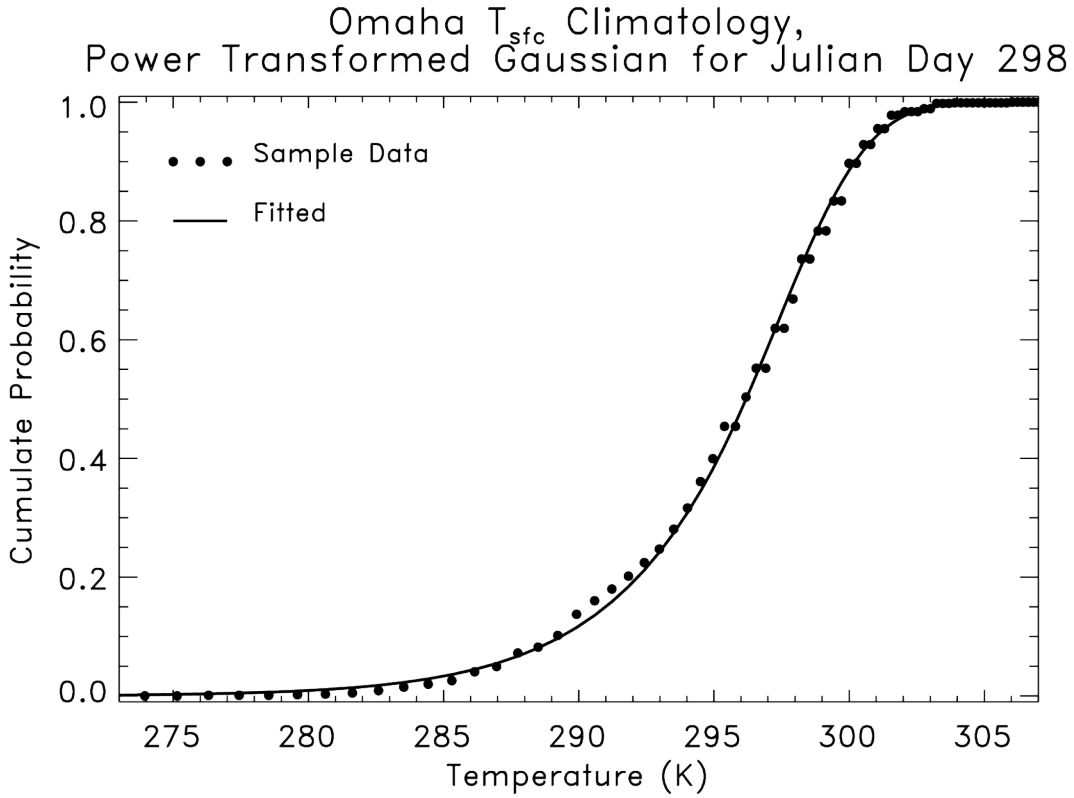
- Why isn't BPF better than NGR at all leads, as suggested from synthetic data results?
- Why is BPF, which a priori ought to be at greatest advantage with small training data sets, comparatively worse with the 30-day training data set relative to multi-decadal reforecast training data set?
- Are there adjustments to the BPF algorithm that can improve it?

Pathological example: Omaha, NE, October 24, 2005



These busts are not frequent, but when they happen, they can make an unbelievably bad forecast.

Is the prior somehow screwy? No.

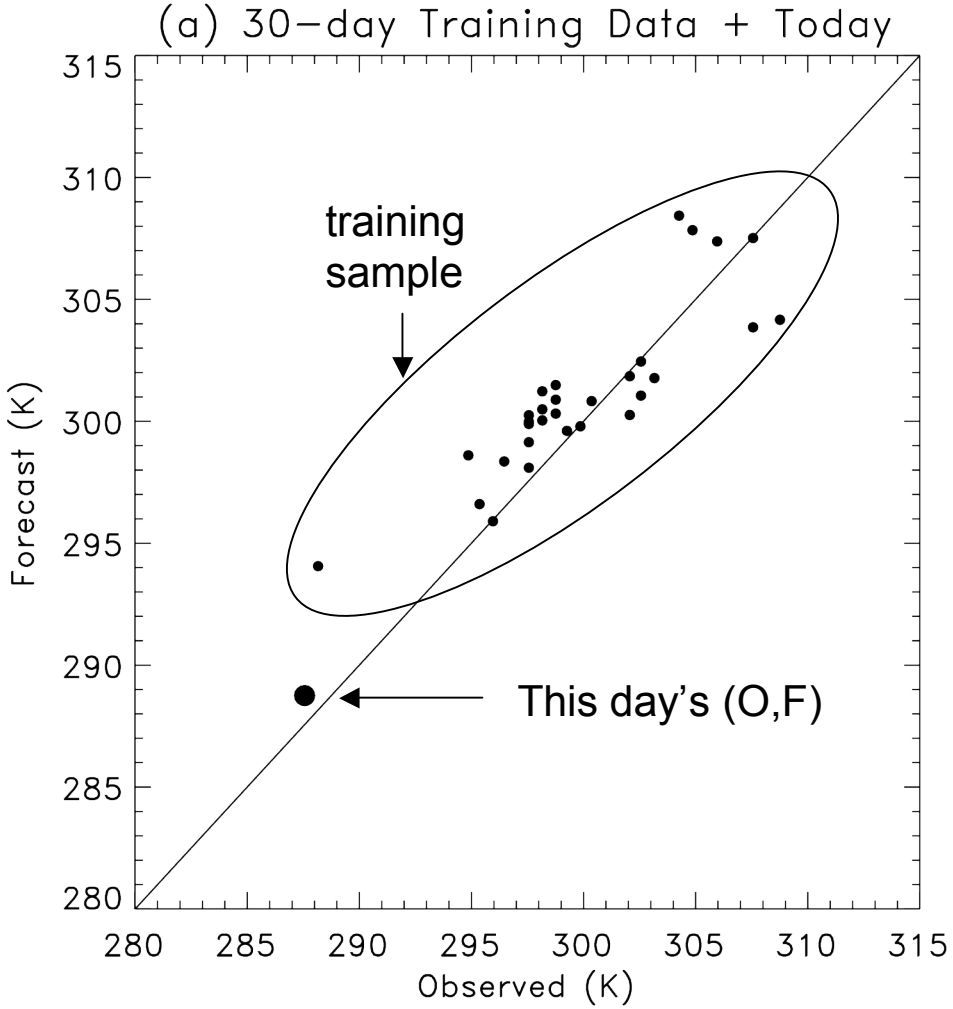


1980-2004 observations, 41 days centered on date of interest (+ / - 20 days)

Start with fitting normal distribution to power-transformed climatological data. Steps:
(1) Normalize data, subtracting mean, dividing by standard deviation.
(2) Test variety of power transformations, choose the one that provides the best fit to standard Gaussian after power transformation and second normalization.

Reasonable fit with exponent of 1.6

Anything obviously wrong with the training data for likelihood?



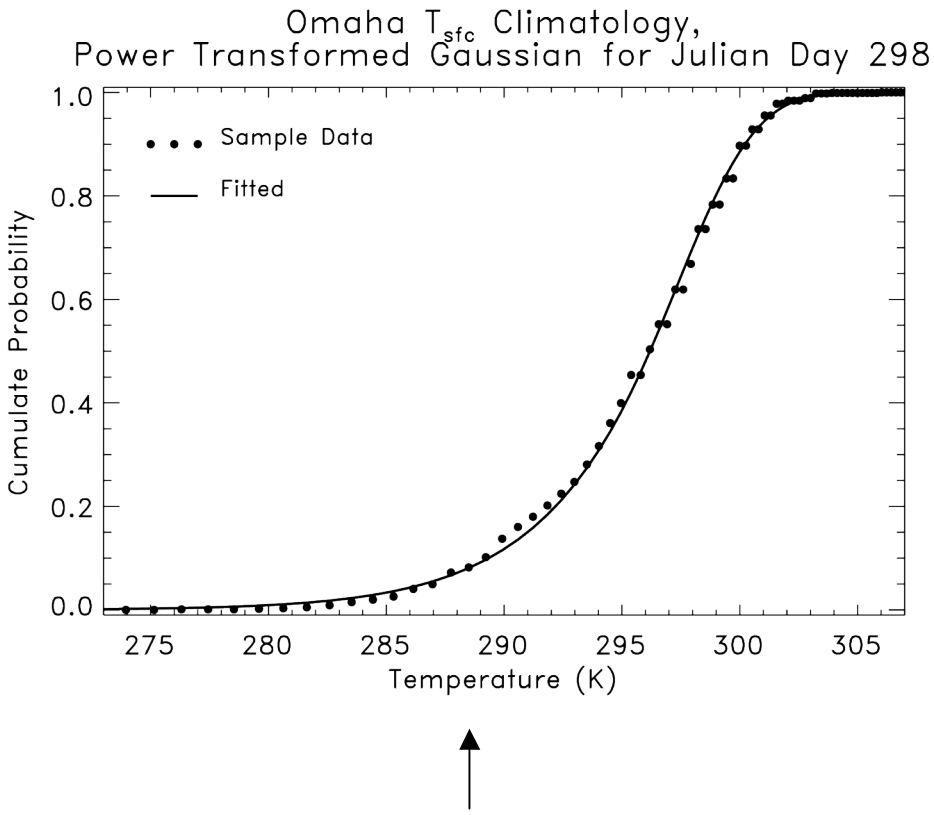
mean F = 297.3

mean O = 295.4

~2-degree warm bias in forecast.

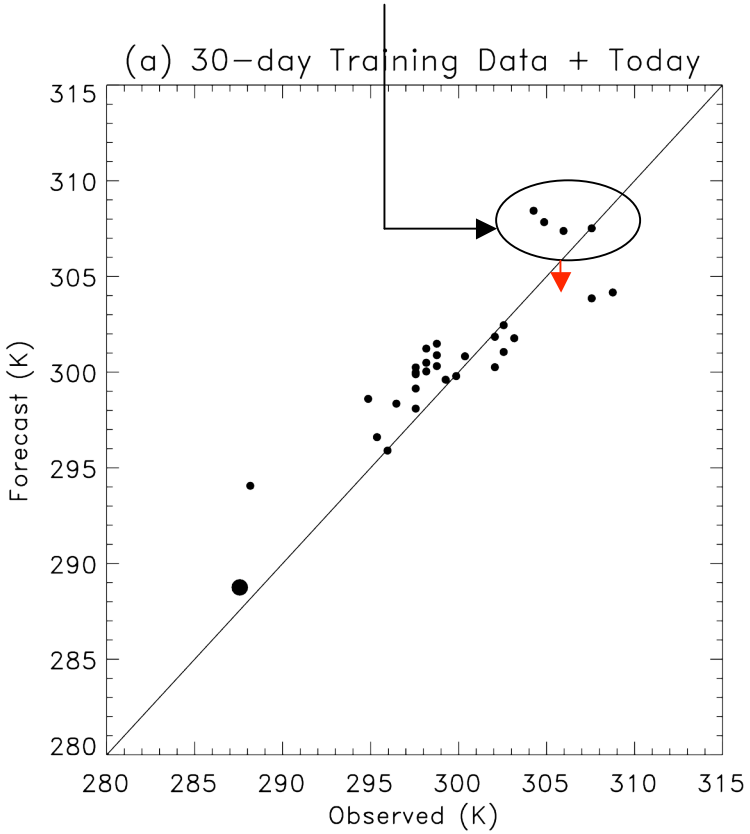
Today's F outside range of training data, though.

Is the remapping of power-transformed variables a source of error?

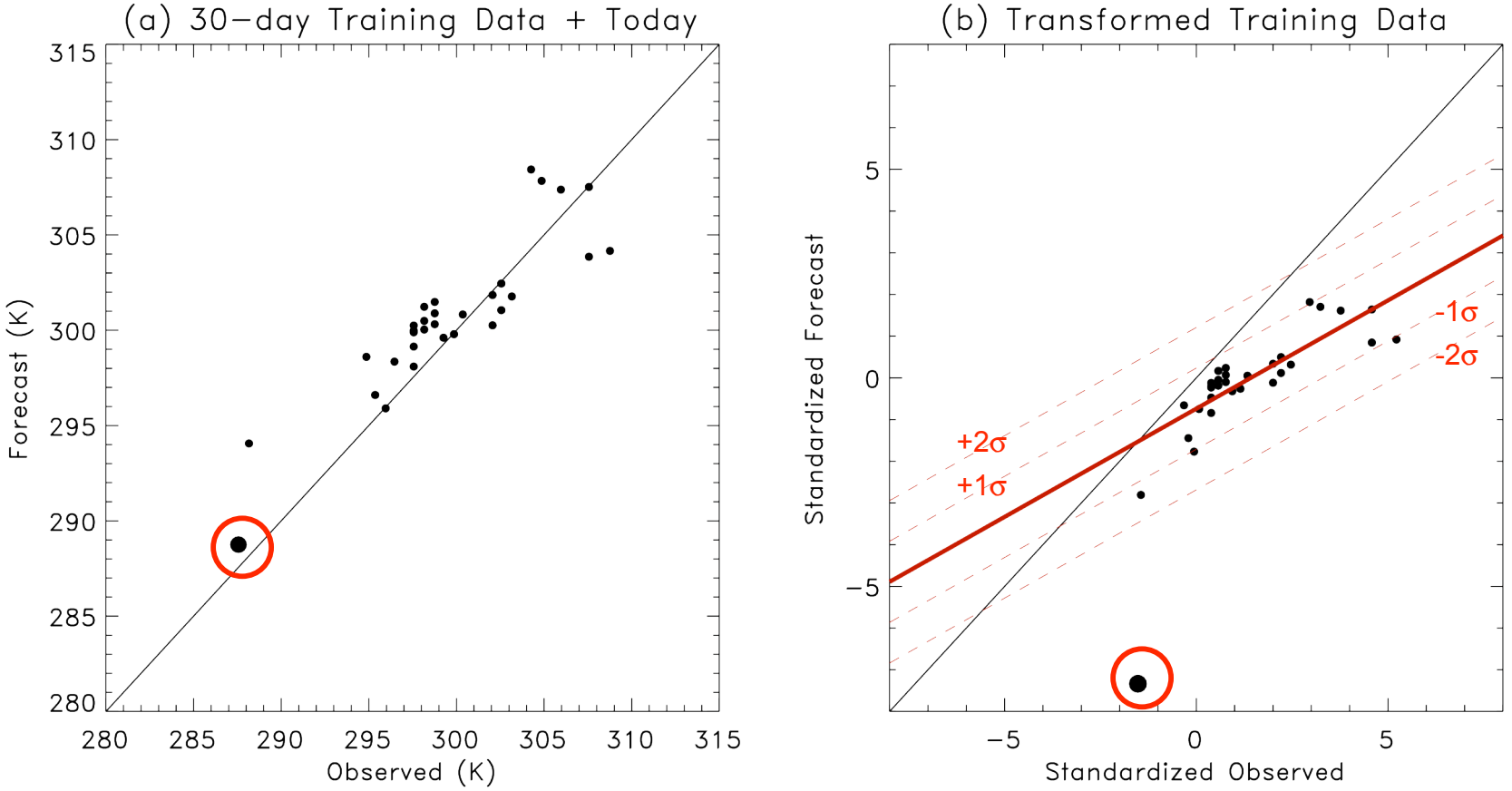


Recall that best power transformation to apply to observed to make ~normal was to raise to power of 1.6

For the forecast training data, a power transformation of $^{0.25}$ selected automatically to pull in these outliers.

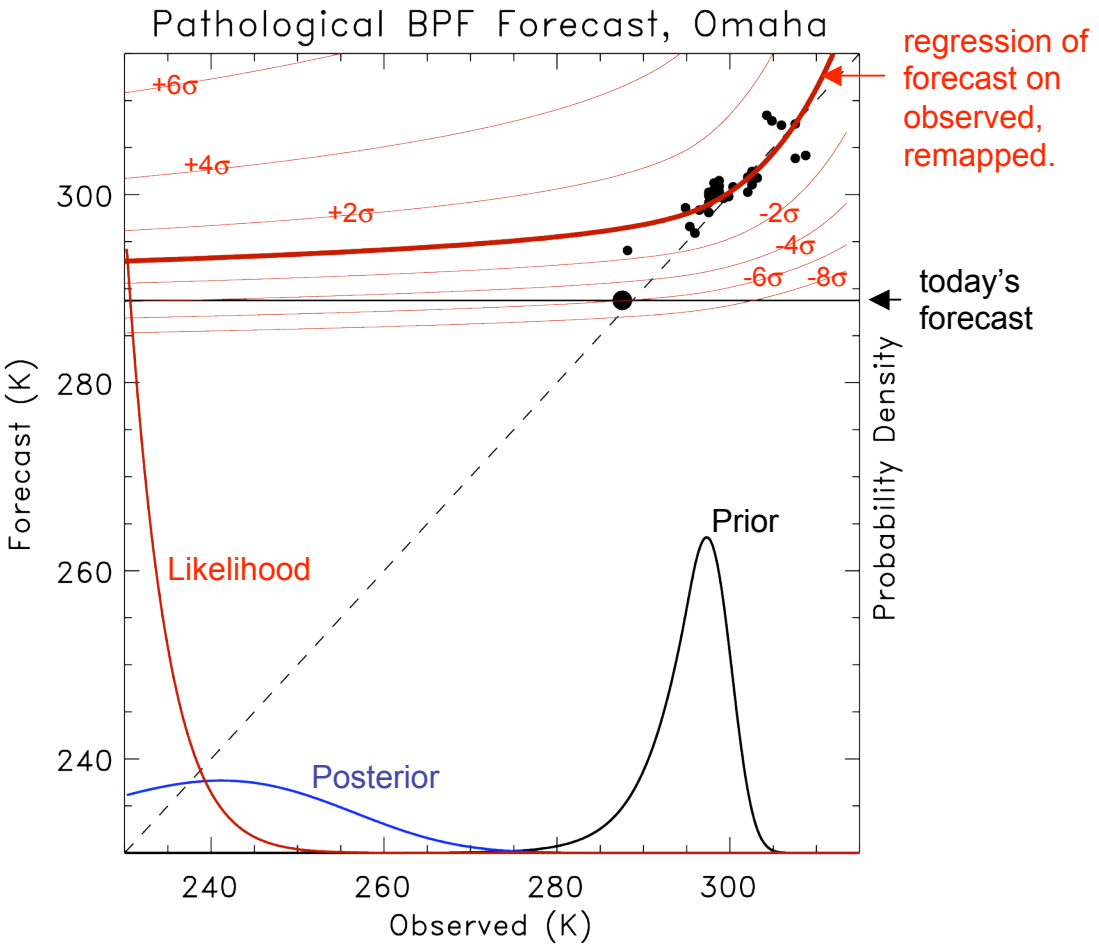


After power transformations and standardization



notice strong warping of data; whereas today's forecast data consistent before transformations to standard normal, inconsistent after.

Plotting power-transformed regression on the original data

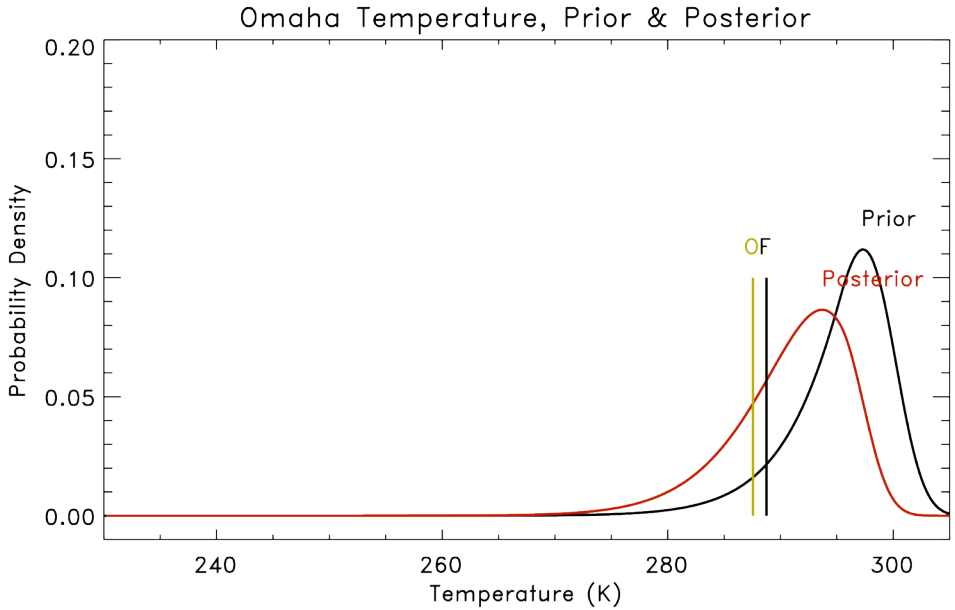


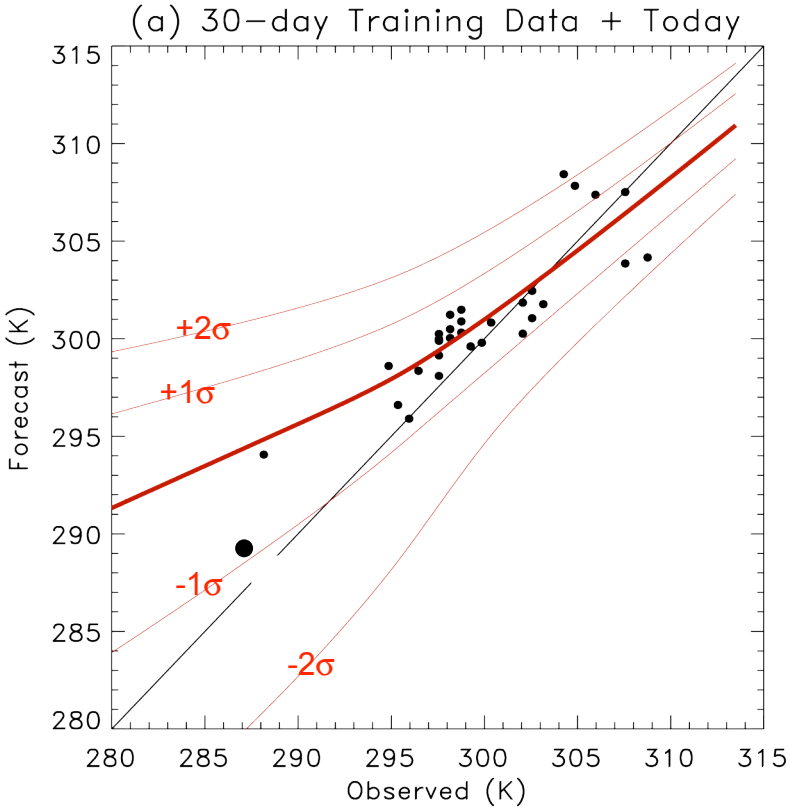
The warping from applying a different power transformation to the forecast relative to the observed (from climatological data) made today's forecast/ observation, which were outliers but relatively consistent before the transformation, into a -6σ forecast outlier.

Possible lessons:
(1) Dangers of fitting non-normal distributions with small training data set.
(2) Dangers of fitting different distribution of forecast relative to observed.

(Though illustrated with power-transformed normals, no reason to suspect that Weibull would be qualitatively any better).

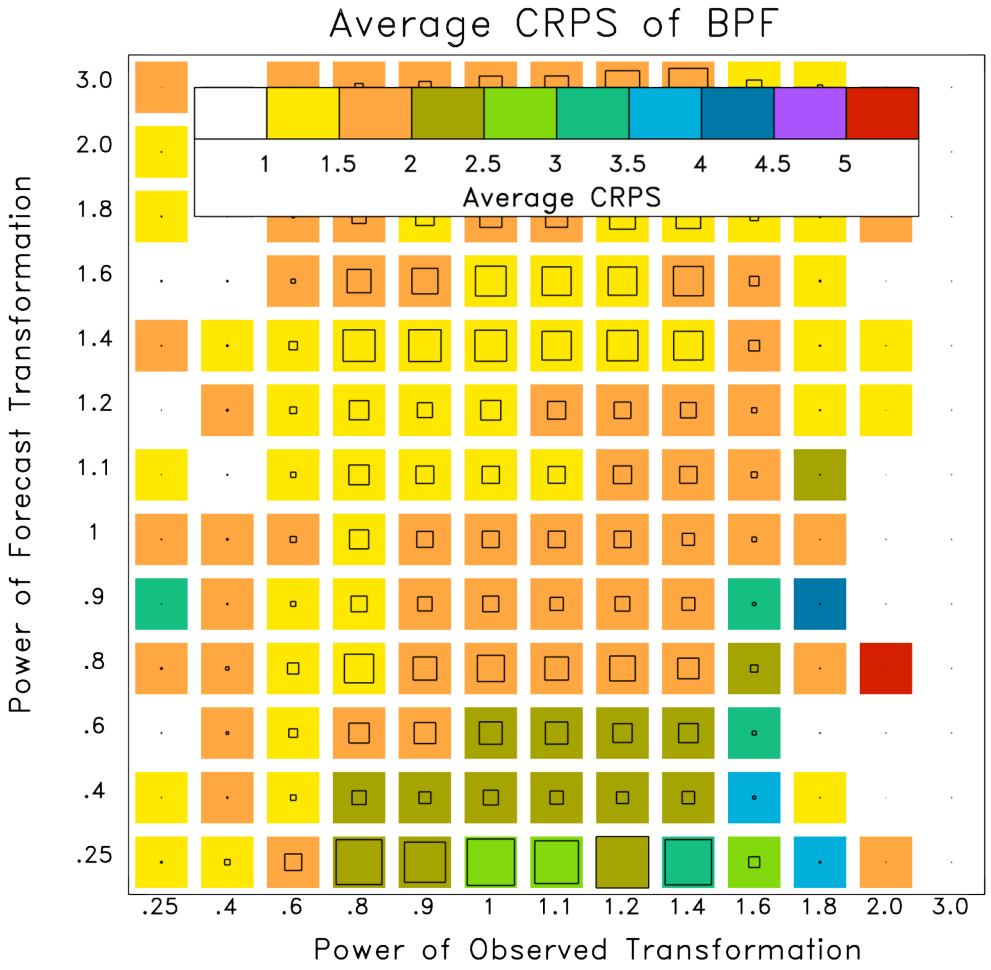
What if we enforce the same 1.6 power transformation on forecast as was used on observed?





perhaps not ideal, but non-pathological now.

How often does the climatology and forecast apply different power transformations? What are the errors?



Colors indicate the magnitude of average forecast error when a given observed/forecast power transformation pair is used.

Black box size indicates the fraction of samples for this transformation pair.

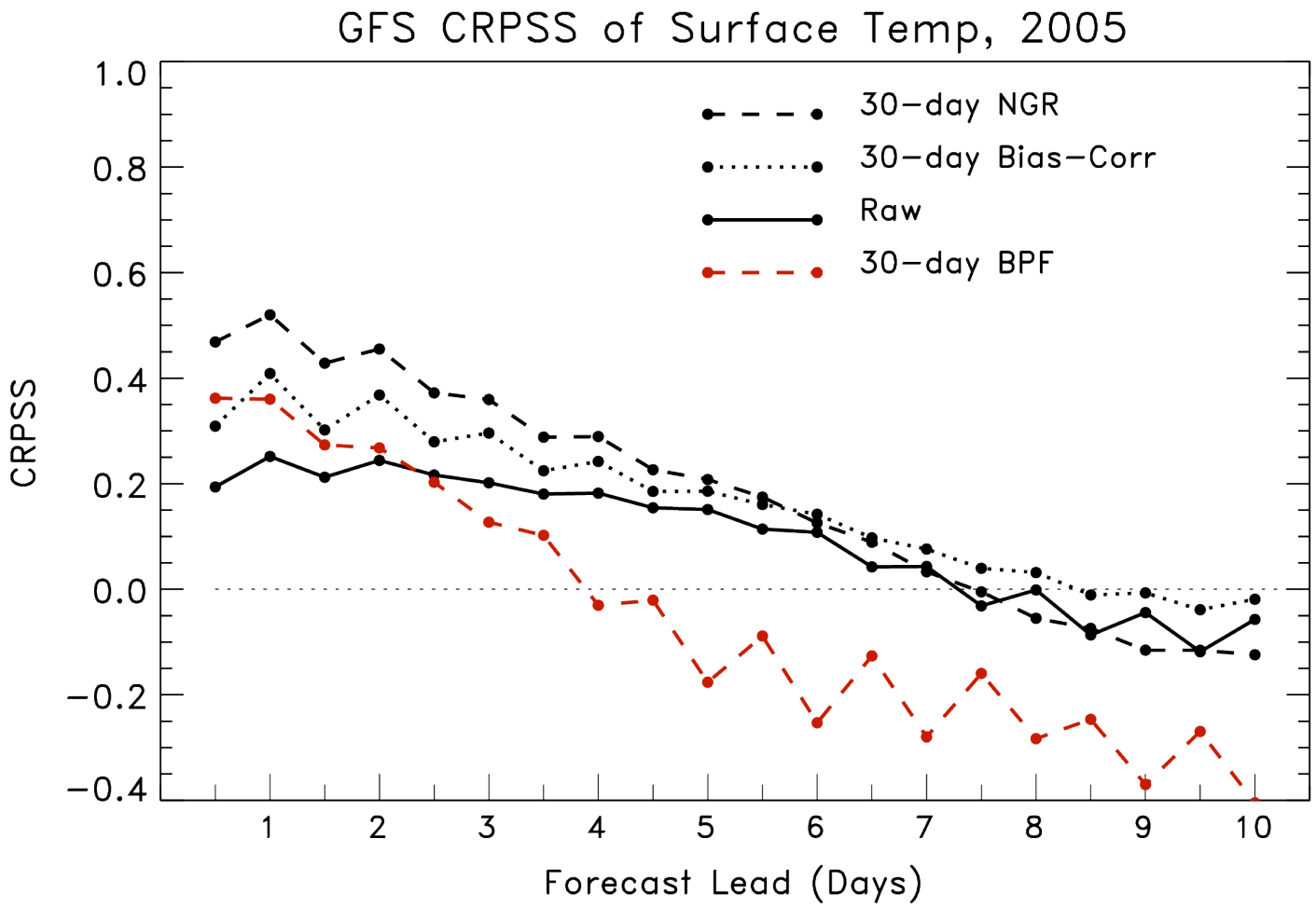
Notes:

(1) Different obs / forecast power transformations are common.

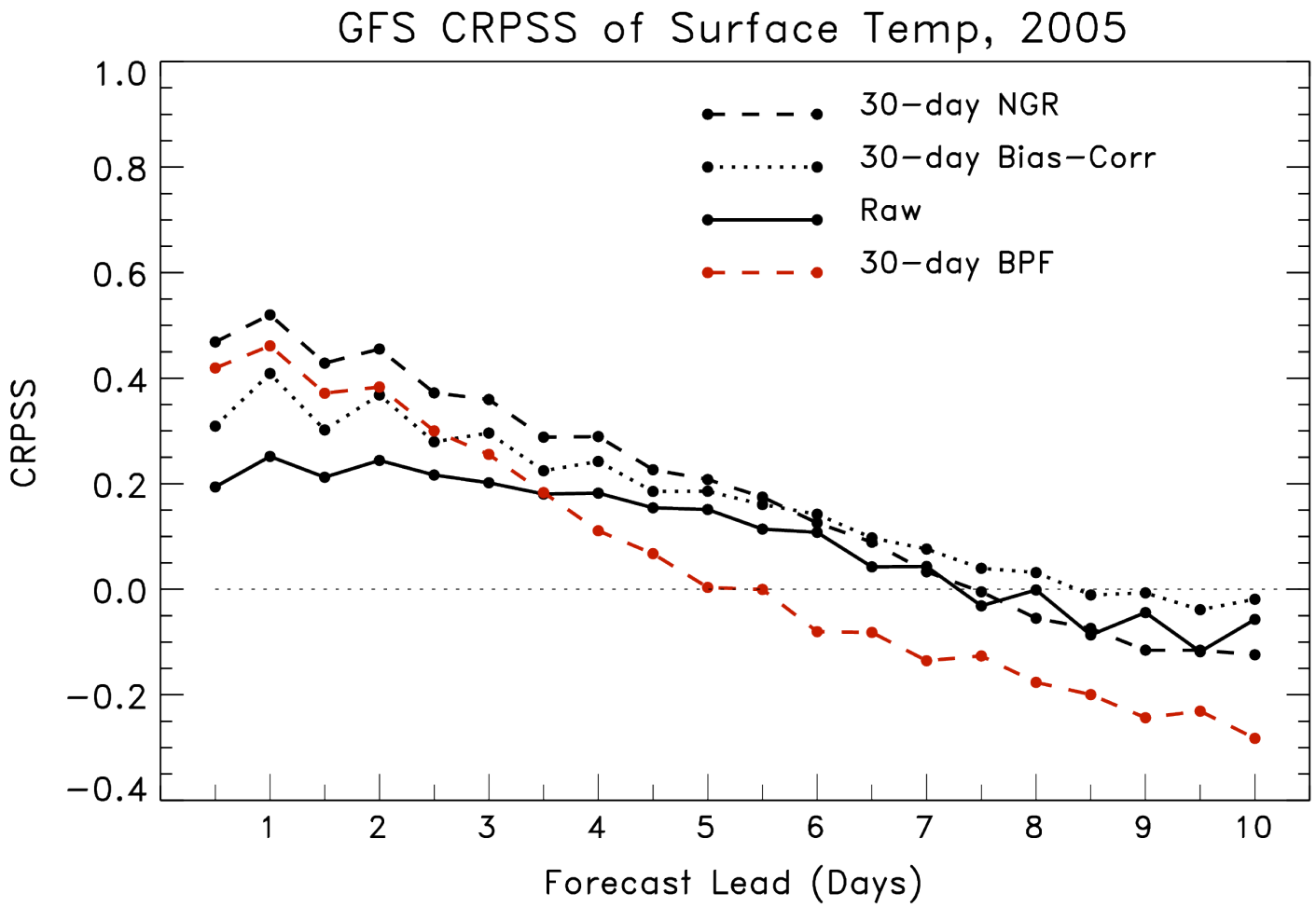
(2) Forecast transformations that are very large/small are common (due to small training sample size?)

(3) Errors larger when obs transform different from forecast transform.

Results, 30-day training data, different observed / forecast transforms

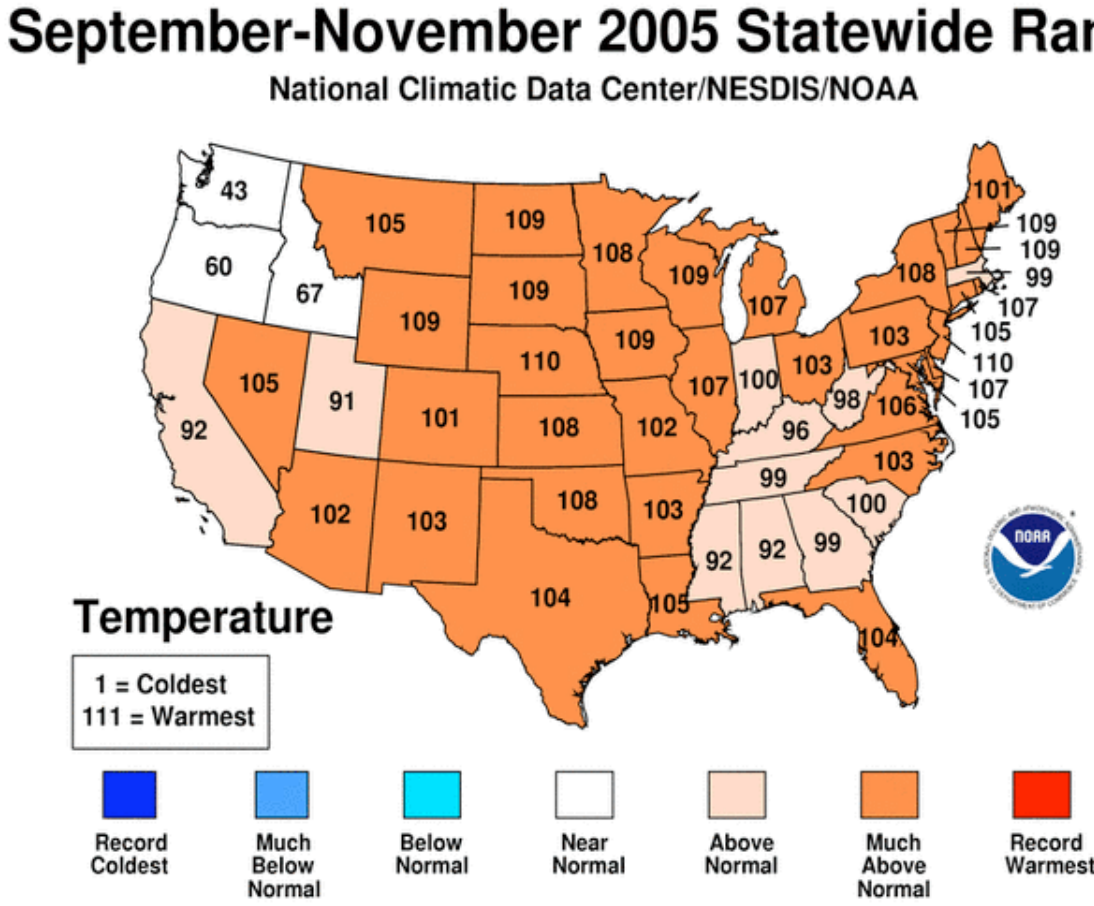


Results, 30-day training data, same observed / forecast transforms



some improvement, but still not competitive with NGR.

Other sources of error in BPF: non-stationary climatology?



Fall 2005 was exceptionally warm, so BPF, modifying a climatological prior from previous colder years (1980-2004), may have consistently underforecast the temperatures.

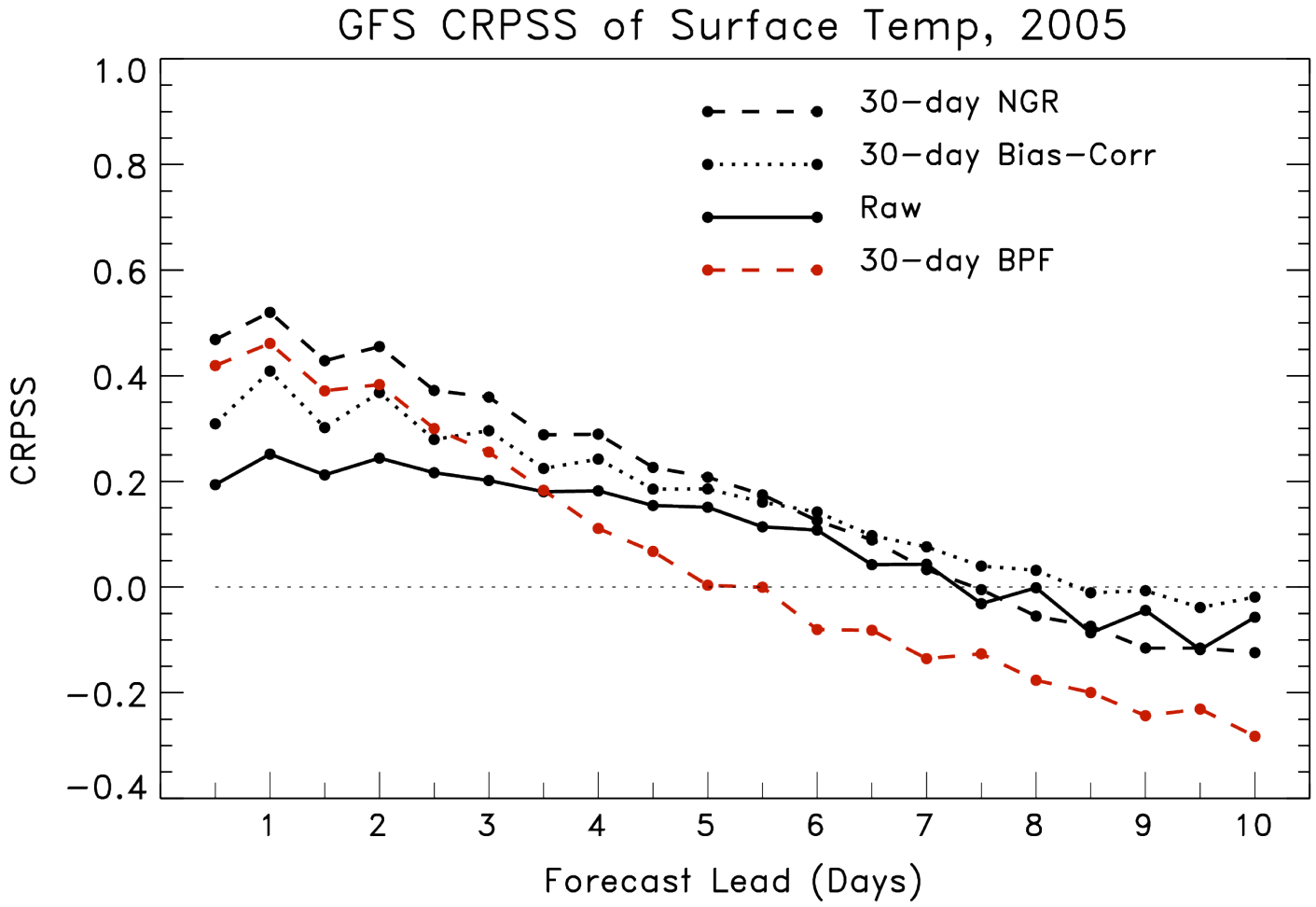
Perhaps climatological prior should include some linear trend?

Source: http://www.ncdc.noaa.gov/oa/climate/research/2005/ann/us-summary.html

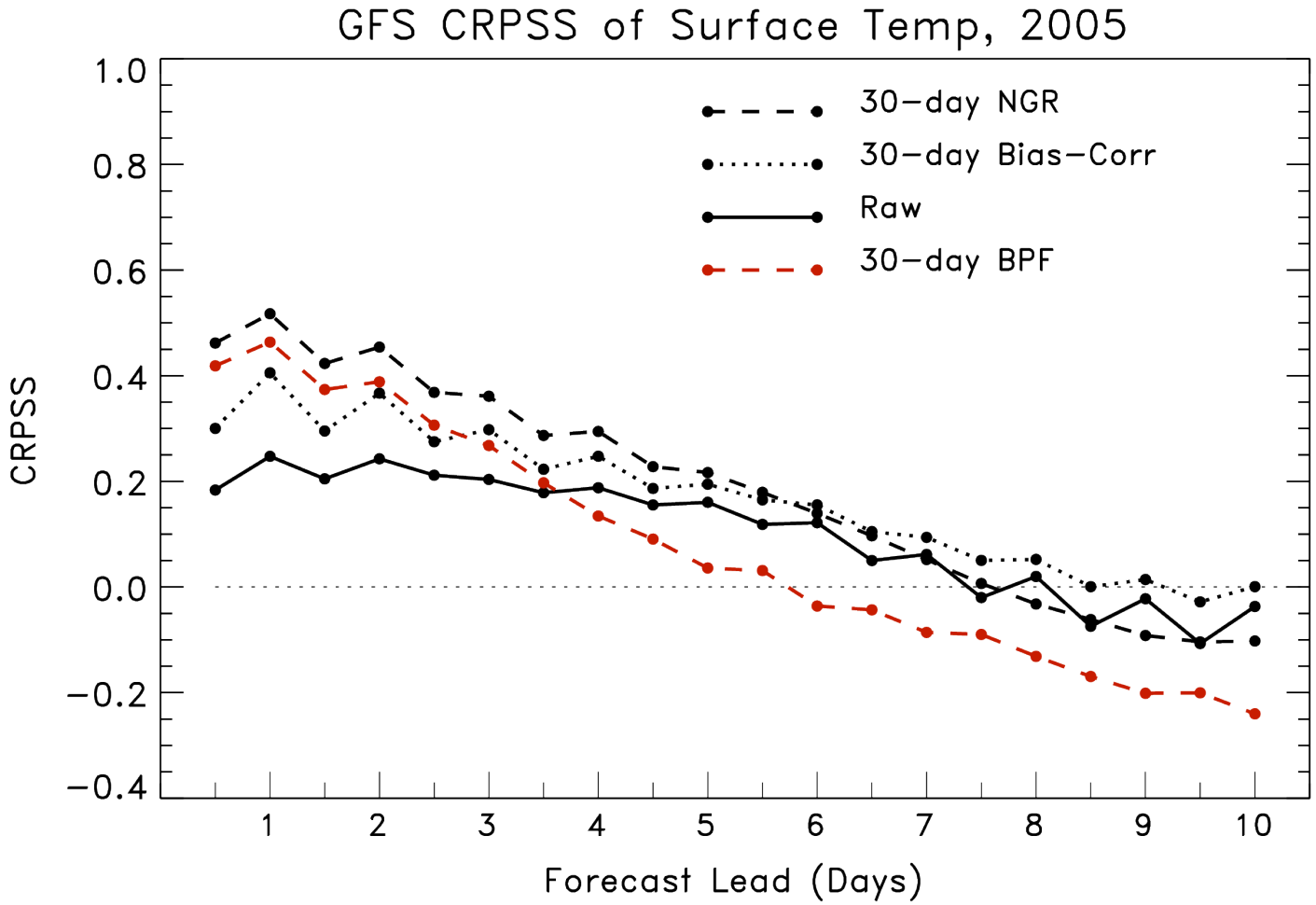
Incorporating changing climate into prior

- Use 1980-2004 temperature trend from regression analysis to change sample values to make them more appropriate for 2005.

Results, 30-day training data, same observed / forecast transforms, no bias correction of climate samples



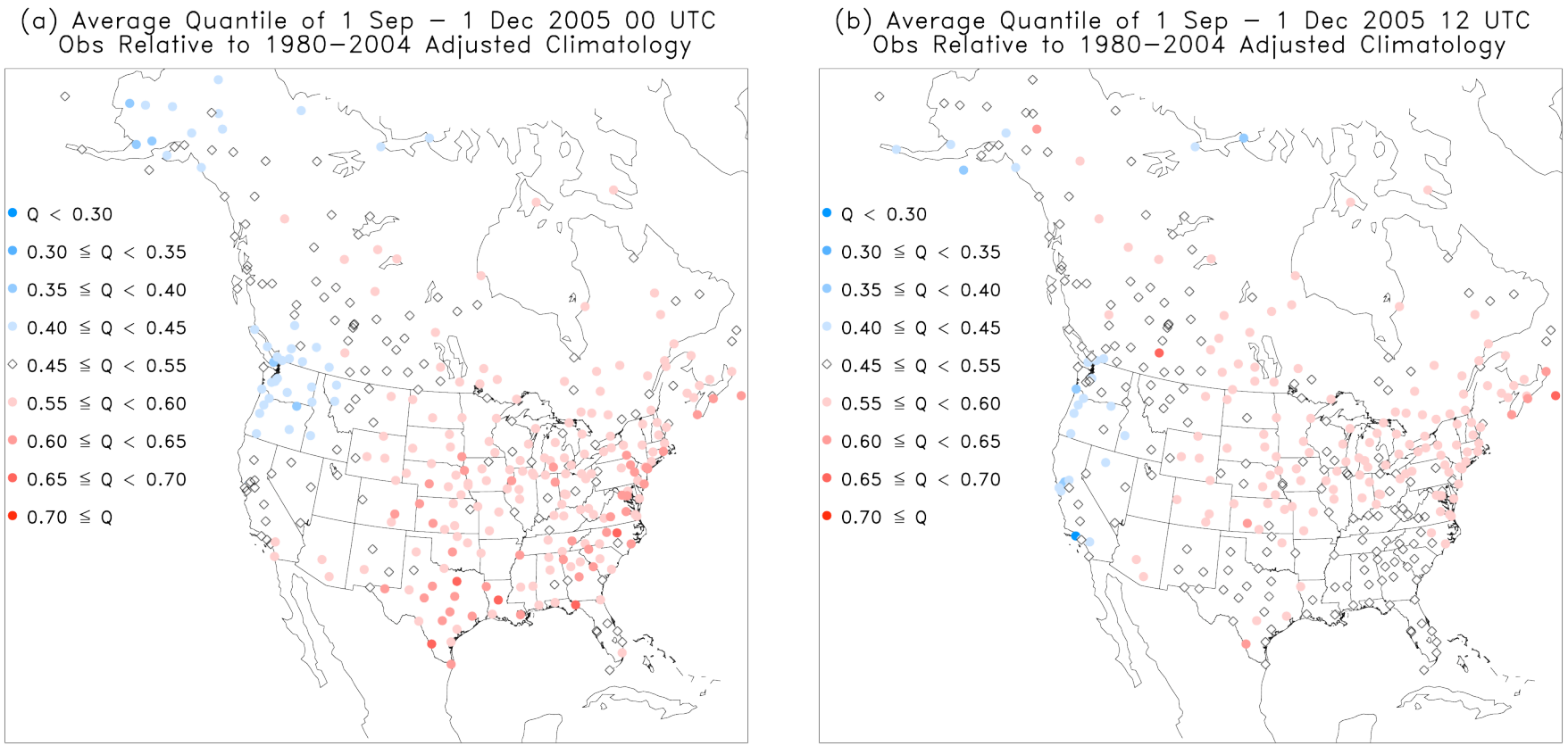
Results, 30-day training data, same observed / forecast transforms, bias correction of climate samples



Slight improvement relative to no bias correction of climatology.

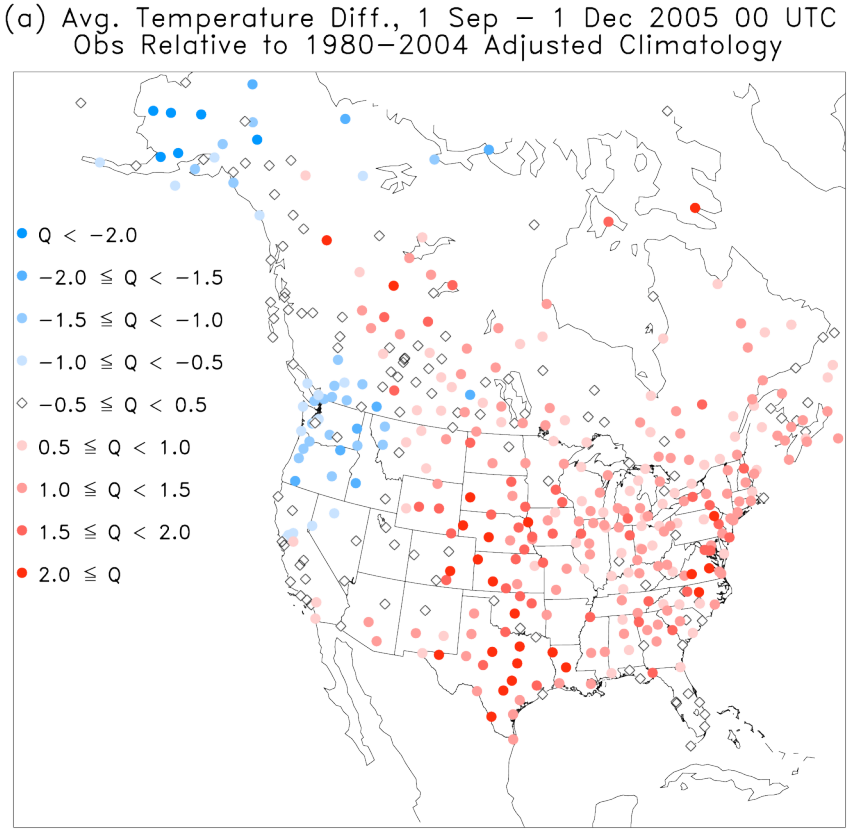
Other skill scores shift, too, since they're calculated relative to slightly changed climatology.

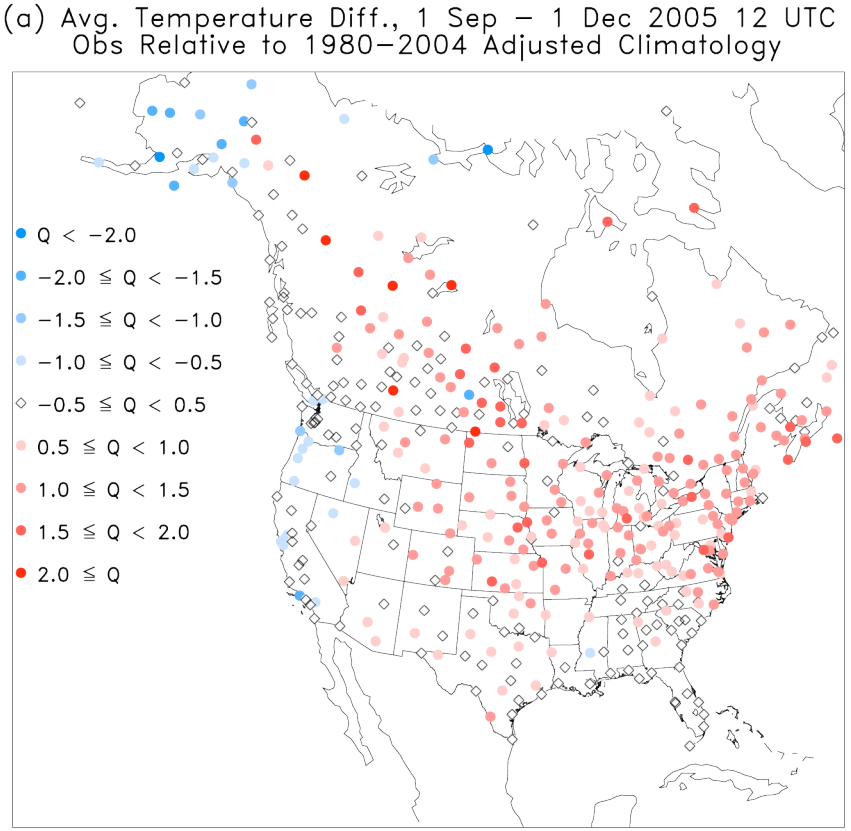
Average quantile of 2005 fall observations relative to bias-corrected climatology



In much of the country, the fall 2005 observed was warmer yet than even the bias-corrected climatology.

Average temperature difference of 2005 fall observations relative to bias-corrected climatology





In much of the country, the fall 2005 observed was warmer yet than even the bias-corrected climatology.