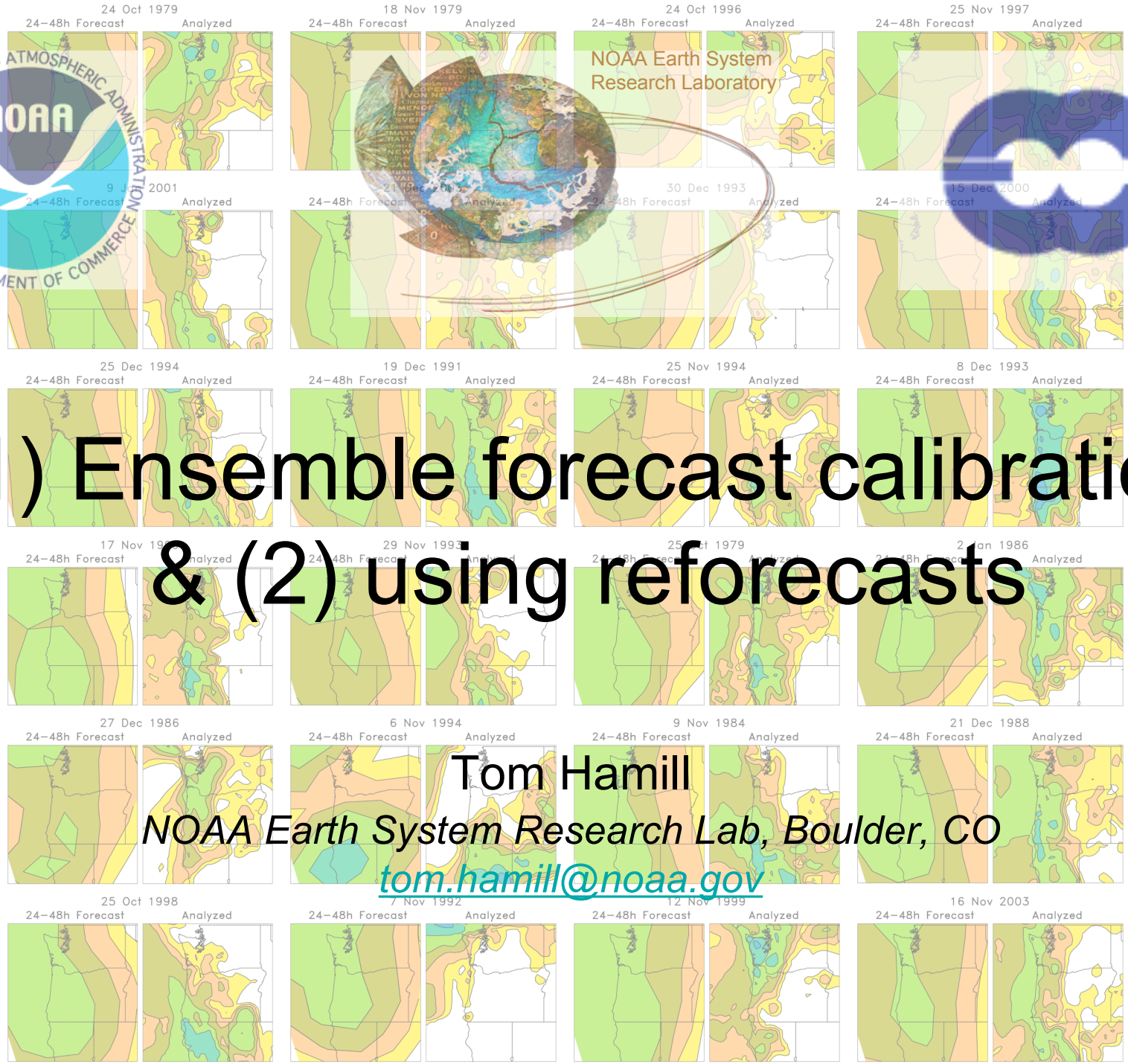


(1) Ensemble forecast calibration & (2) using reforecasts

Tom Hamill

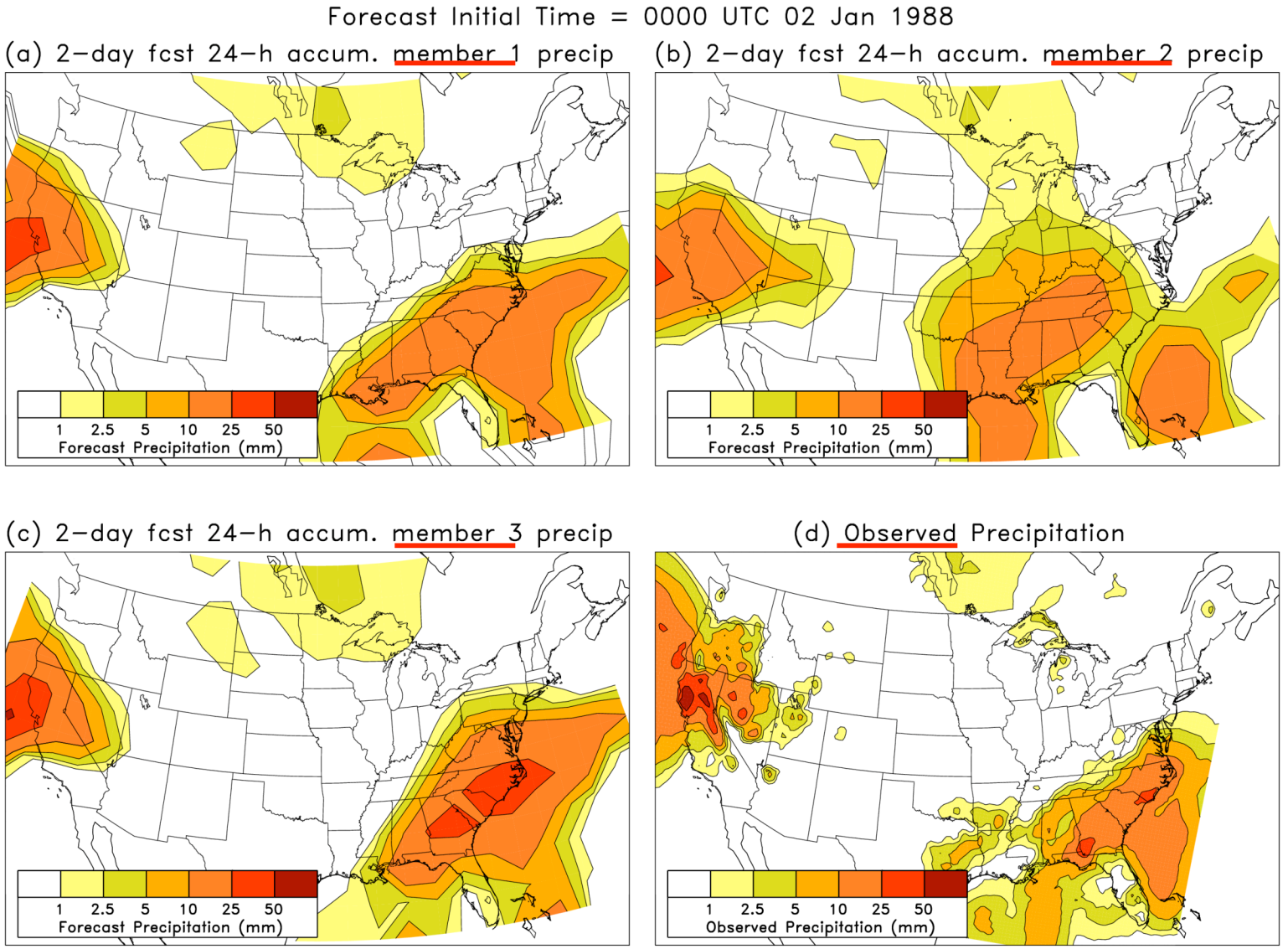
NOAA Earth System Research Lab, Boulder, CO

tom.hamill@noaa.gov

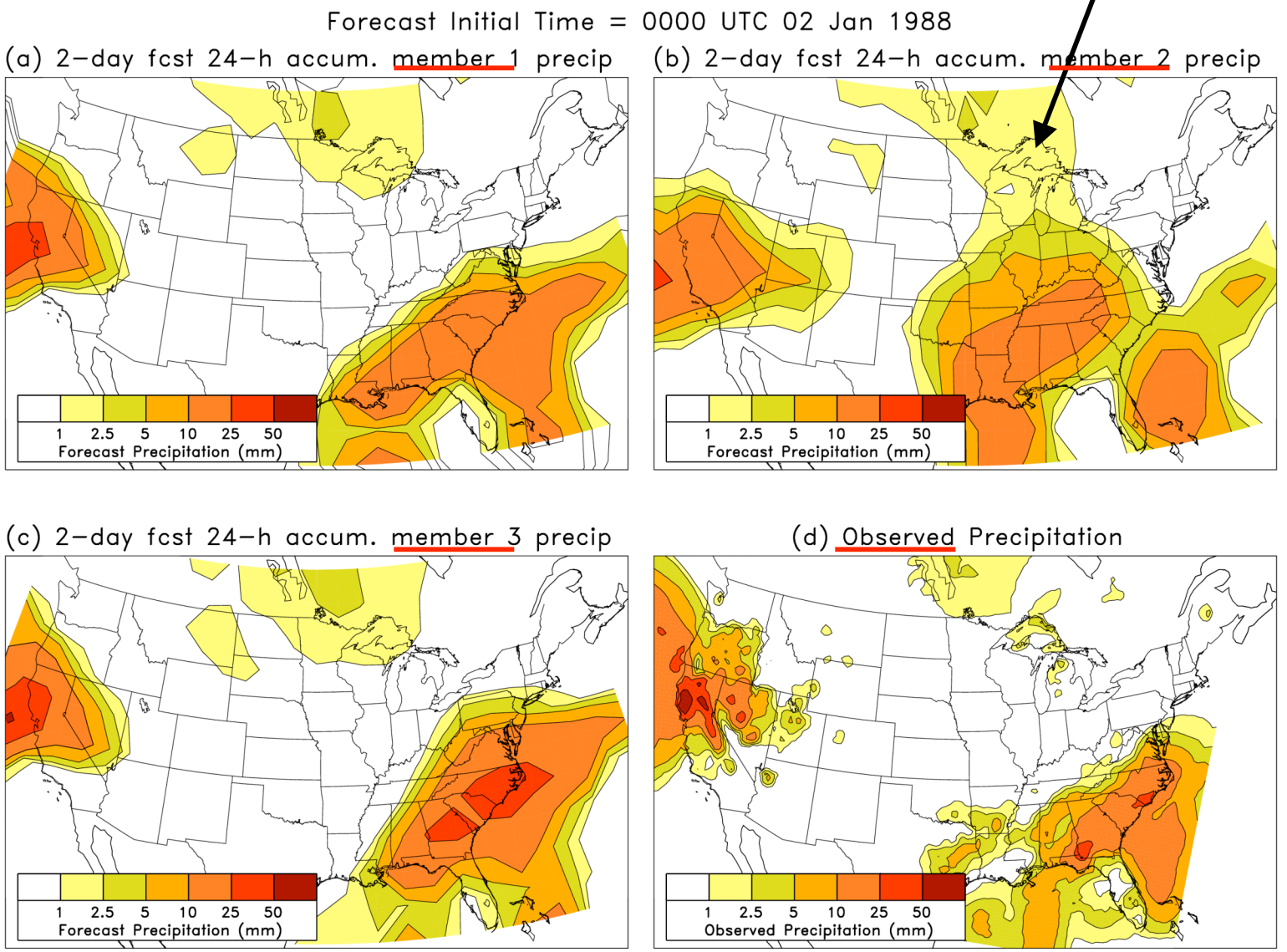
Definition

- **Calibration**: $f\left(\mathbf{x}^t \middle| \mathbf{x}^f\right)$; the statistical adjustment of the (ensemble) forecast
  - Rationale 1: Infer large-sample probabilities from small ensemble.
  - Rationale 2: Remove bias, increase forecast reliability while preserving as much sharpness as possible. Guided by discrepancies between past observations and forecasts.

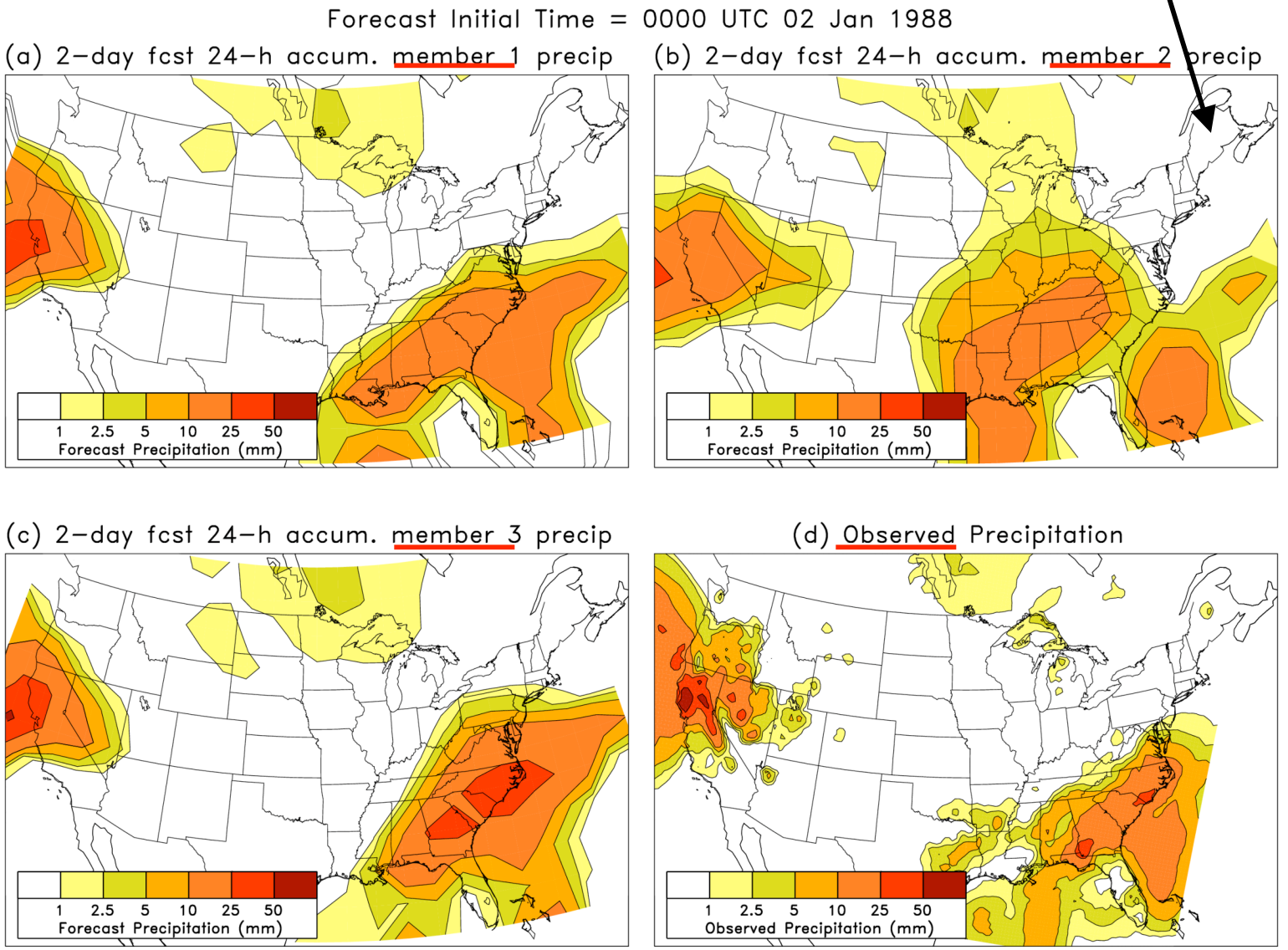
Ensemble-base probabilistic forecasts: problems we'd like to correct through calibration



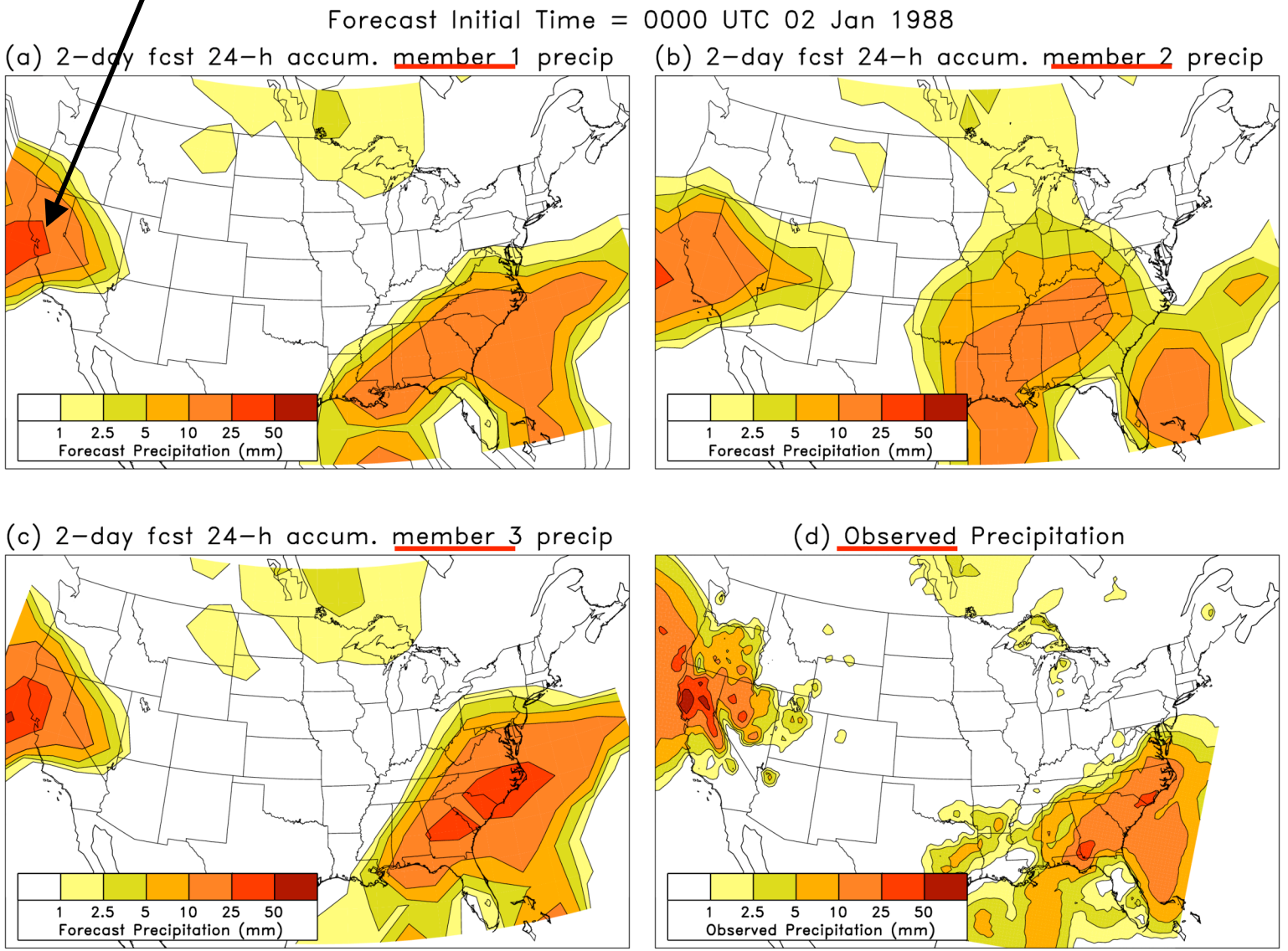
(1) bias (drizzle over-forecast)



(2) ensemble members too similar to each other.



(3) Ensembles are too smooth, not capturing intense local precipitation due to orographic forcing. *Downscaling* needed.



Calibration questions

- Is there a best technique, or best for this particular forecast problem? Different techniques may be needed for:
  - Errors are ~normally distributed, ~stationary, vs.
  - Distributions with long tails

- How much training data (past forecasts & observations) do you have / need?
  - More needed to do good job with rare events.
  - Lots more work involved in trying to get a good result with a short training data set.

Disadvantages to calibration

- Calibration won't correct the underlying problem. Prefer to achieve unbiased, reliable forecasts by doing numerical modeling correctly in the first place.
- No one general approach that works best for all applications.
- Corrections may be model-specific; the calibrations for NCEP v 2.0 may not be useful for ECMWF, or even NCEP v 3.0.
- Could constrain model development. Calibration ideally based on long database of prior forecasts (reforecasts, or hindcasts) from same model. Upgrading model good for improving raw forecasts, may be bad for skill of post-processed forecasts.
- Users beware: Several calibration techniques that have been recently proposed are conceptually flawed / only work properly in certain circumstances.

Calibration review

- Adjusting for sample size, no model-error correction
- Simple methods
  - Gross bias correction
  - Linear regression
  - Kalman filters
- More complex methods
  - Logistic regression
  - Rank histogram-based calibration
  - Dressing
  - Bayesian model averaging
  - CDF corrections
  - Non-homogeneous Gaussian regression

Inferring large-sample pdf from small ensemble: fitting parametric distributions

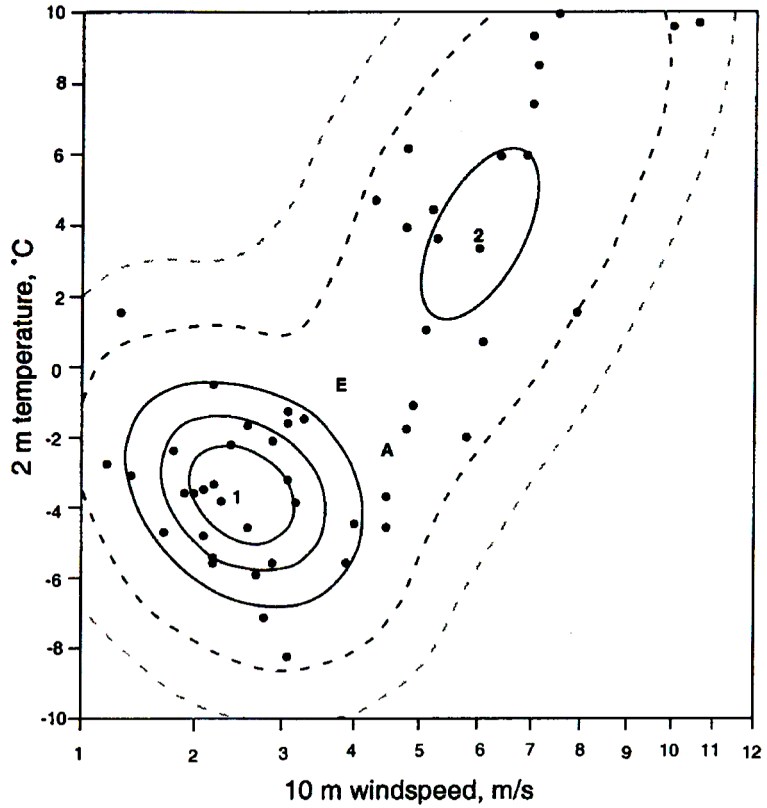
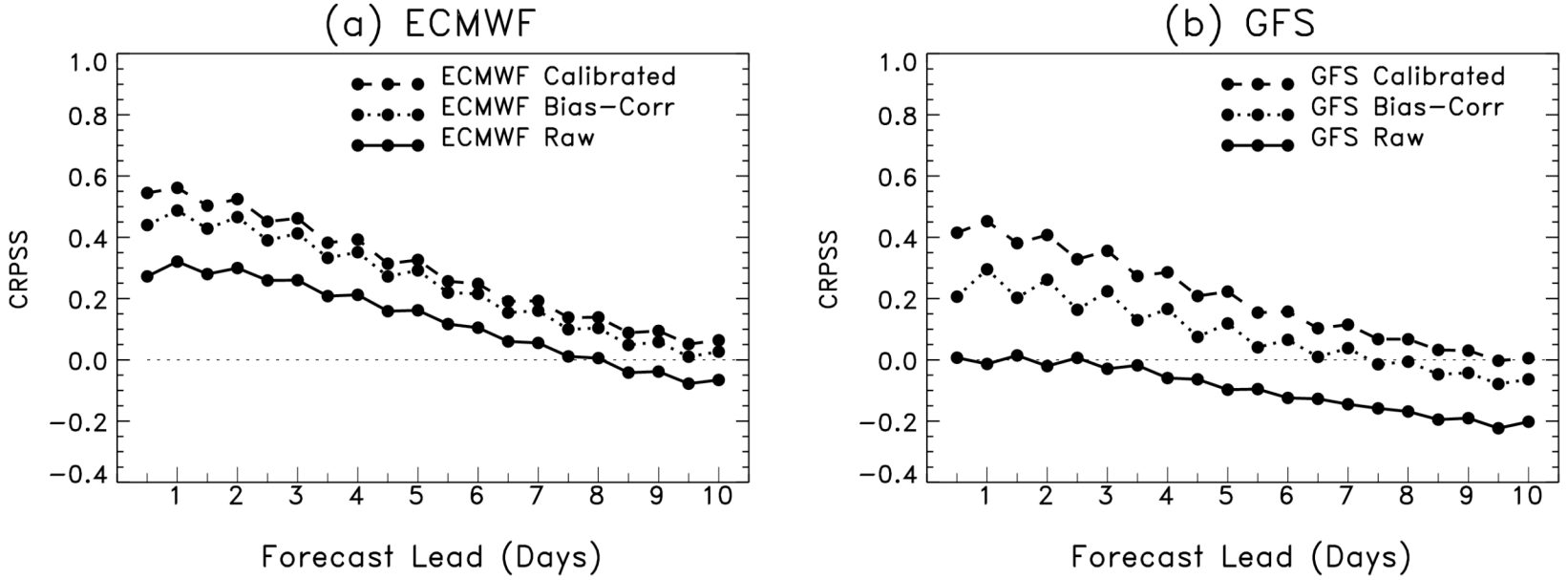


Figure 2. Example ensemble distribution with fitted Gaussian mixture, jointly for the temperature and wind-speed forecast at 12 UTC 10 January 1997 at Manchester, made at the 180 h lead time. Dots indicate individual forecasts made by the 51 ensemble members, with the ensemble mean located at 'E'. The two bivariate Gaussian densities $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are centred at '1' and '2', respectively, and the smooth lines indicate level curves of their mixture $f_{\text{mix}}(\mathbf{x})$, formed with $\alpha = 0.57$ (see text). Contour interval is 0.05, and the thick and thin dashed lines are for 0.01 and 0.001, respectively. Subsequent verifying analysis is 'A'.

Wilks (*QJRMS*, **128**, p 2821) explored fitting parametric distributions, or mixtures thereof, to ECMWF forecasts in perfect-model context. Power-transformed non - Gaussian variables prior to fitting. Goal was smooth pdfs, not bias/spread corrections.

Gross bias correction

- Given sample of past forecasts $x_1, \ldots, x_n$ and observations $y_1, \ldots, y_n$, gross bias correction is simply $\overline{y} - \overline{x}$

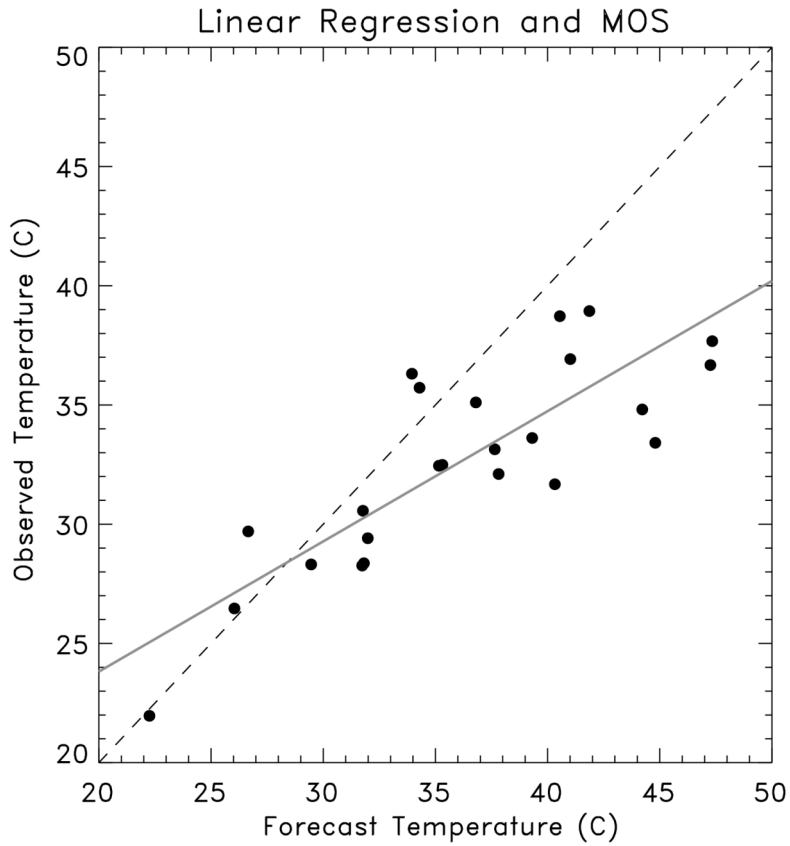


In surface-temperature calibration experiments with NCEP's GFS and ECMWF, simple gross bias correction achieved a large percentage of the improvement that was achieved through more sophisticated, bias+spread correction.

Ref: Hagedorn et al., MWR, 2008, in press.

Linear regression

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$



Corrects for bias; when no skill, regresses to sample climatology.

Diagnostics include statistics on error, so can infer pdf.

Multiple linear regression, with multiple predictors, often used.

Ref: any applied statistics textbook

Model Output Statistics (“MOS”)

many elements based on multiple linear regression

```
KBID   GFS MOS GUIDANCE    2/16/2005  1800 UTC
 DT /FEB  17                   /FEB  18                 /FEB  19
 HR   00 03 06 09 12 15 18 21 00 03 06 09 12 15 18 21 00 03 06 12 18
 N/X              32          40          25          35       19
 TMP  42 39 36 33 32 36 38 37 35 33 30 28 27 30 32 31 28 25 23 19 27
 DPT  34 29 26 22 19 18 17 17 17 17 17 15 14 13 11  8  7  6  5  2  4
 CLD  OV FW CL CL SC BK BK BK BK BK BK BK SC BK BK BK BK FW CL CL CL
 WDR  26 30 32 32 32 31 29 28 30 32 31 31 31 31 30 29 31 32 33 33 27
 WSP  12 12 12 11 08 08 09 08 09 09 10 10 10 12 13 13 15 16 15 09 08
 P06        17     0     0     0     4     0    10     6     8  0  0
 P12              17           0          10          17        8
 Q06         0     0     0     0     0     0     0     0     0  0  0
 Q12              0           0          0           0        0
 T06       0/ 2  0/ 0  1/ 0  1/ 2  0/ 1  0/ 1  1/ 0  0/ 1  0/ 0  0/ 0
 T12                 1/ 0        1/ 2        1/ 1        0/ 1  0/ 0
 POZ   0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 POS  13 47 70 84 91100 96100100100100 92100 98100100100 94 92100100
 TYP   R  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S  S
 SNW                                     0                    0
 CIG   7  8  8  8  8  8  8  8  8  7  7  7  8  7  7  7  8  8  8  8  8
 VIS   7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7  7
 OBV   N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N  N
```

US: Statistical corrections to operational US NWS models, some fixed (NGM), some not (Eta, GFS). Refs: http://www.nws.noaa.gov/mdl/synop/index.htm, Carter et al., *WAF*, **4**, p 401, Glahn and Lowry, *JAM*, **11**, p 1580. **Canadian** models discussed in Wilson and Vallee, *WAF*, **17**, p. 206, and *WAF*, **18**, p 288. **Britain**: Met Office uses “updateable MOS” much like perfect prog.

Kalman filter

Today's forecast bias estimate

Yesterday's bias estimate

Yesterday's observed bias

$$\hat{b}_t^f = \hat{b}_{t-1}^f + K_t\left(\varepsilon_t - \hat{b}_{t-1}^f\right)$$

Kalman gain: weighting applied to residual

Pro:

- memory in system, amount tunable through K_t
- adaptive

Con:

- takes time to adapt after regime change

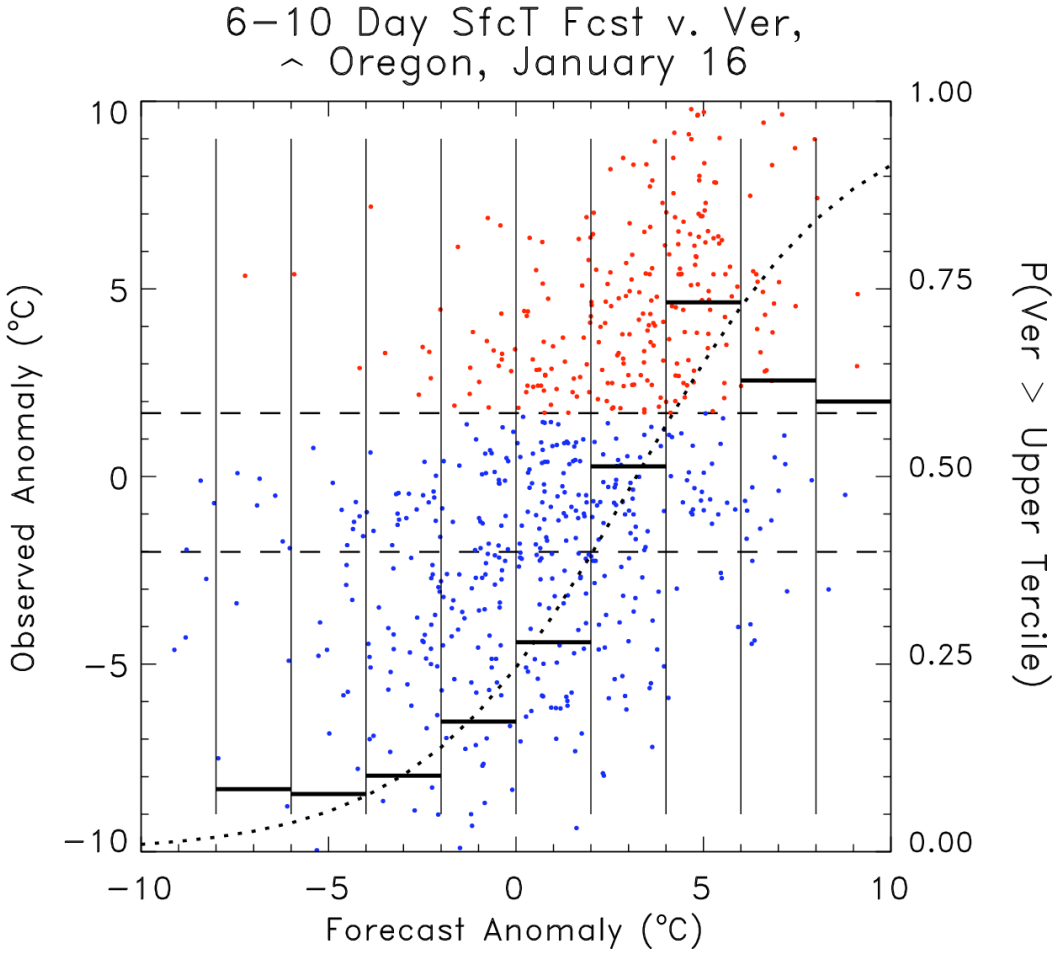
Ref: Cheng and Steenburg, conferences.dri.edu/WxPrediction/Weather12/Cheng_Steenburgh.ppt

Logistic regression

- Define event, for example, Temp > $Q_{0.67}$
- For each grid point (or station) let *x* = continuous predictor data (ens. mean forecast value), *y* = binary predictand data (1.0 if predicted event happened, 0.0 if not).
- Problem: Compute P(*y* =1.0 | *x*) as a continuous function of *x*.
- Logistic Regression: $$P = 1 - \frac{1}{1 + exp\ (\beta_0 + \beta_1 x)}$$

Ref: any applied statistics text.

Logistic regression using a long data set of observed and forecast anomalies

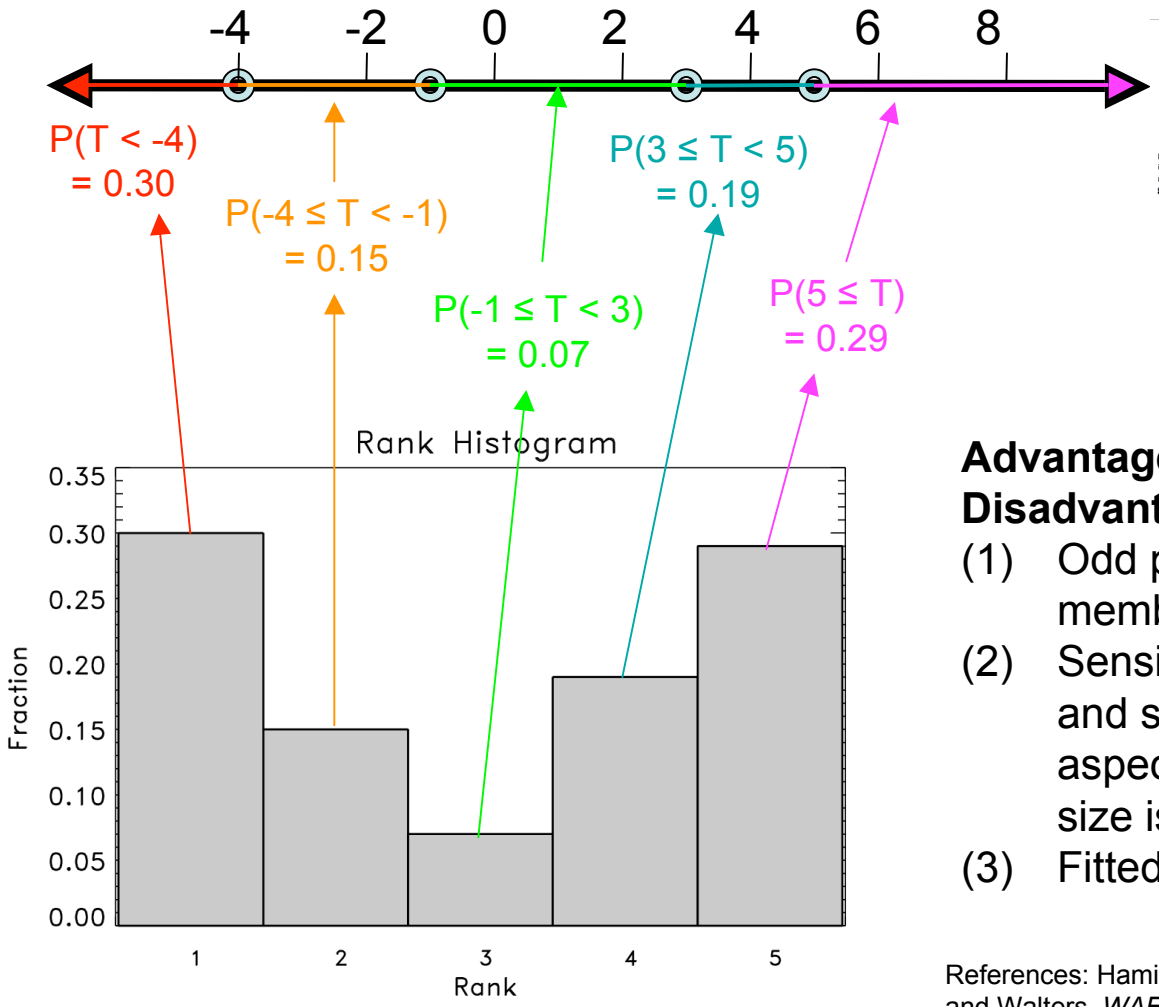


Seeking to predict probability of warmer than normal conditions (upper tercile of observed). Using reforecasts, we have 23 years of data. Let's use old data in a 31-day window around the date of interest to make statistical corrections.

Dashed lines: tercile boundaries
Red points: samples above upper tercile
Blue points: samples below upper tercile
Solid bars: probabilities by bin count
Dotted line: logistic regression curve

Ref: Hamill et al. MWR, June 2004

Ensemble calibration: rank histogram techniques



NCEP MRF precipitation forecasts, from Eckel and Walters, 1998

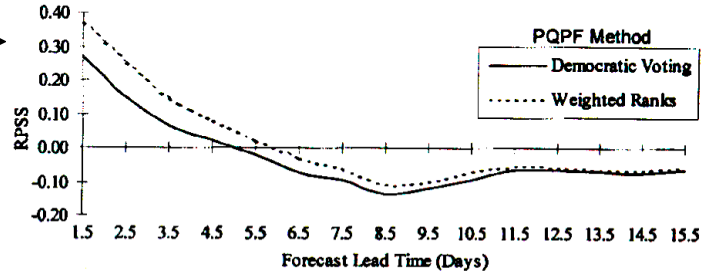


FIG. 10. Ranked probability skill score (RPSS) results for all forecast lead times.

Advantages: Demonstrated skill gain

Disadvantages:

(1) Odd pdfs, especially when two ensemble members close in value.

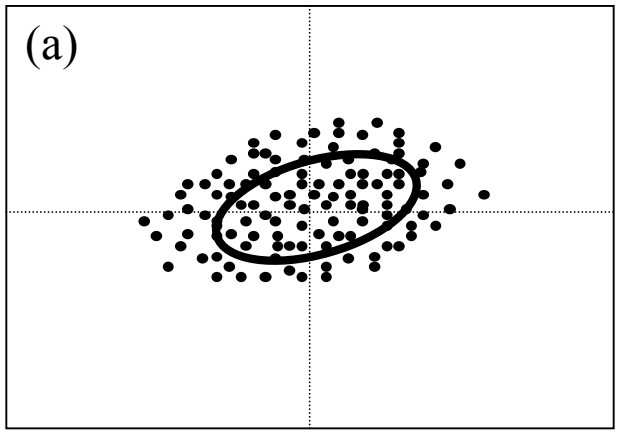
(2) Sensitive to shape of rank histogram, and shape of histogram may vary with aspects like precip amount --> sample size issues.

(3) Fitted parametric distributions as skillful

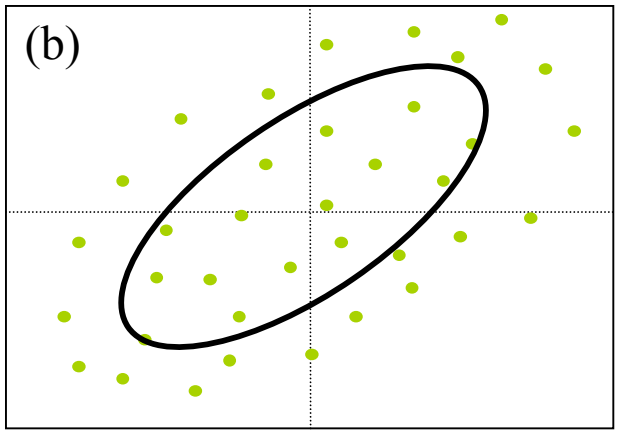
References: Hamill and Colucci (*MWR*, 1997, 1998; Eckel and Walters, *WAF*, 1998; used at Met Office)

Dressing methods

Original Ensemble



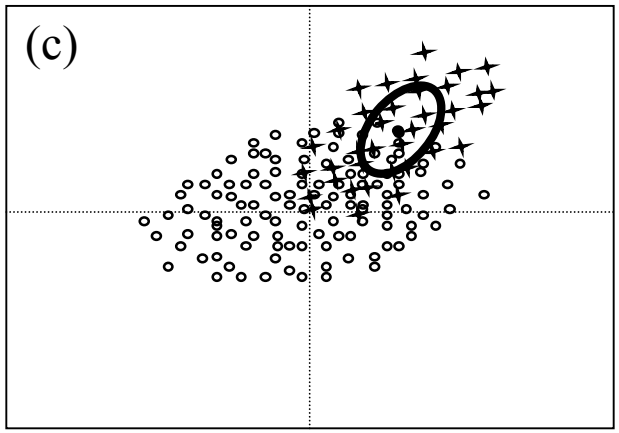
Cov(ens mean errors)



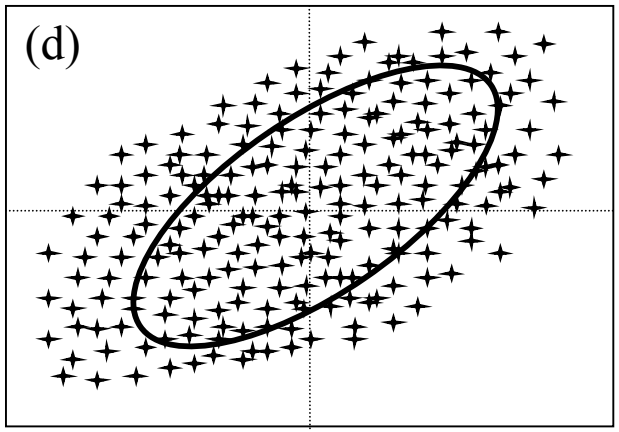
Method of correcting spread problems. Assume prior bias correction.

Adv: Demonstrated improvement in ETKF ensemble forecasts in NCAR model.

Dressing Samples



Dressed Ensemble



Dis: Only works if too little spread, not too much.

Ref: Roulston and Smith (Tellus, 55A, p 16); Wang and Bishop (QJRMS, 2005; picture above)

Bayesian model averaging (BMA)

$$p(y \mid f_1, \ldots, f_K) = \sum_{k=1}^{K} w_k \; g_k(y \mid f_k)$$

Weighted sum of kernels centered around individual, bias-corrected forecasts.

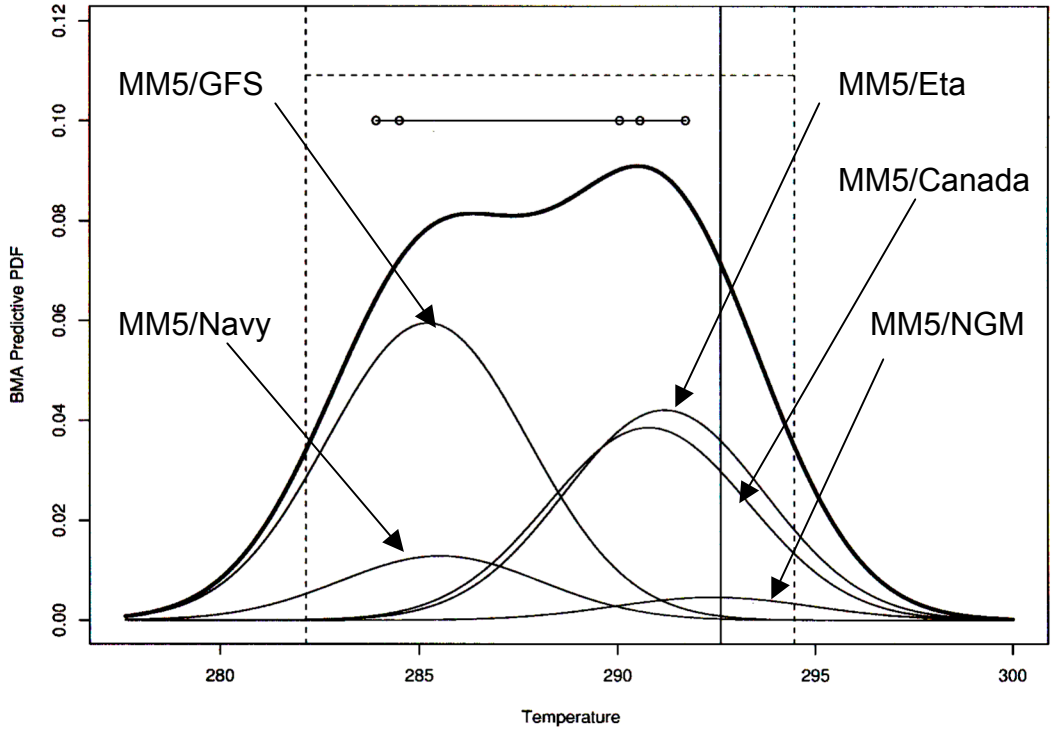


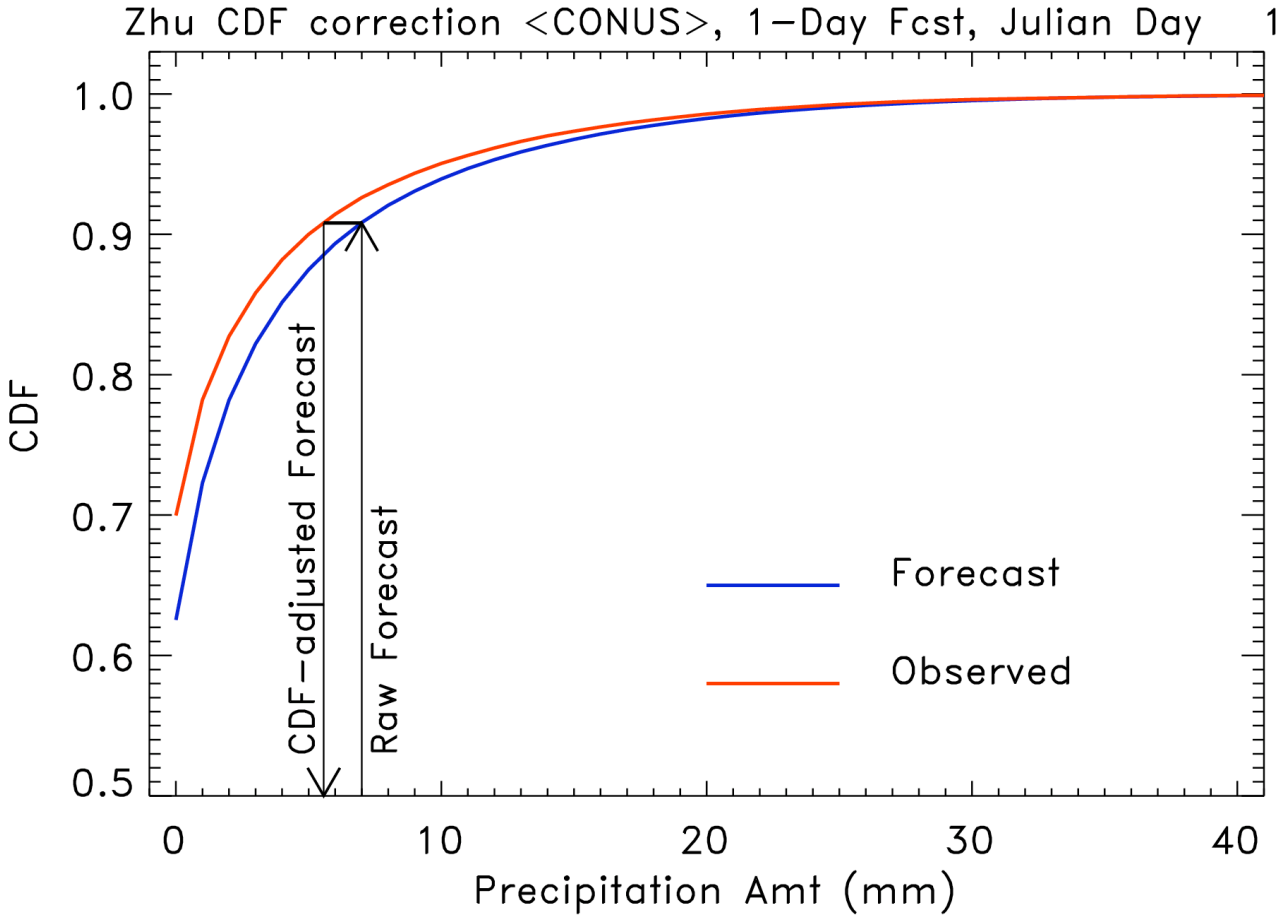
Figure 3: BMA predictive PDF (thick curve) and its five components (thin curves) for the 48-hour surface temperature forecast at Packwood, Wash., initialized at 0000 UTC on June 12, 2000. Also shown are the ensemble member forecasts and range (solid horizontal line and bullets), the BMA 90% prediction interval (dotted lines), and the verifying observation (solid vertical line).

Advantages: Theoretically appealing. No parameterized distribution assumed, weights applied proportional to their independent information (in concept).

Disadvantages: When trained with small sample, BMA radically de-weighted some members due to "overfitting" See Hamill, *MWR*, Dec. 2007.

Ref: Raftery et al., MWR, 2005. Wilson et al., MWR, 2007

Another problematic method: CDF-based corrections



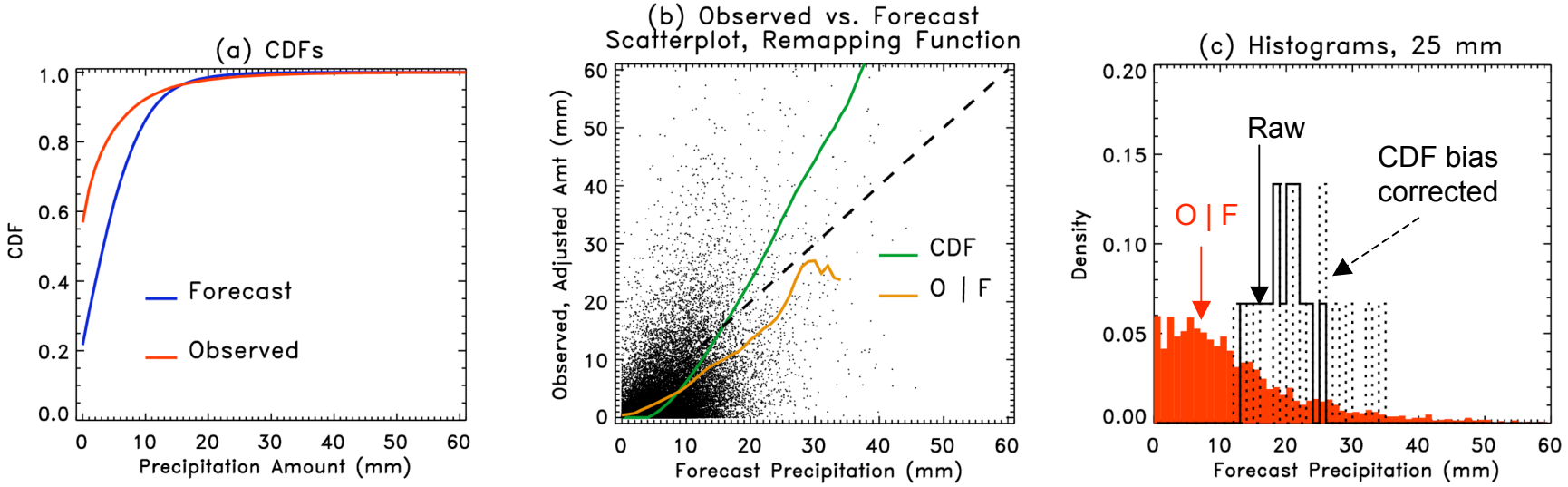
Use difference in CDFs to correct each ensemble member's forecast. In example shown, raw 7-mm forecast corrected to ~5.6 mm forecast.

NOTE: bias only, not spread correction or downscaling.

Ref: Zhu and Toth, 2005 AMS Annual Conf., and many others

CDF corrections: example of problem

1-day forecasts in Northern Mississippi (US), mid-August.
Consider a forecast precipitation of 25 mm.



CDF-based corrections at high amounts suggest further increasing precipitation amount forecast. O|F indicates decrease.

At root of problem is assumption that Corr (F,O) ≅1.0

Non-homogeneous Gaussian regression

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

Is there a “best” calibration technique?

Using Lorenz ‘96 toy model, direct model output (DMO), rank histogram technique, MOS applied to each member, dressing, logistic regression, non-homogeneous Gaussian regression (NGR), “forecast assimilation”, and Bayesian model averaging (with perturbed members assigned equal weights) were compared. Comparisons generally favored logistic regression and NGR, though differences were not dramatic, and results may not generalize to other forecast problems such as ones with non-Gaussian errors.

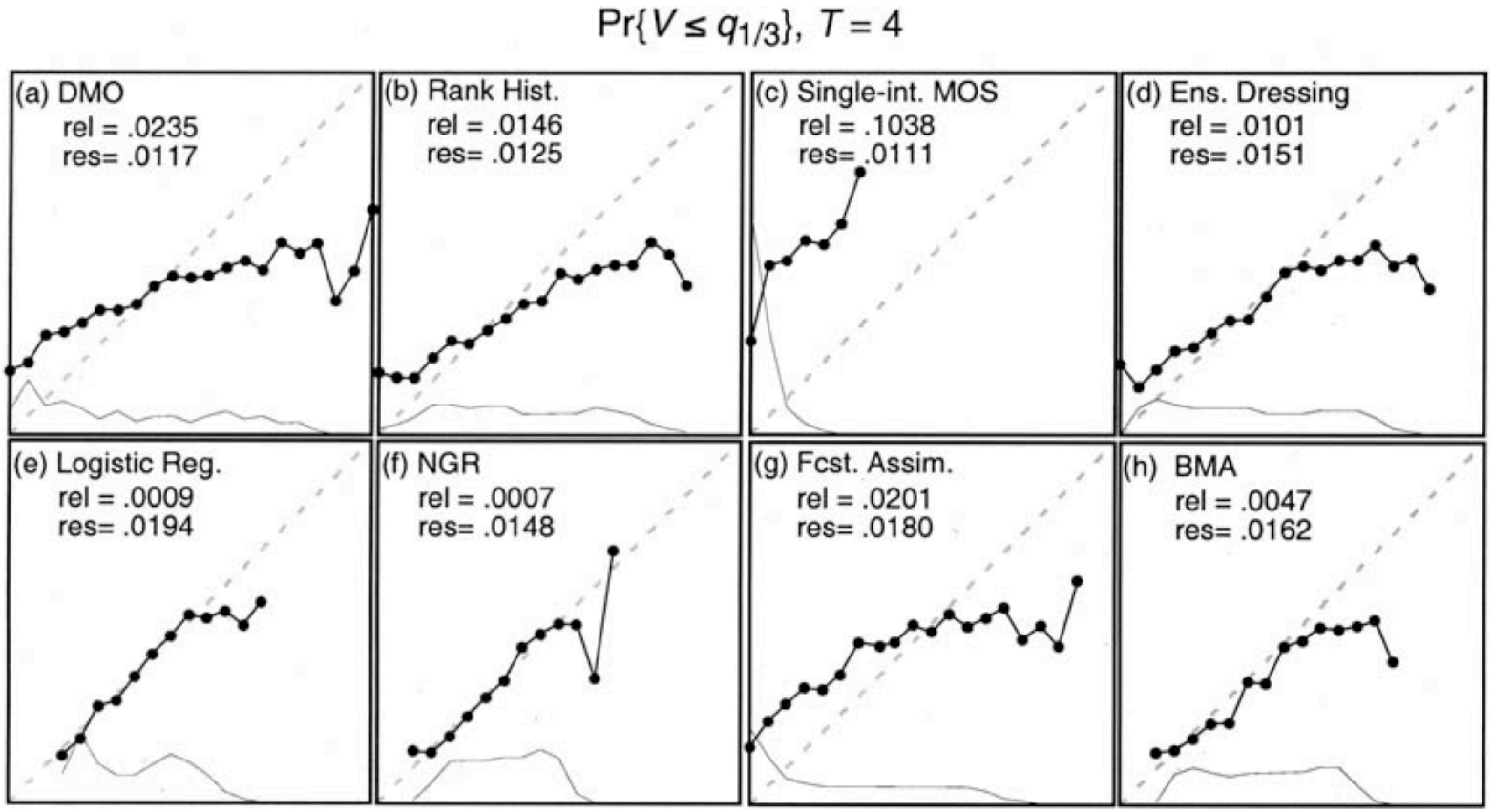


Figure 8. *As Figure 5, for Pr{V ≤ $q_{1/3}$} at lead time T = 4.*

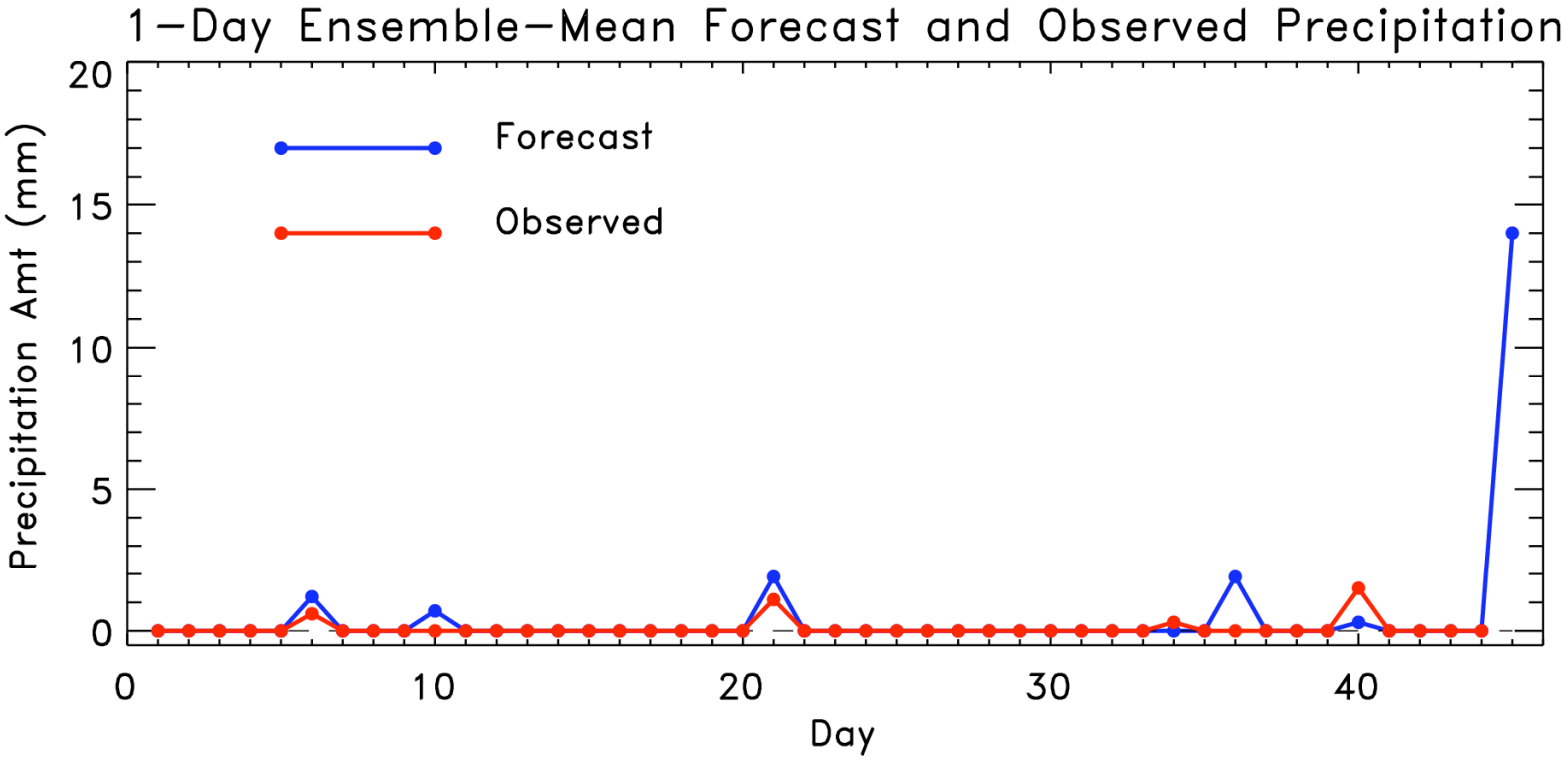
Ref: Wilks, *Met. Apps*, 2006, **13**, p. 243

Part II: “Reforecasting”
Q: What is a reforecast?

- A hindcast, a numerical prediction for a date in the past *using the model and data assimilation system that is currently operational*.

Why compute reforecasts?

- For many forecast problems, such as long-lead forecasts or high-precipitation events, a few past forecasts may be insufficient for calibrating the probabilistic forecasts



NOAA's reforecast data set

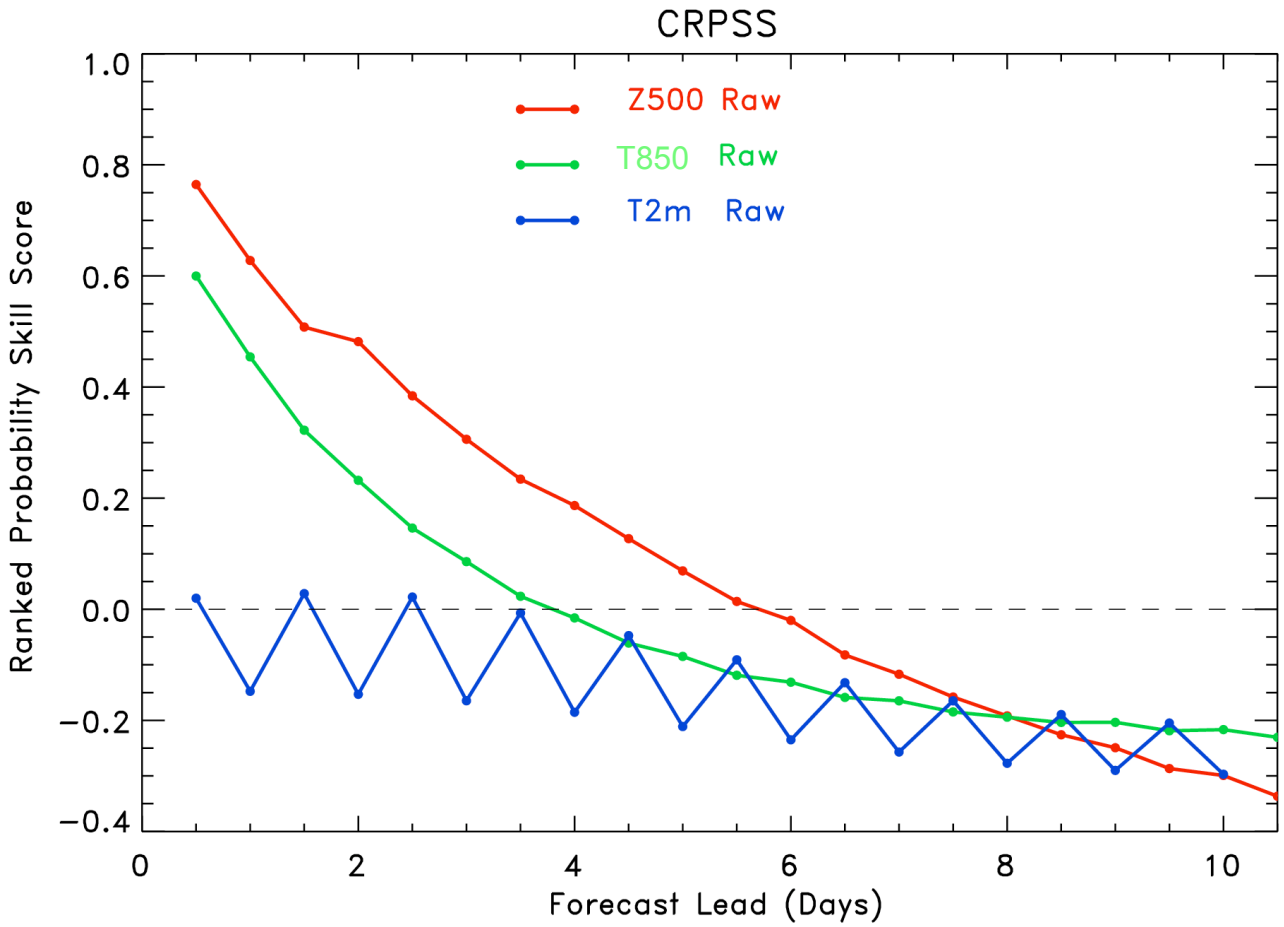
- **Model**: T62L28 NCEP GFS, circa 1998
- **Initial States**: NCEP-NCAR Reanalysis II plus 7 +/- bred modes.
- **Duration**: 15 days runs every day at 00Z from 19781101 to now. (*http://www.cdc.noaa.gov/people/jeffrey.s.whitaker/refcst/week2*).
- **Data**: Selected fields (winds, hgt, temp on 5 press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

Outline

- Part 2a: Several applications of 1998 GFS reforecasts.
  - Comparison of Z500, T850, T2m
  - 6-10 day forecasts over US
  - Downscaled PQPF in US
  - Monsoon PQPF in India
  - Tornado forecasts

- Part 2b: An exploration of whether reforecasts from a much-improved 2005 ECMWF model provide similar benefits as were achieved for 1998 GFS

Skill of 500-hPa Z, 850-hPa T, and 2-m T from raw 1998 GFS reforecast ensemble

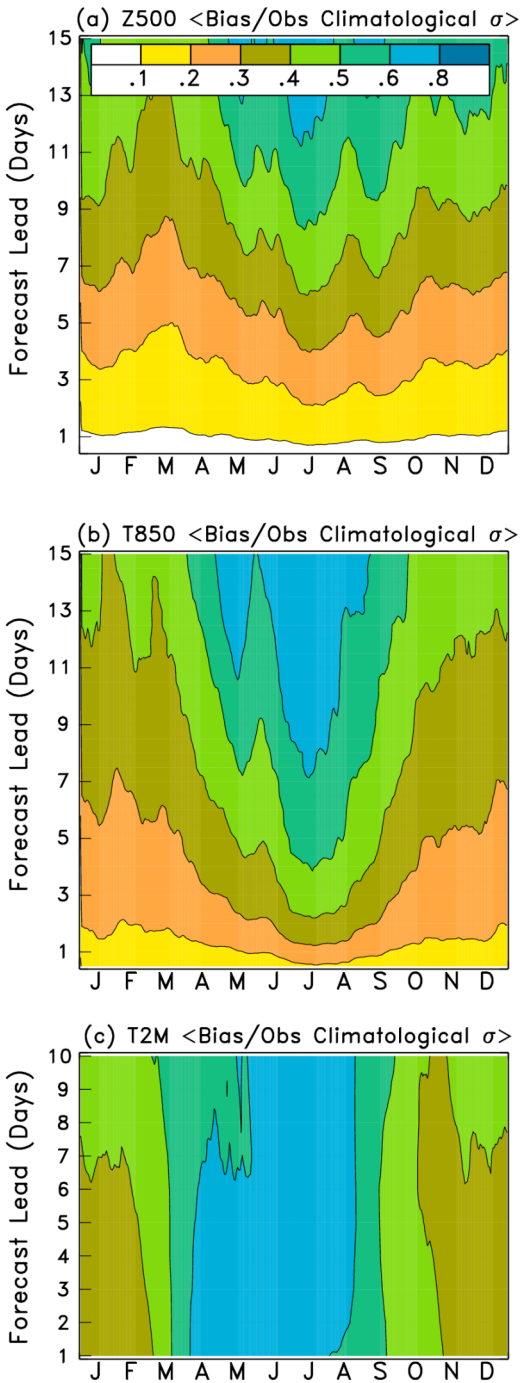


1998 T62 GFS much less accurate than current models, but qualitatively still the same with current models.

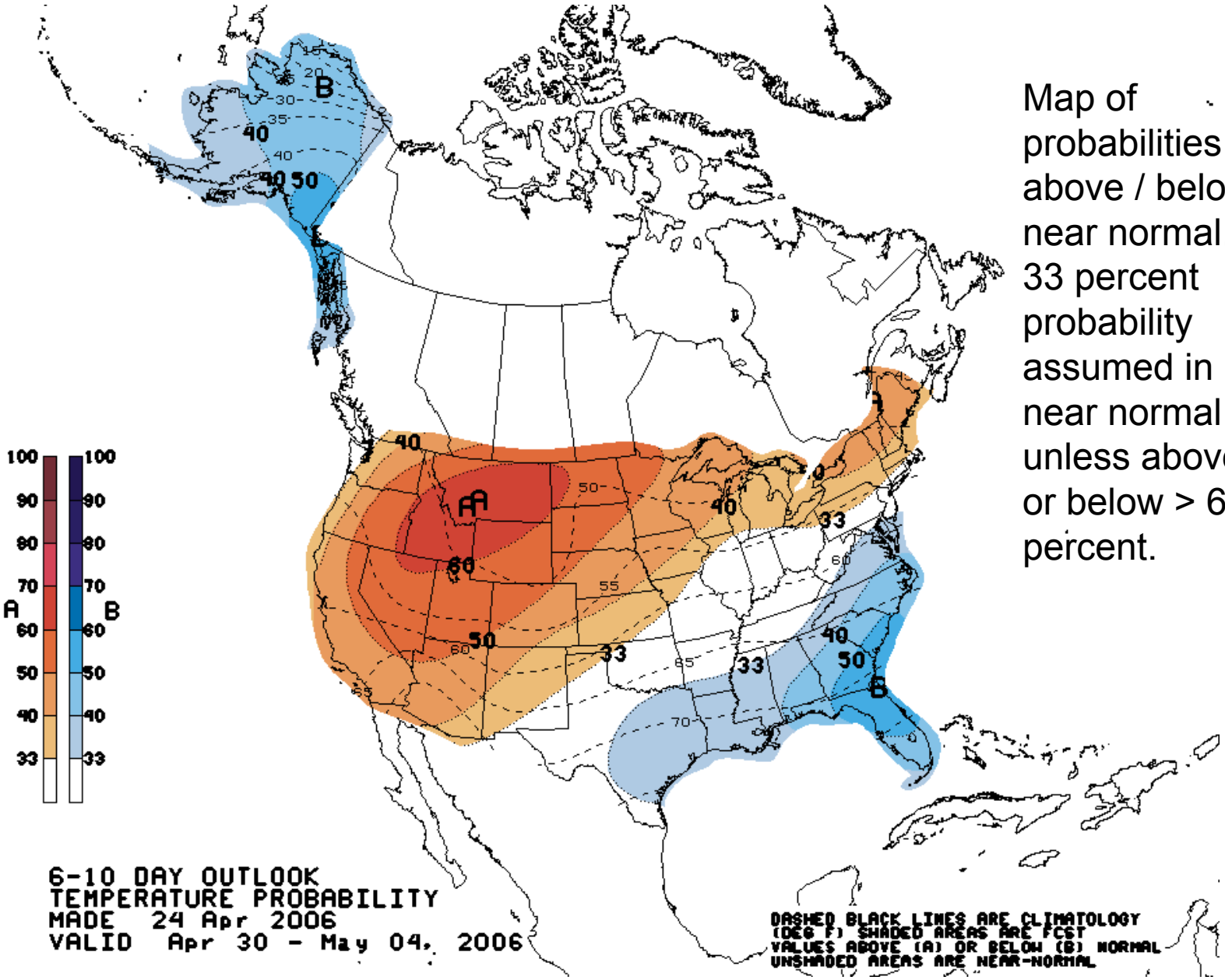
The one we probably care about the most, T_{2m}, scores the worst.

(1979-2004 data)

Forecast bias contaminates T_{2m} much more than Z_{500}

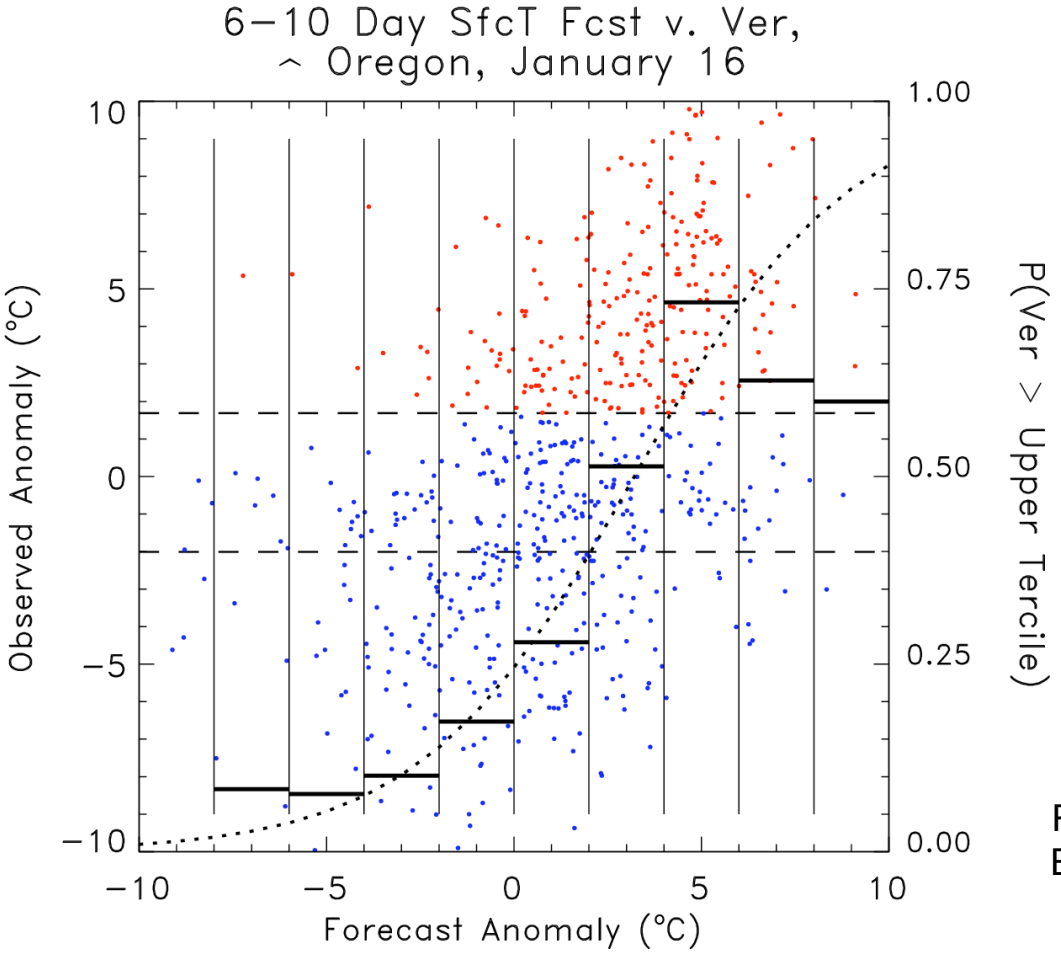


Application: NCEP/CPC's 6-10 day outlook



Map of probabilities of above / below / near normal. 33 percent probability assumed in near normal unless above or below > 67 percent.

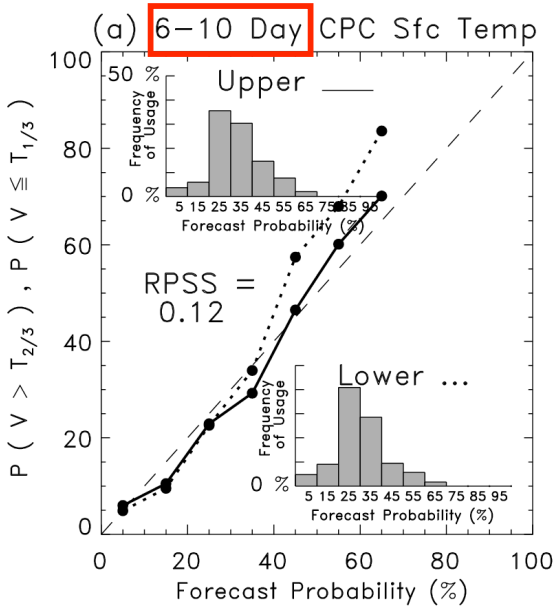
Using a long reforecast data set of observed and forecast anomalies



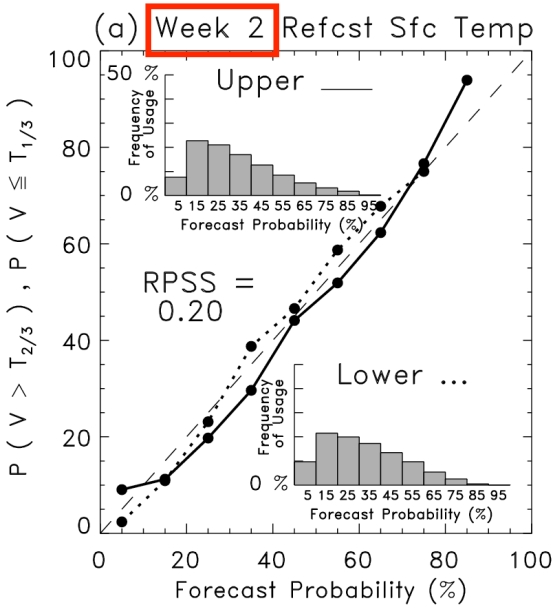
With our reforecasts, we have 25+ years of data. Let's use old data in a 31-day window around the date of interest to make statistical corrections.

Dashed lines: tercile boundaries
Red points: samples above upper tercile
Blue points: samples below upper tercile
Solid bars: probabilities by bin count
Dotted line: a fitted model, TBD

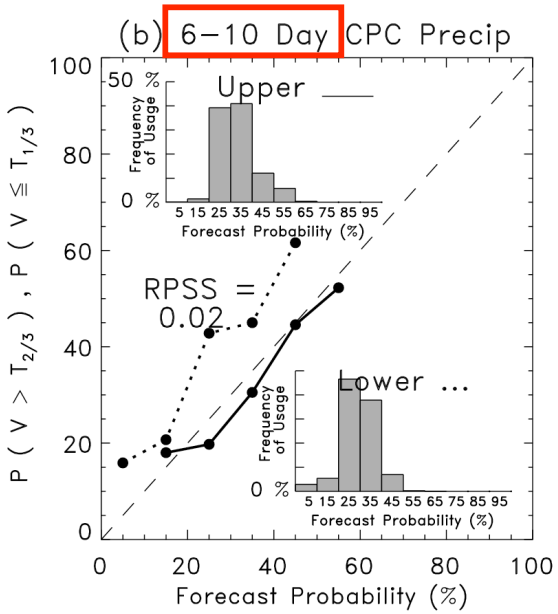
Comparison against NCEP / CPC forecasts at 155 stations, 100 days in winter 2001-2002



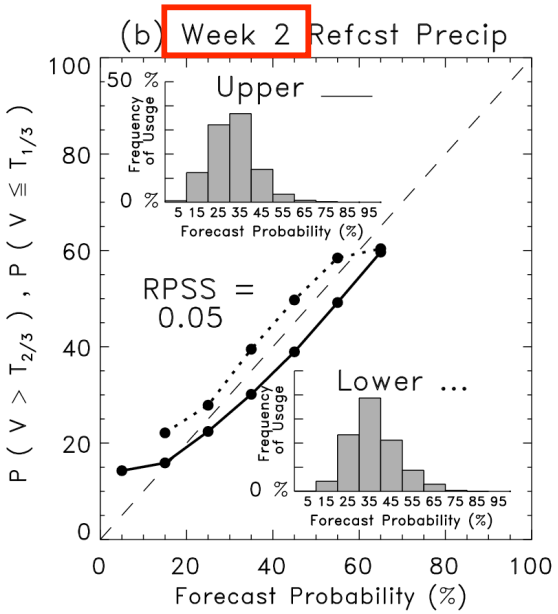
temperature forecasts



Reforecast calibrated **Week-2** forecasts more skillful than operational NCEP/CPC **6-10 day,** which was based on human blending of NCEP, ECMWF, other tools.

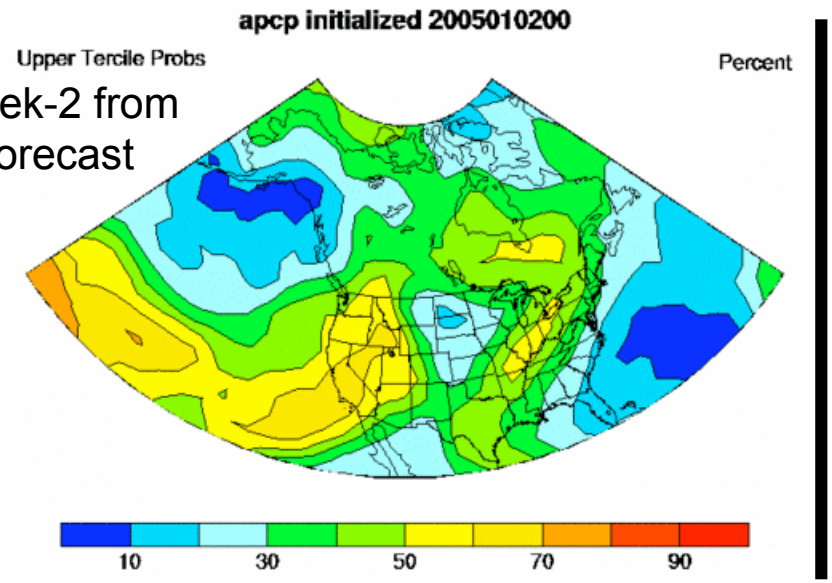


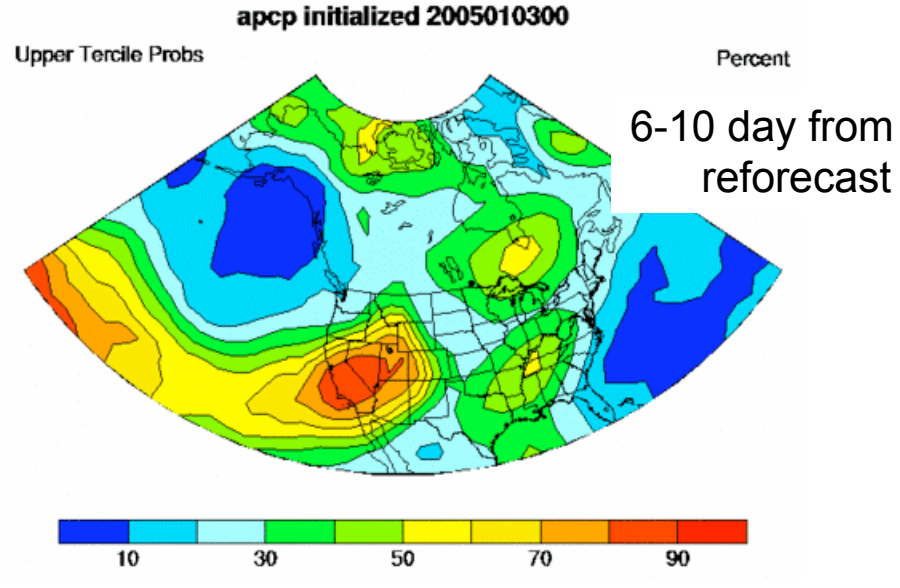
precipitation forecasts



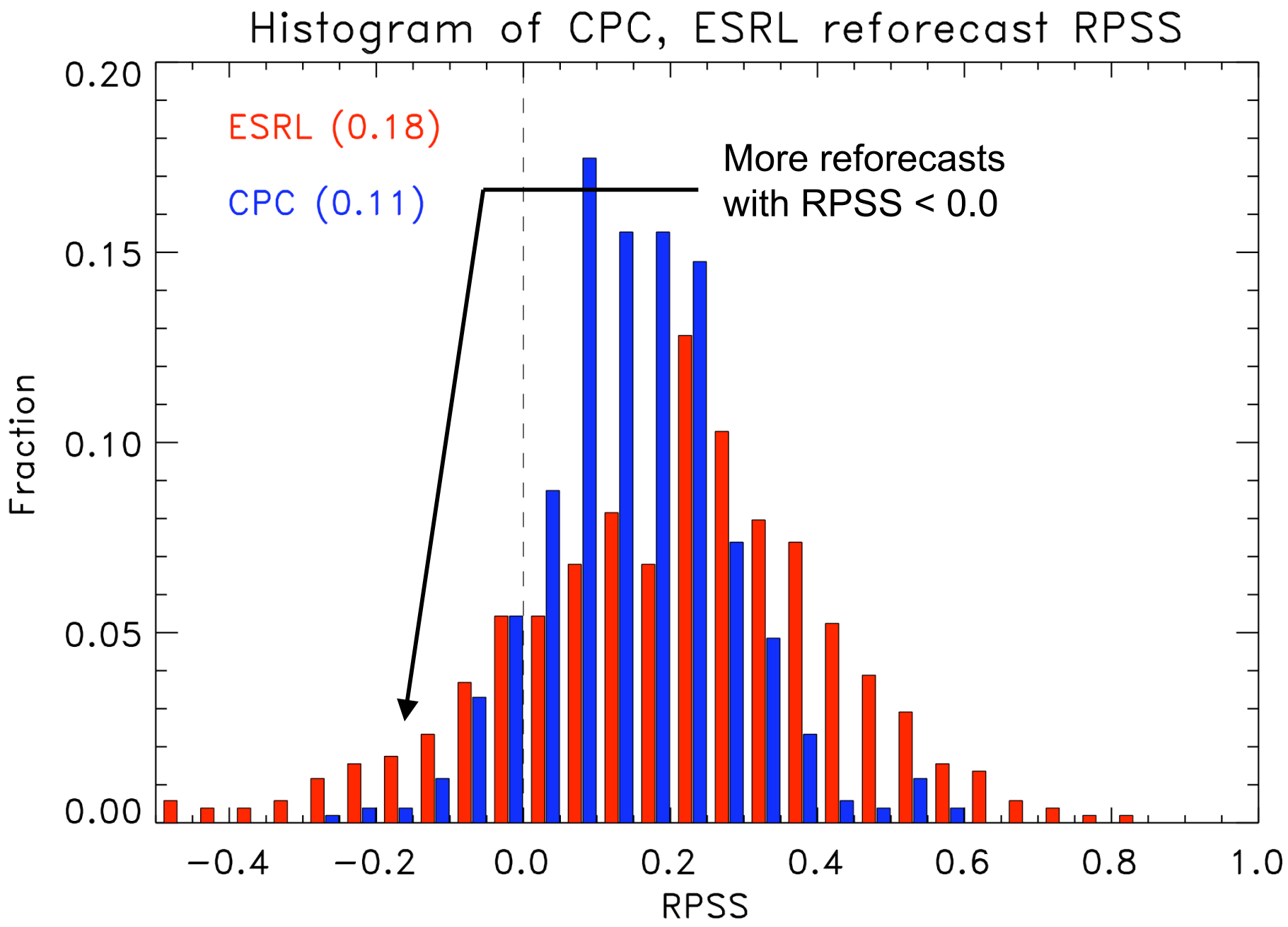
Reforecast-based example: floods causing La Conchita, California landslide, 12 Jan 2005

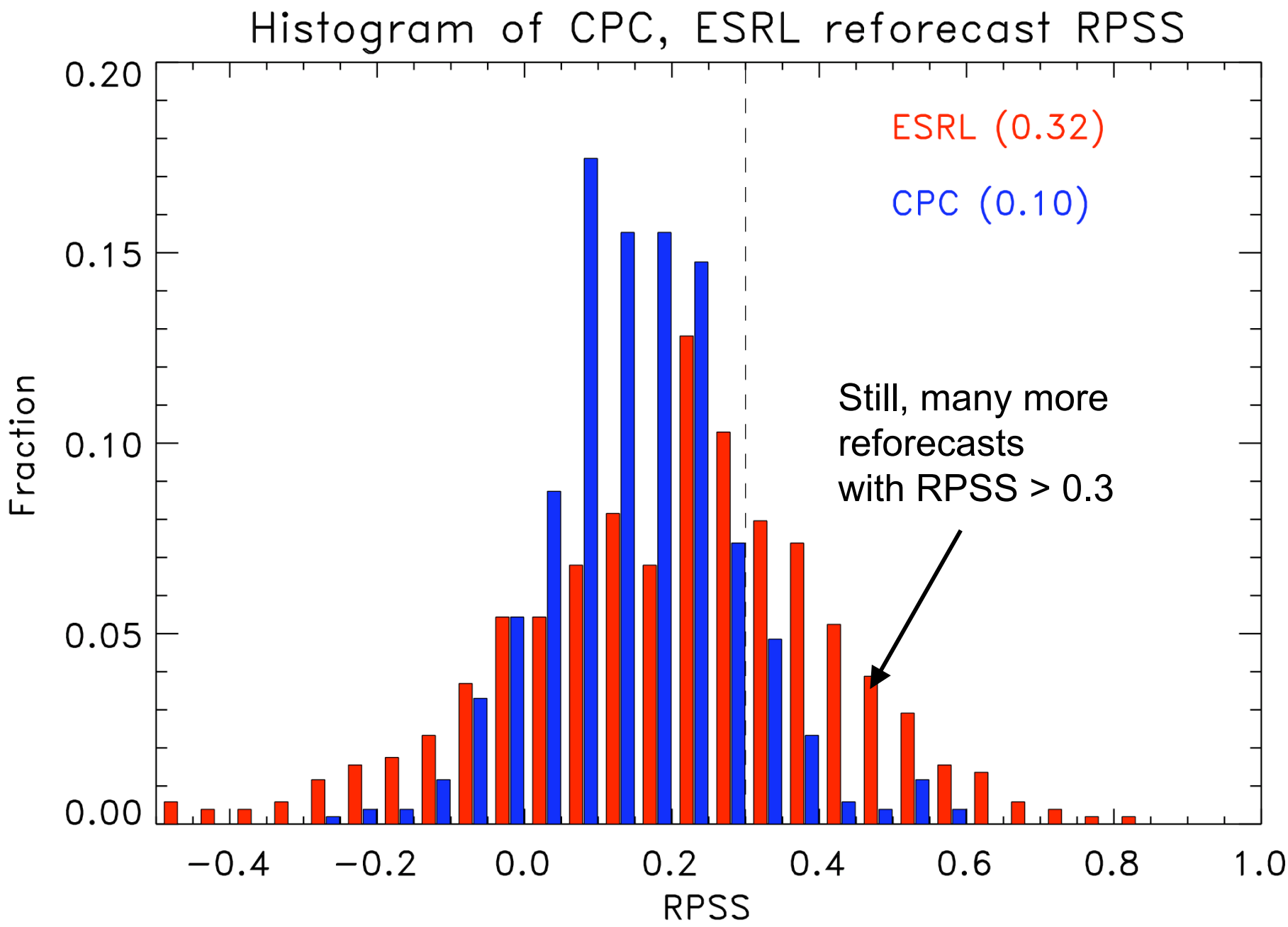
week-2 from reforecast



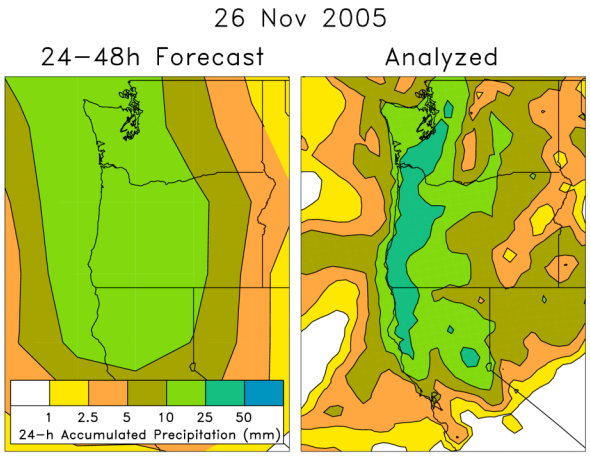


6-10 day from reforecast


Histogram of CPC, ESRL reforecast RPSS
ESRL (0.18)
CPC (0.11)
More reforecasts
with RPSS < 0.0
Fraction
0.20
0.15
0.10
0.05
0.00
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
RPSS

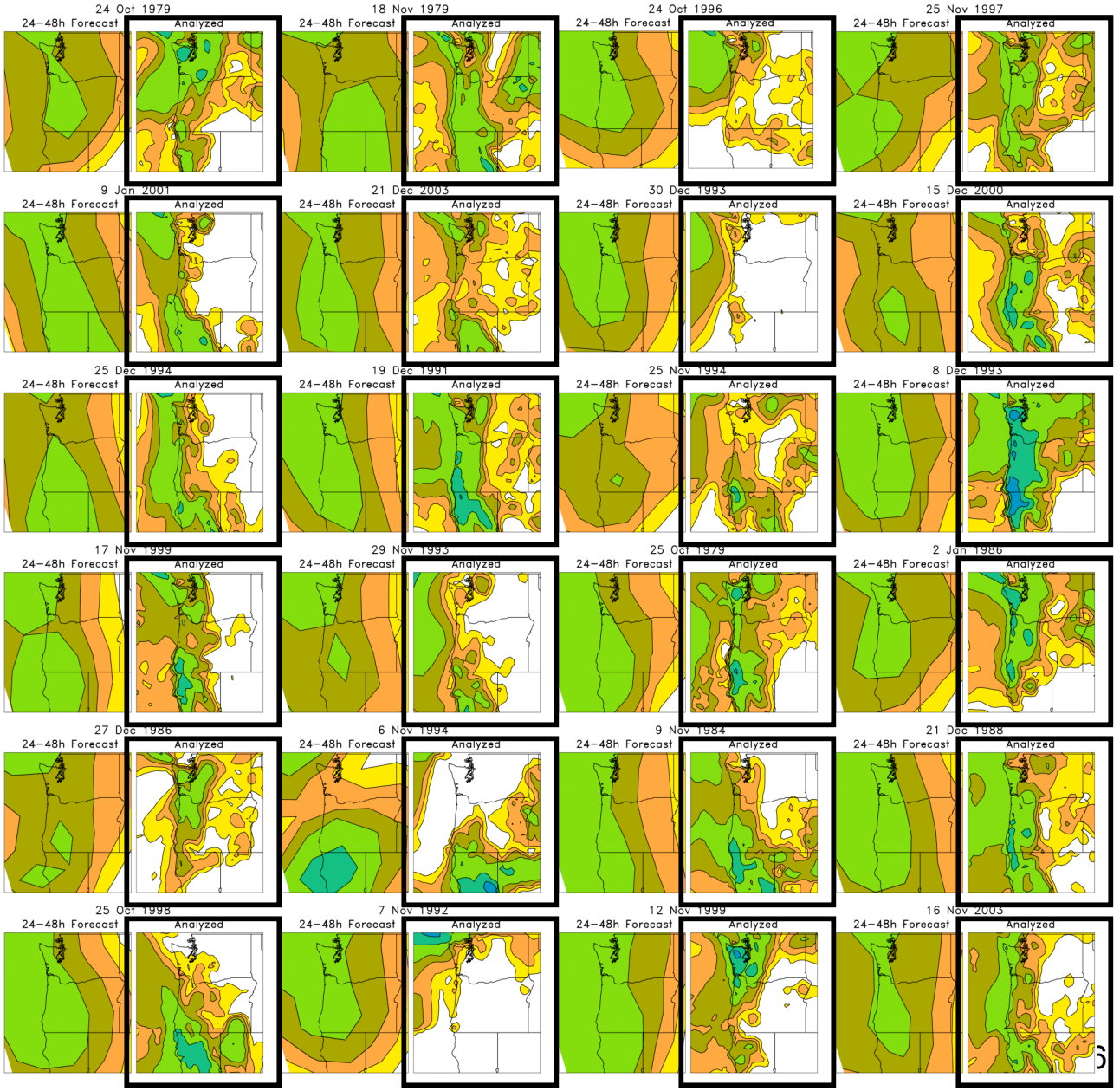

Histogram of CPC, ESRL reforecast RPSS
ESRL (0.32)
CPC (0.10)
Still, many more
reforecasts
with RPSS > 0.3
Fraction
0.20
0.15
0.10
0.05
0.00
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
RPSS

Application: downscaled precipitation forecasts using analog technique

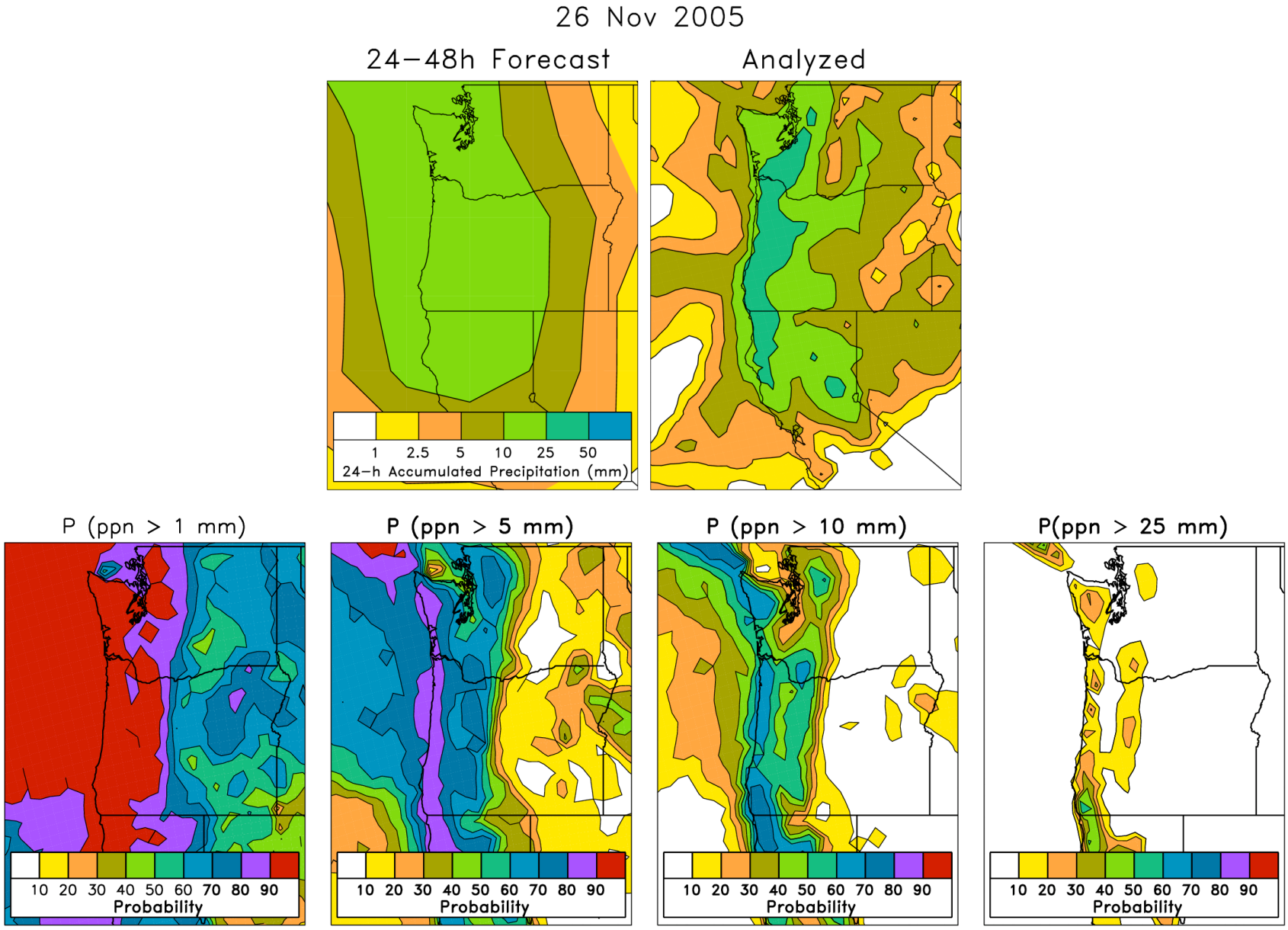


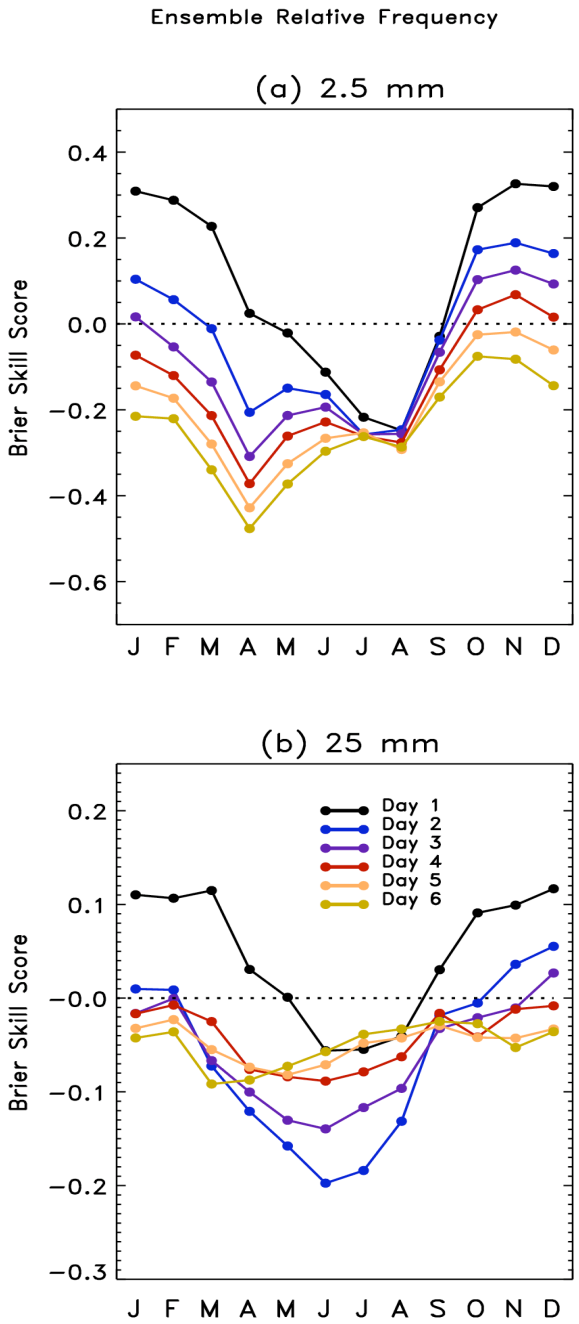
On the left are old forecasts similar to today's ensemble-mean forecast. The data on the right, the analyzed precipitation conditional upon the forecast, can be used to statistically adjust and downscale the forecast.

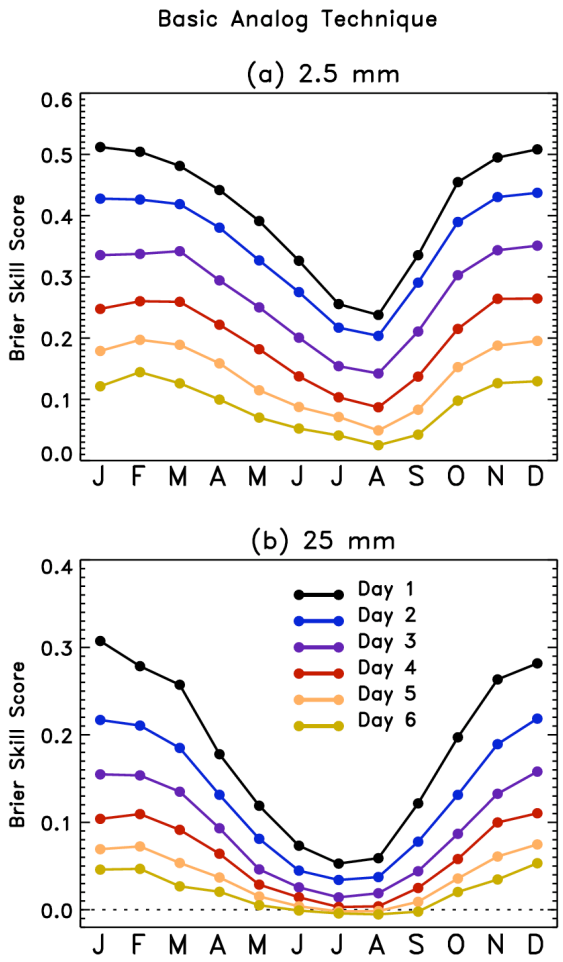
Analog approaches like this may be particularly useful for hydrologic ensemble applications, where an ensemble of realizations is needed.



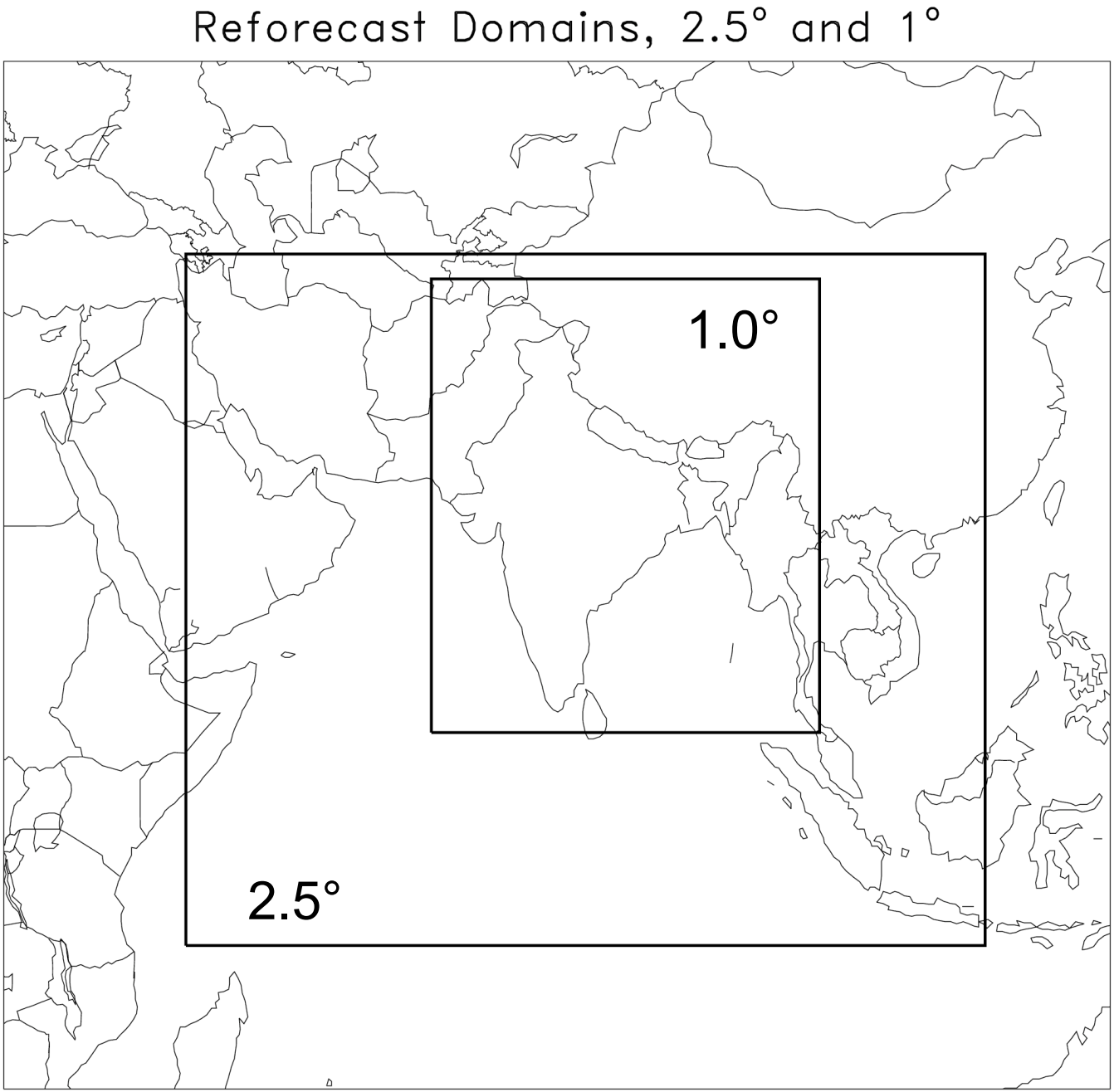
Downscaled analog probability forecasts







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

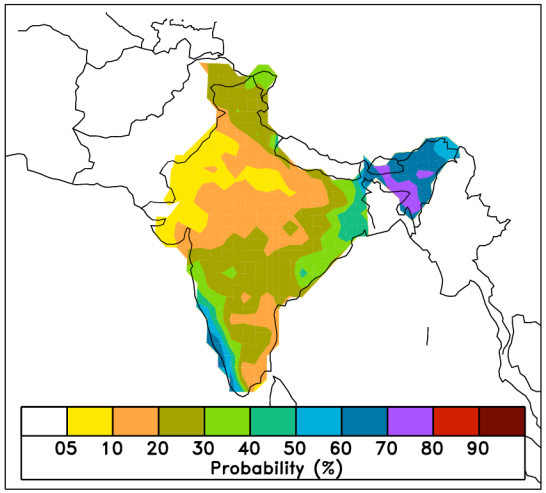


Application: monsoon forecasts over India

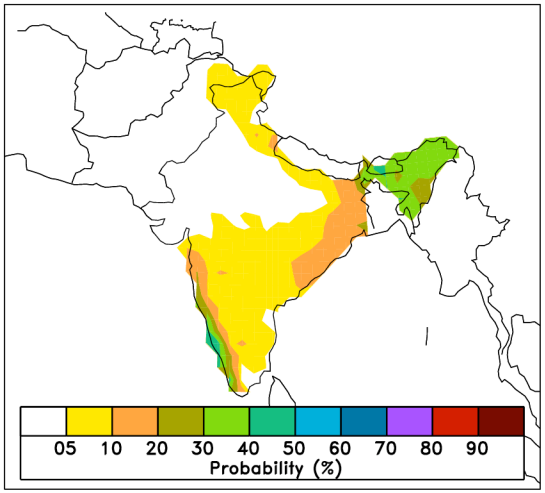
For this experiment we saved forecast total precipitation, column precipitable water, and sea-level pressure tendency on coarse and fine grids, as shown, for May 15 - Oct 15, 1979-2007. 1-degree precipitation analyses available over India.

Monsoon precipitation climatology

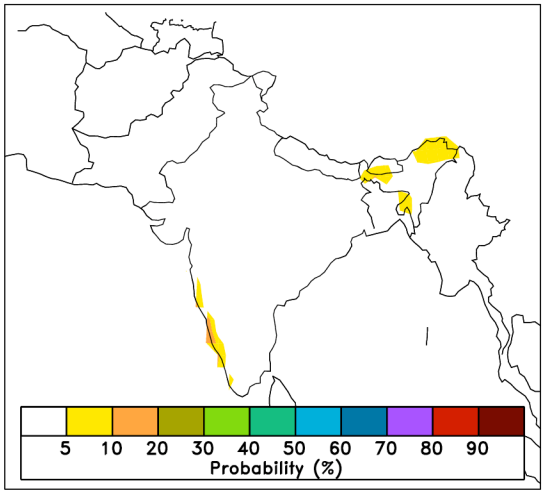
(a) Climatological
P(Obs > 1 mm) Jun 01



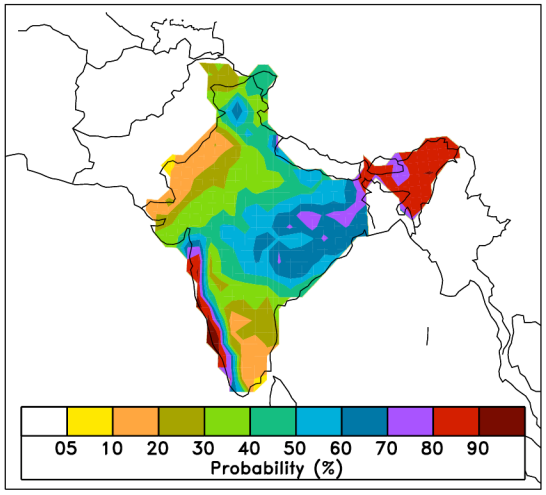
(b) Climatological
P(Obs > 10 mm) Jun 01



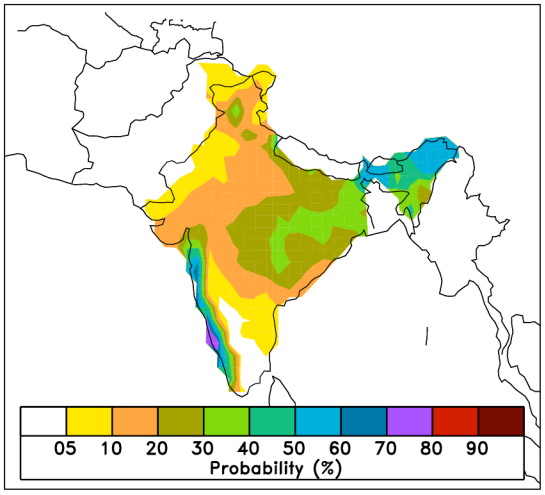
(c) Climatological
P(Obs > 50 mm) Jun 01



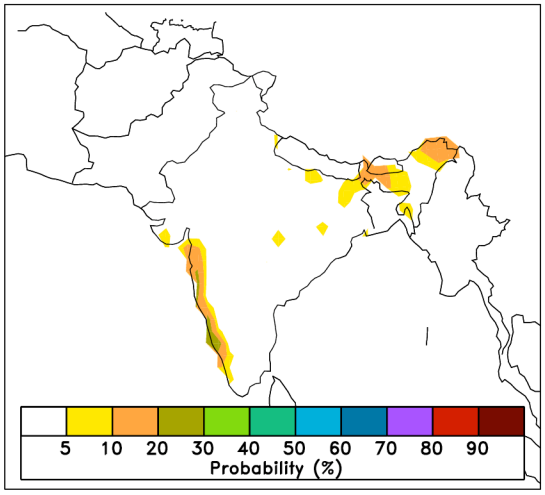
(a) Climatological
P(Obs > 1 mm) Jul 01



(b) Climatological
P(Obs > 10 mm) Jul 01

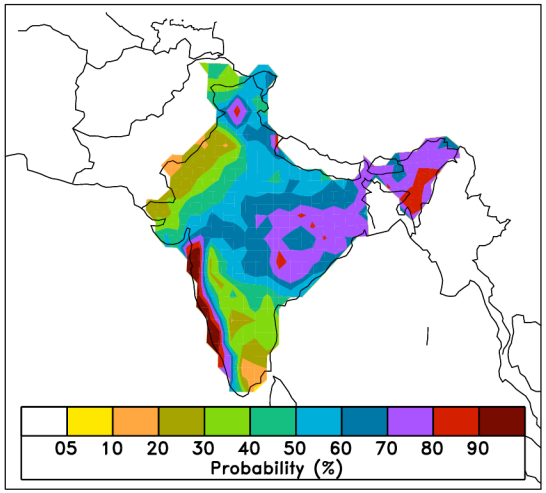


(c) Climatological
P(Obs > 50 mm) Jul 01

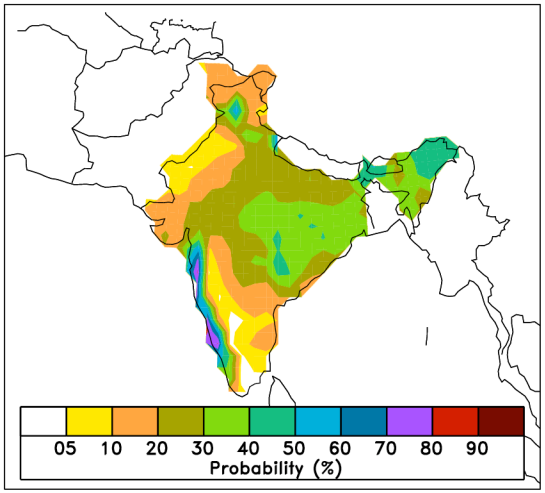


Monsoon precipitation climatology

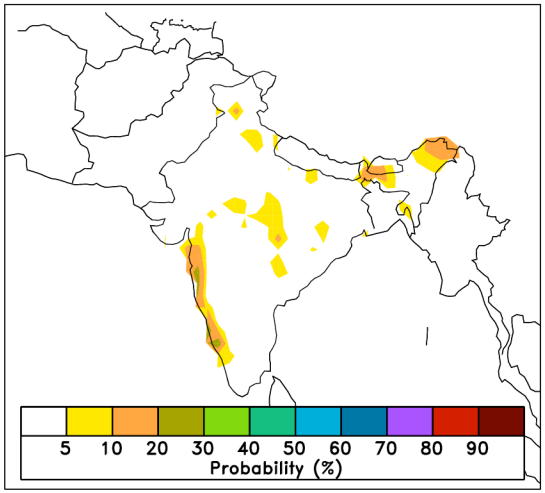
(a) Climatological
P(Obs > 1 mm) Aug 01



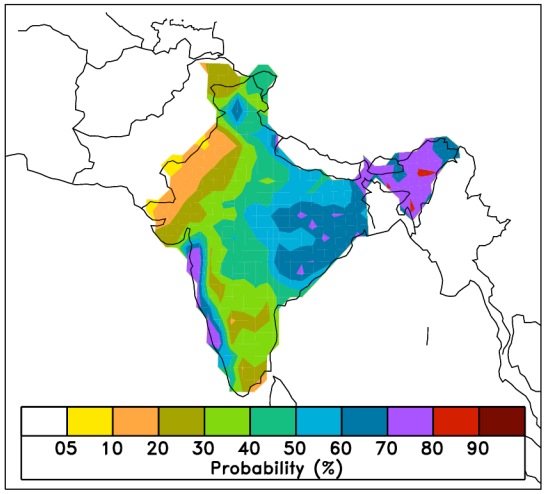
(b) Climatological
P(Obs > 10 mm) Aug 01



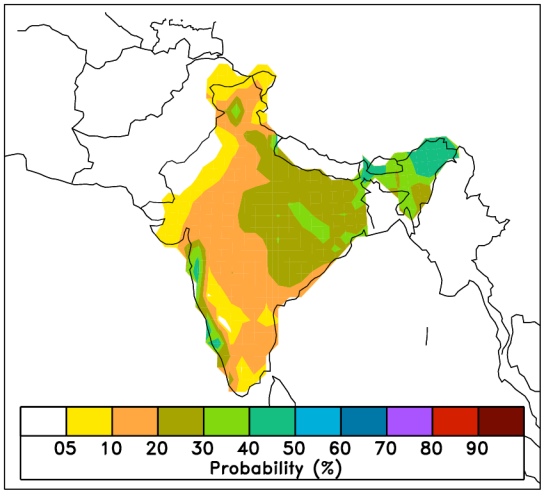
(c) Climatological
P(Obs > 50 mm) Aug 01



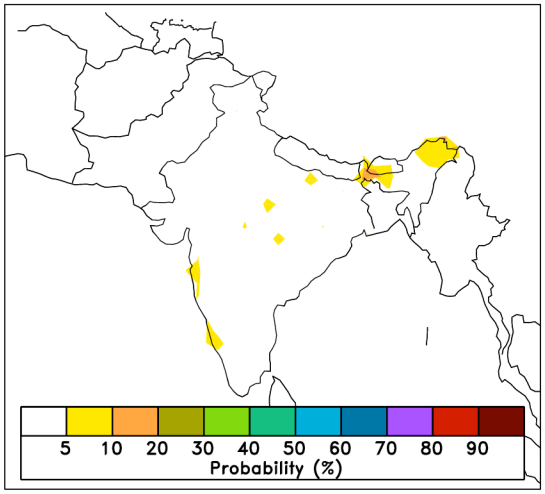
(a) Climatological
P(Obs > 1 mm) Sep 01



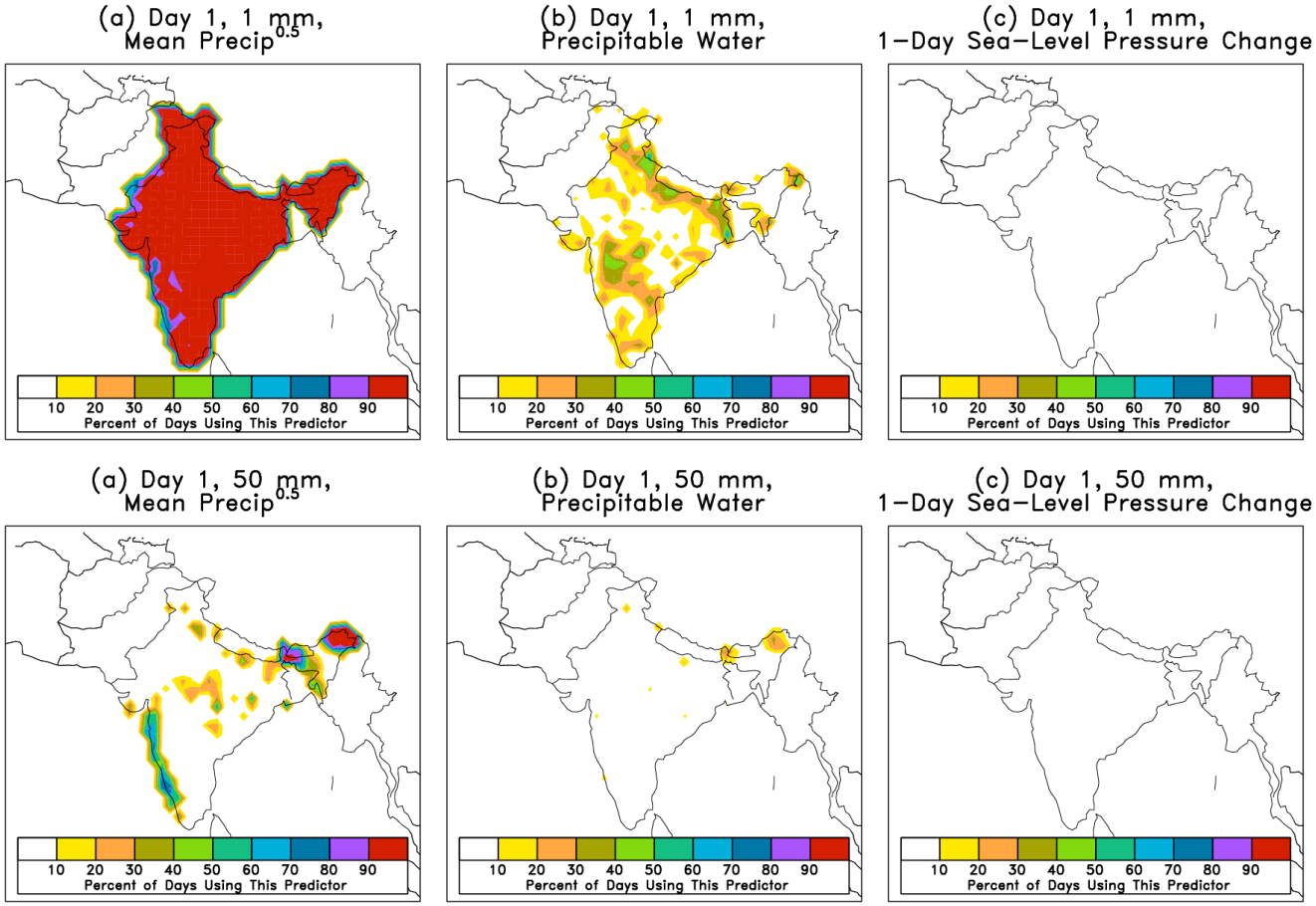
(b) Climatological
P(Obs > 10 mm) Sep 01



(c) Climatological
P(Obs > 50 mm) Sep 01

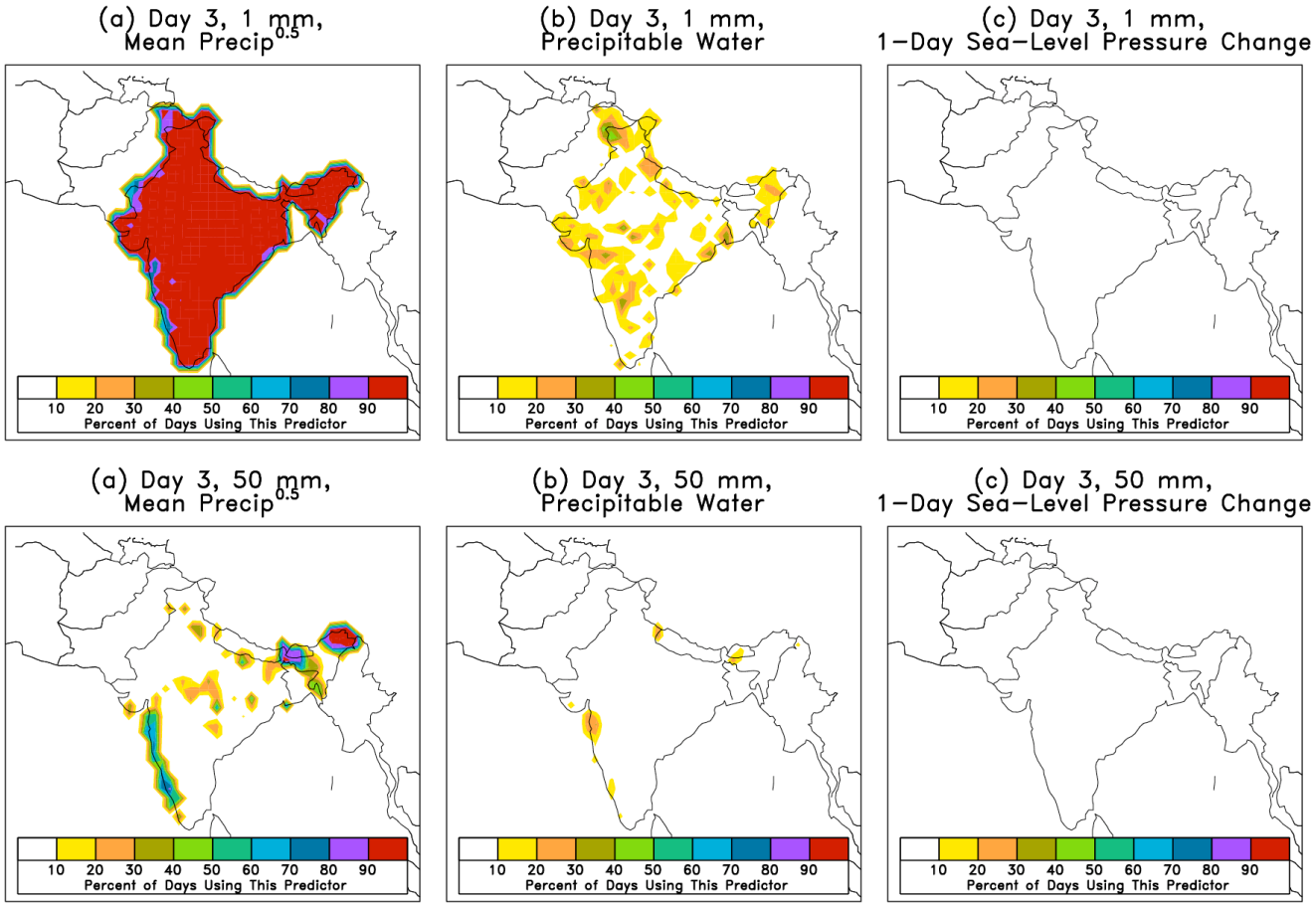


Which predictors in logistic regression with stepwise elimination? Day 1



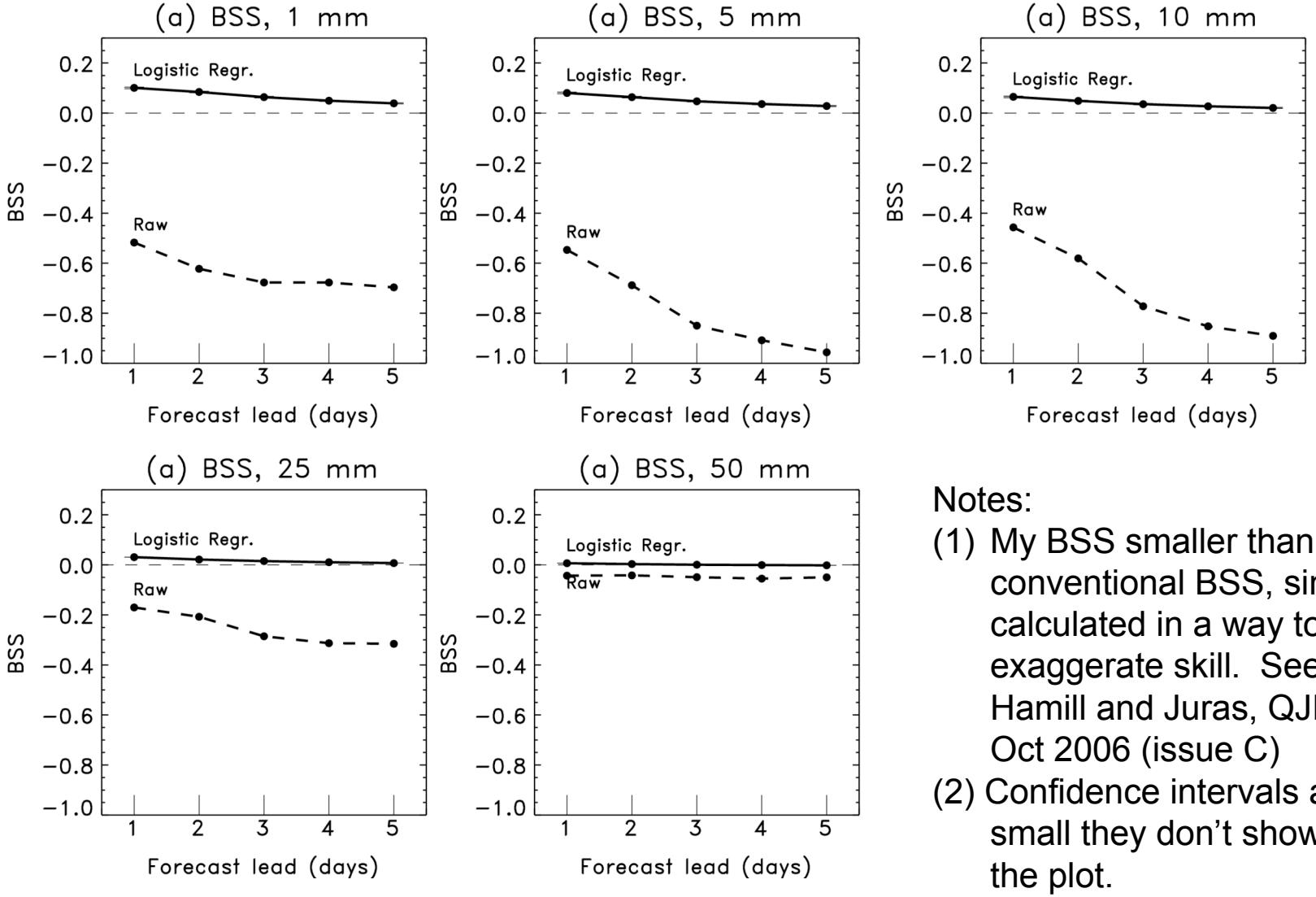
For every day of the monsoon season, a stepwise linear regression was run to determine which predictors provided a reduction in error. As shown, a power-transformed ensemble-mean forecast precipitation was uniformly selected as an important predictor. Precipitable water was occasionally selected, and sea-level pressure change was virtually never selected. Based on these results, all subsequent logistic regression analyses will be based on using only one predictor, the power-transformed ensemble-mean precipitation amount.

Which predictors in logistic regression with stepwise elimination? Day 3



The same conclusion is reached when considering other forecast leads.

Brier Skill Scores

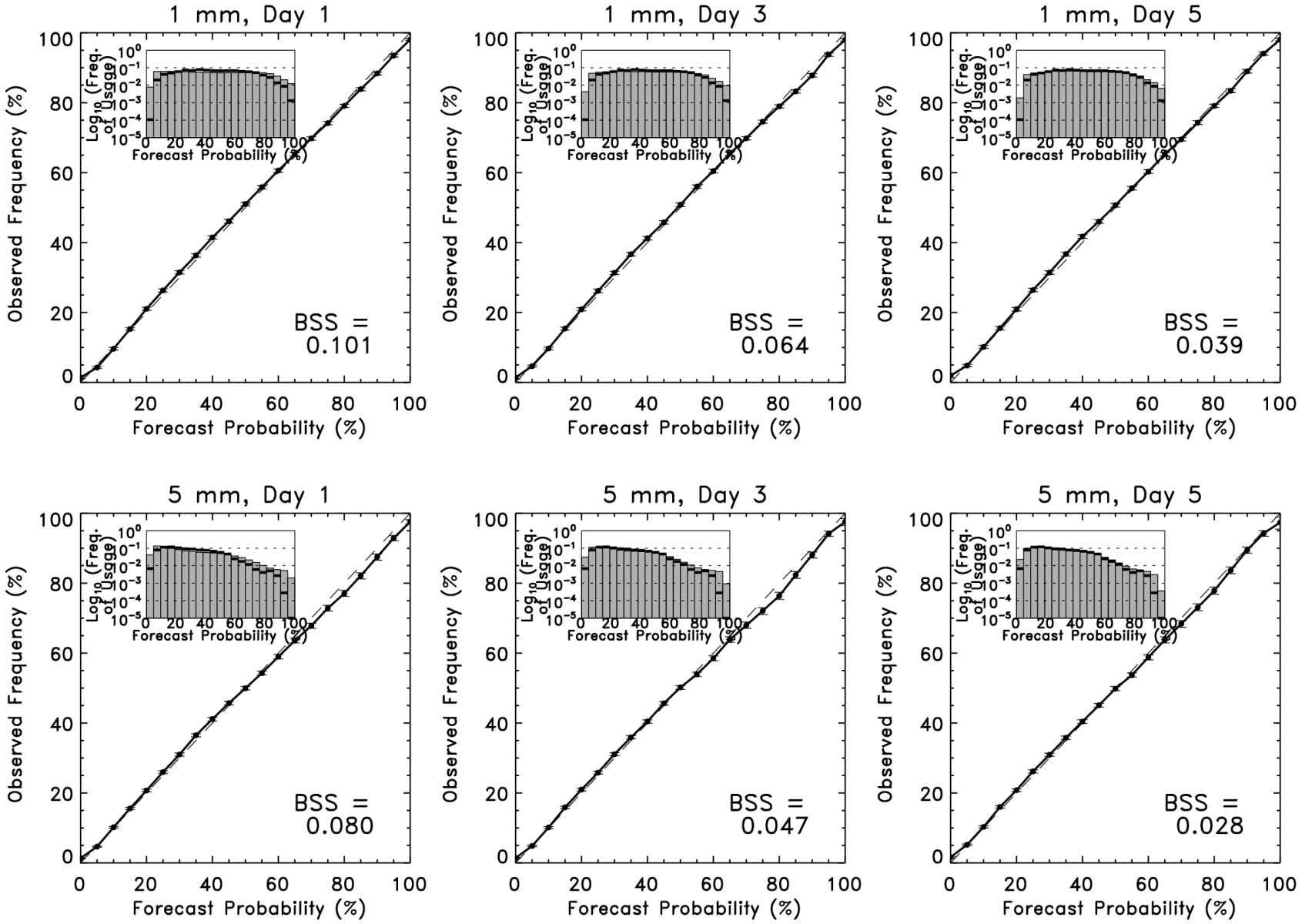


Notes:

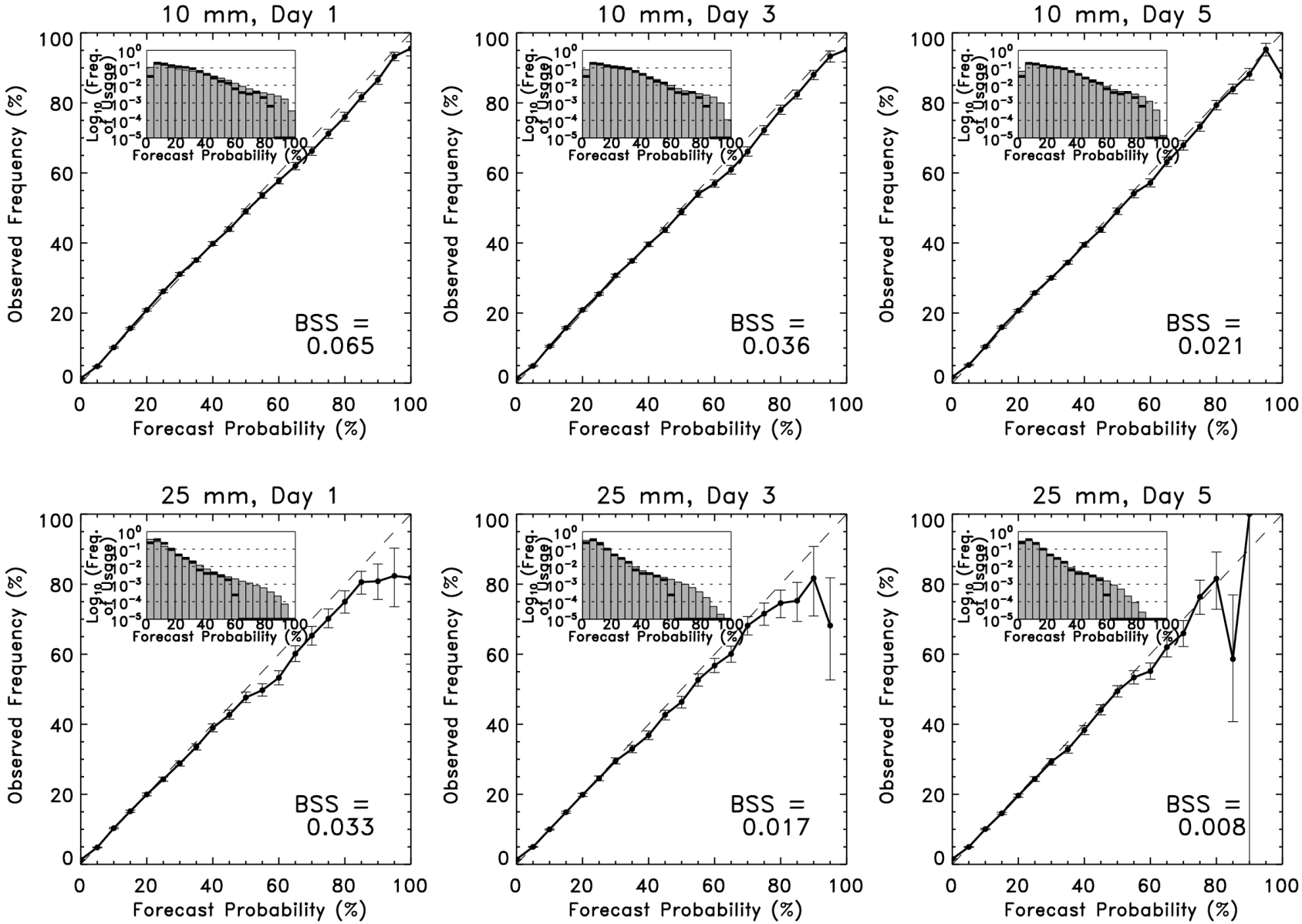
(1) My BSS smaller than conventional BSS, since calculated in a way to not exaggerate skill. See Hamill and Juras, QJRMS, Oct 2006 (issue C)

(2) Confidence intervals are so small they don't show up on the plot.

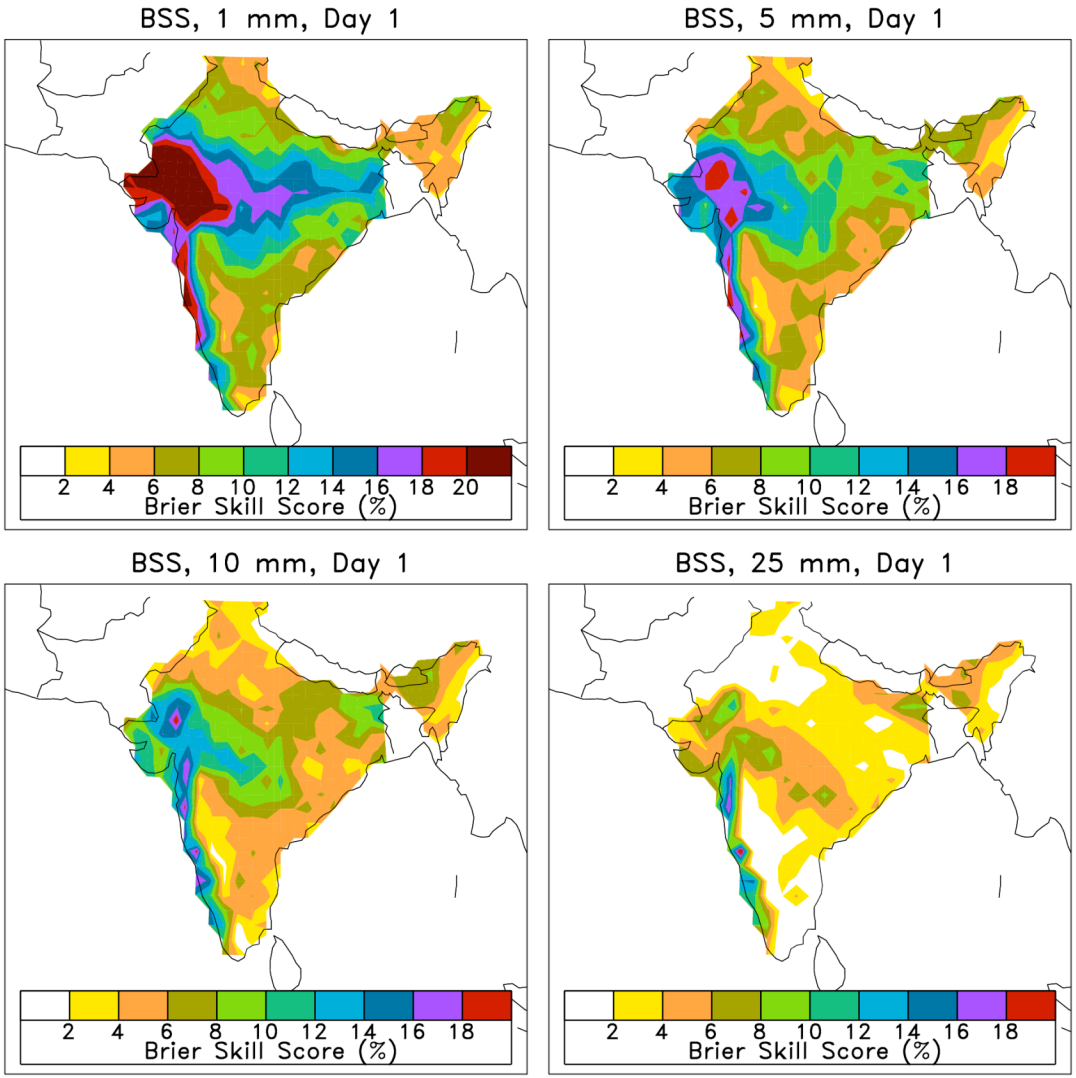
Reliability, logistic regression, 1 and 5 mm



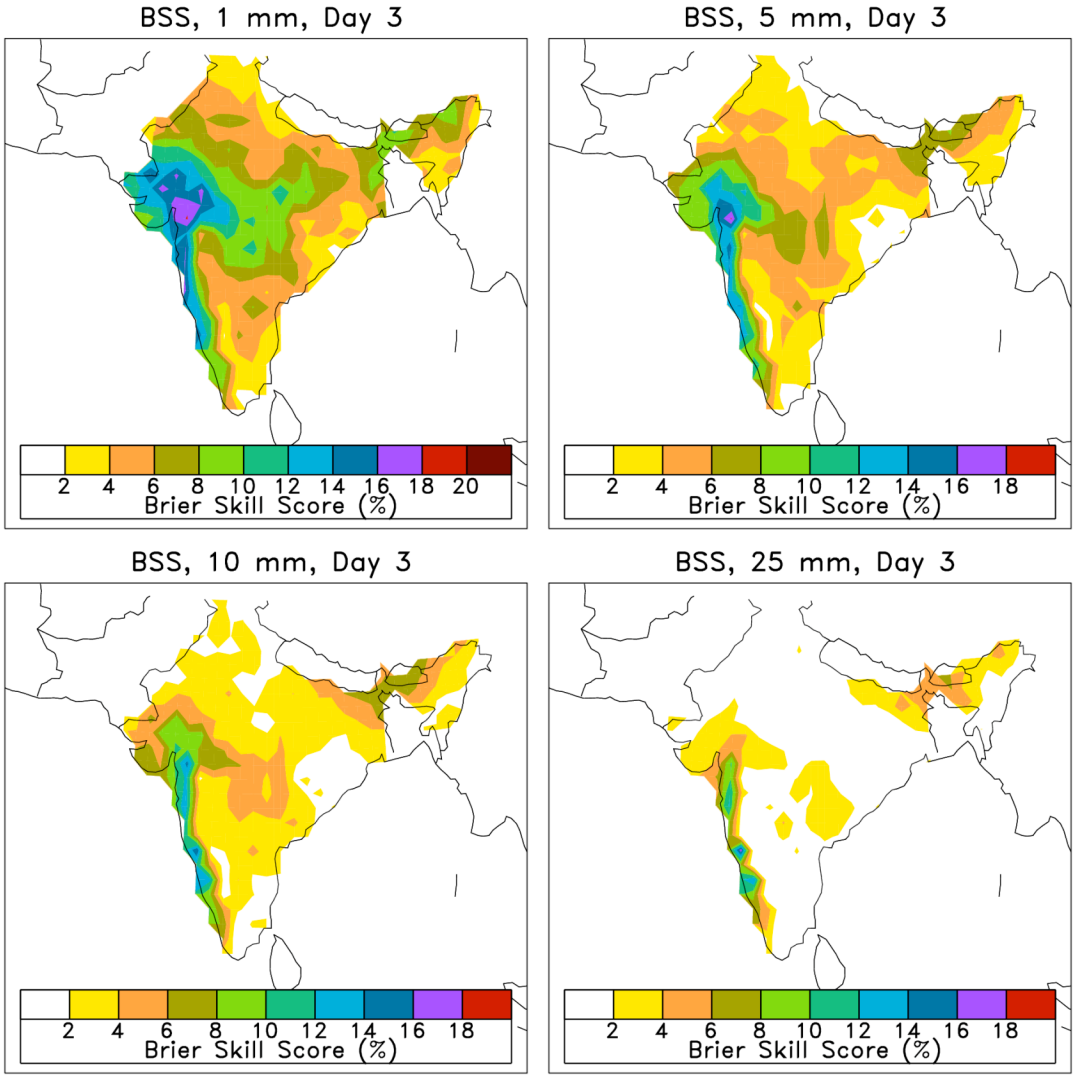
Reliability, logistic regression, 10 and 25 mm



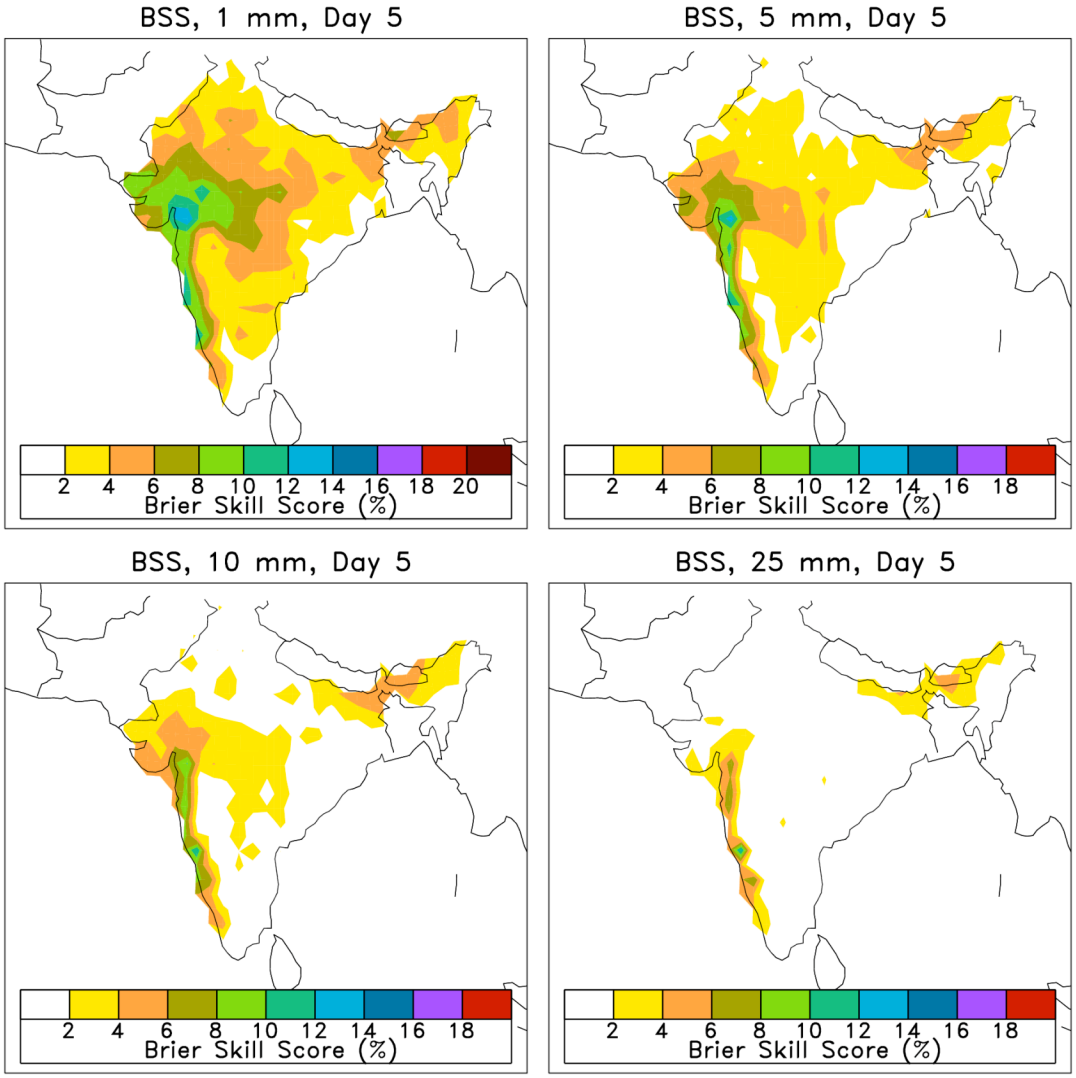
Map of logistic regression BSS, day 1



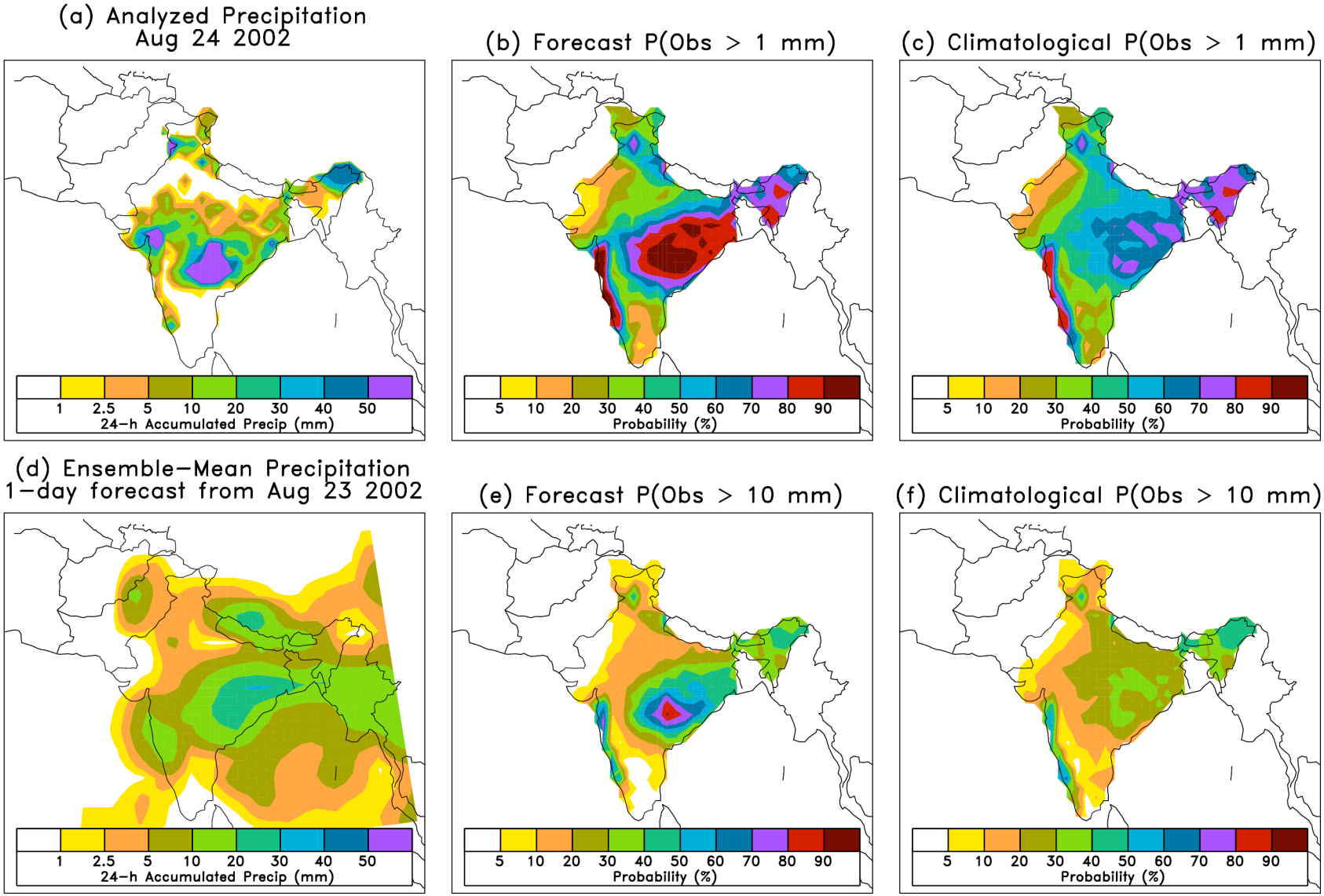
Map of logistic regression BSS, day 3



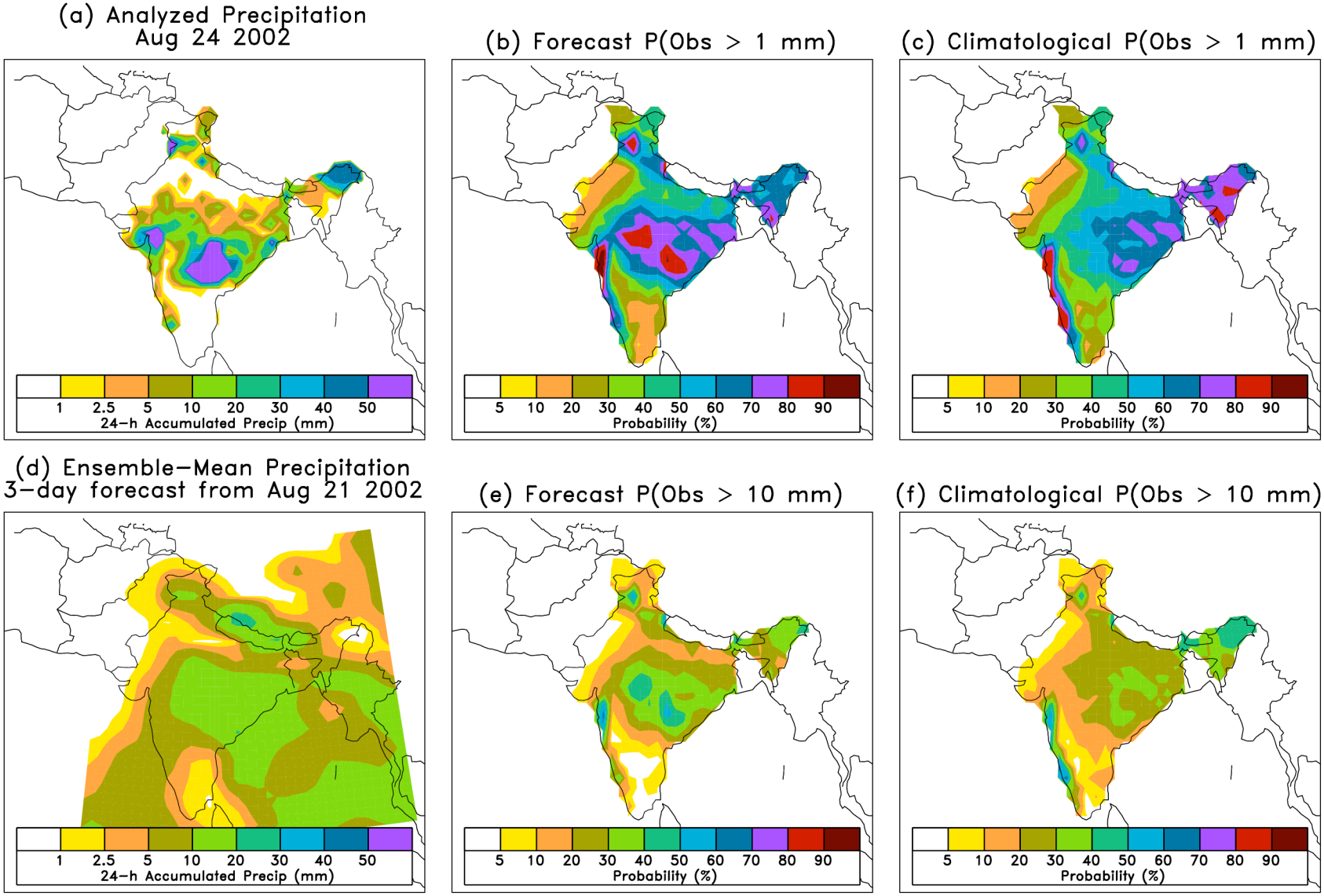
Map of logistic regression BSS, day 5



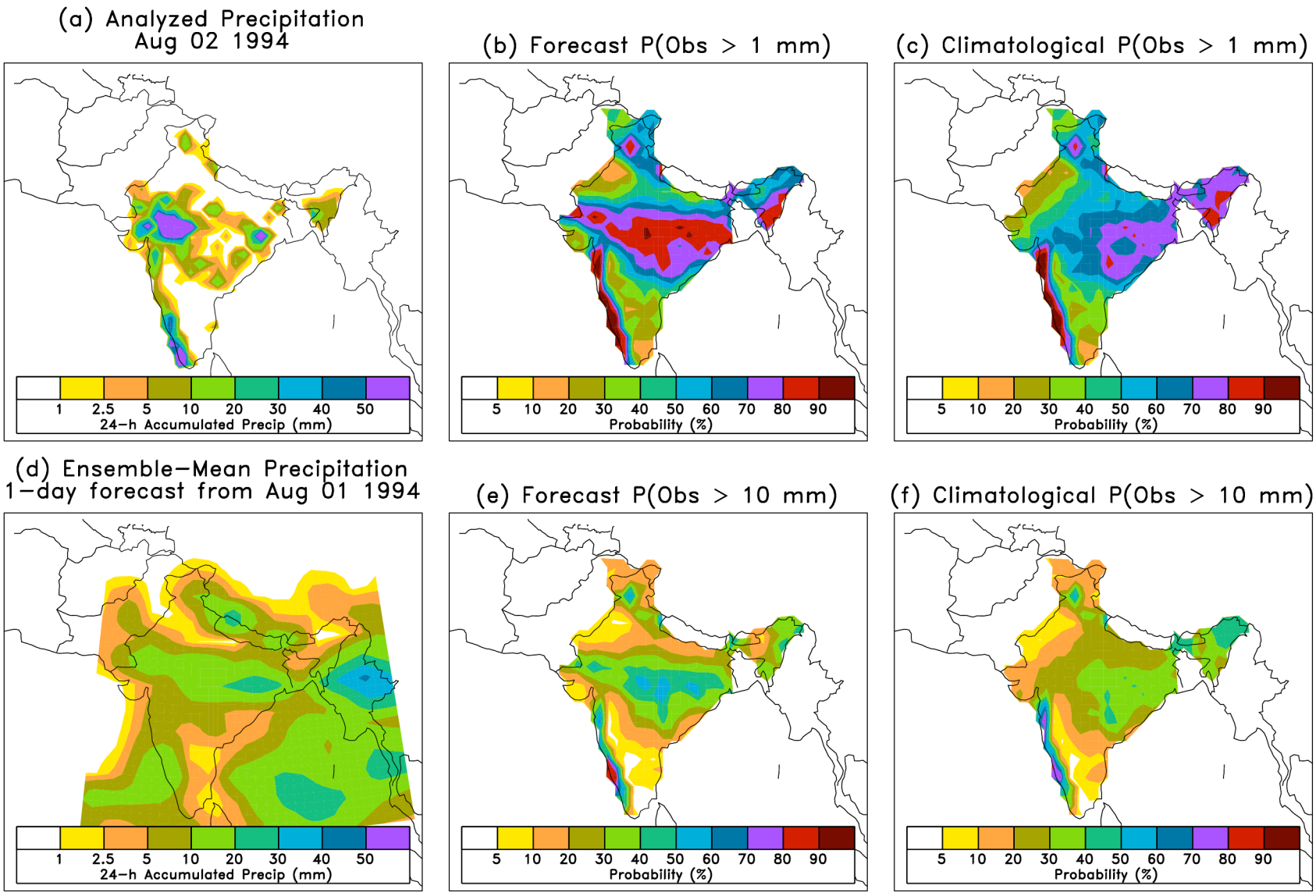
Logistic regression forecast example #1, 1-day lead



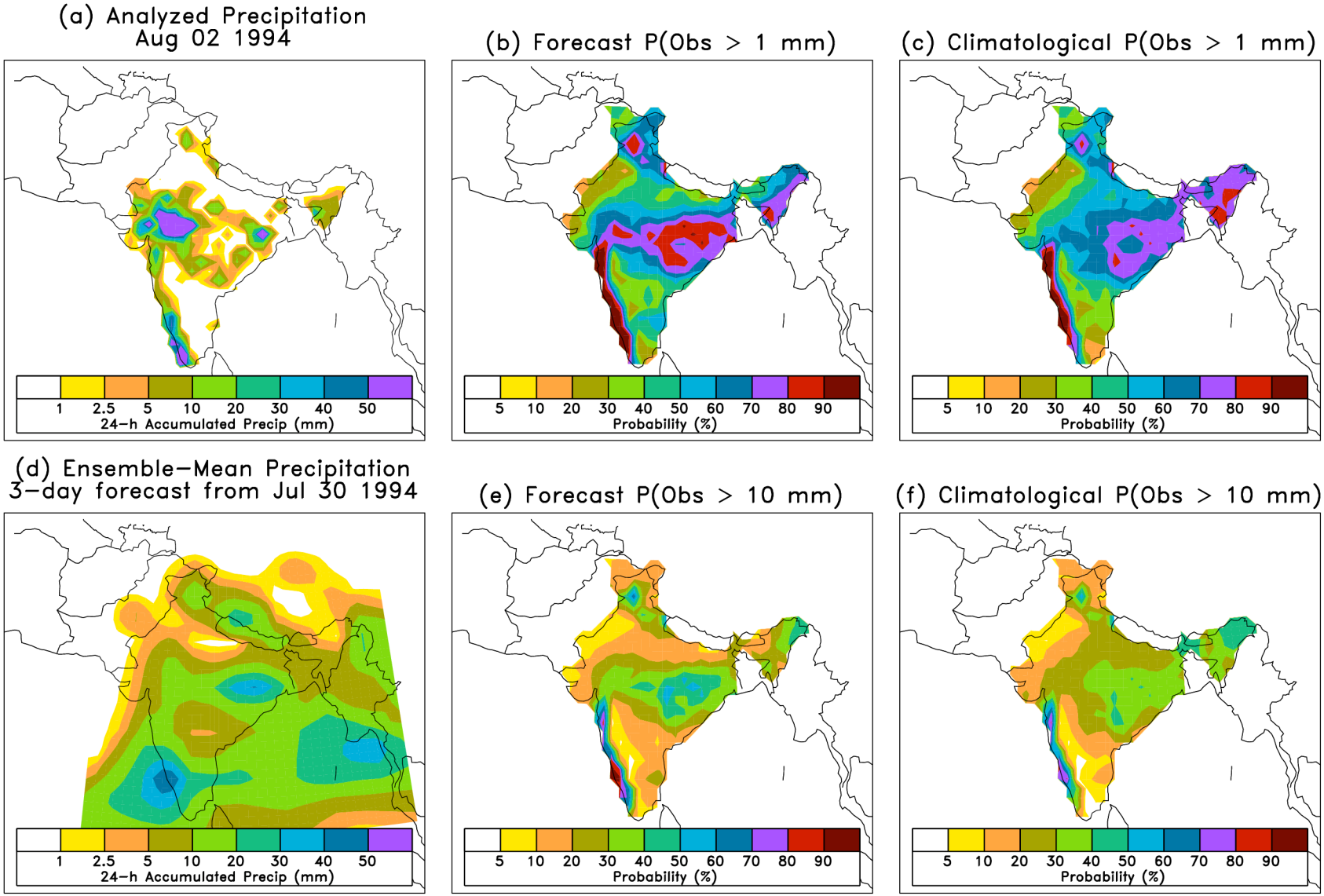
Logistic regression forecast example #1, 3-day lead



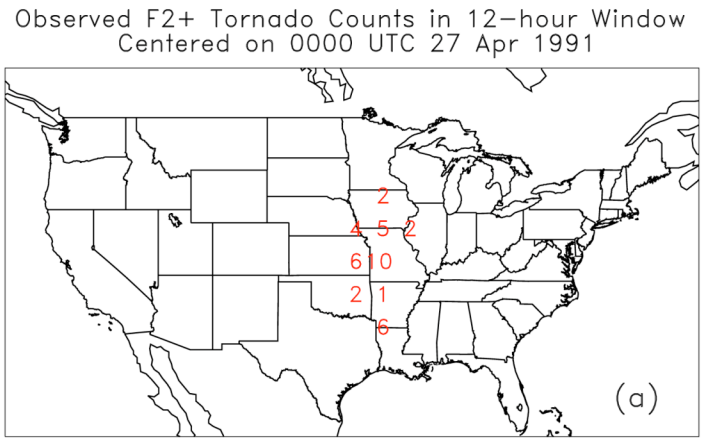
Logistic regression forecast example #2, 1-day lead

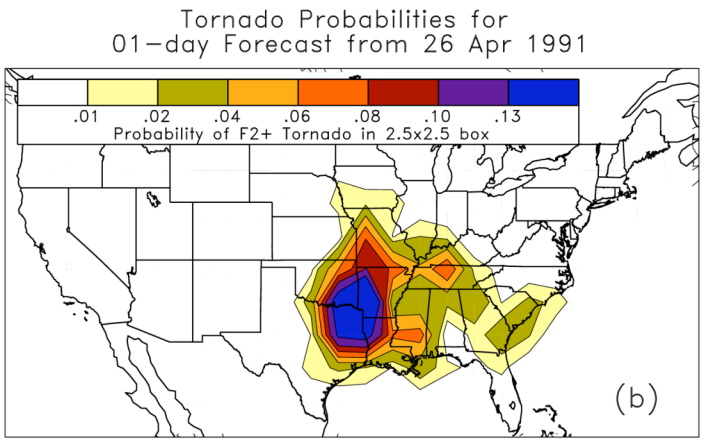


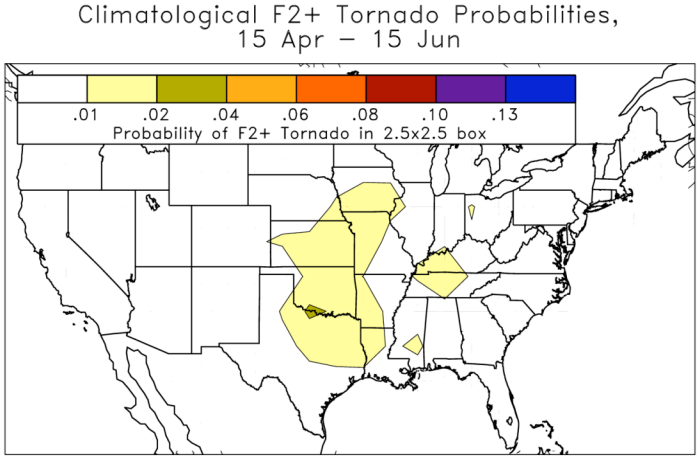
Logistic regression forecast example #2, 3-day lead



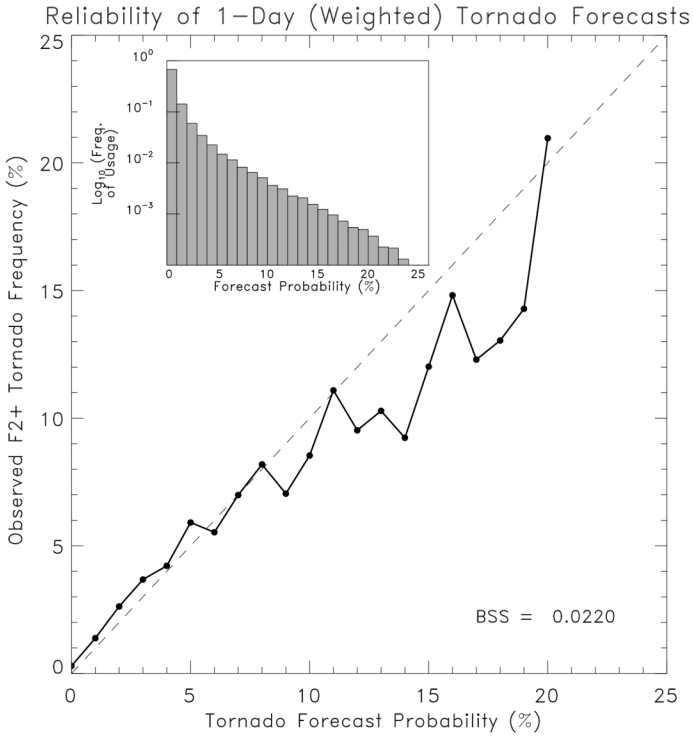
Tornado probability forecasting







forecast wind shear and instability were used as predictors in an analog approach.



Part IIb:

Calibration using ECMWF reforecast data set

Questions

- Will reforecasts benefit calibration of a state-of-the art model like ECMWF's as much as with now outdated GFS model?

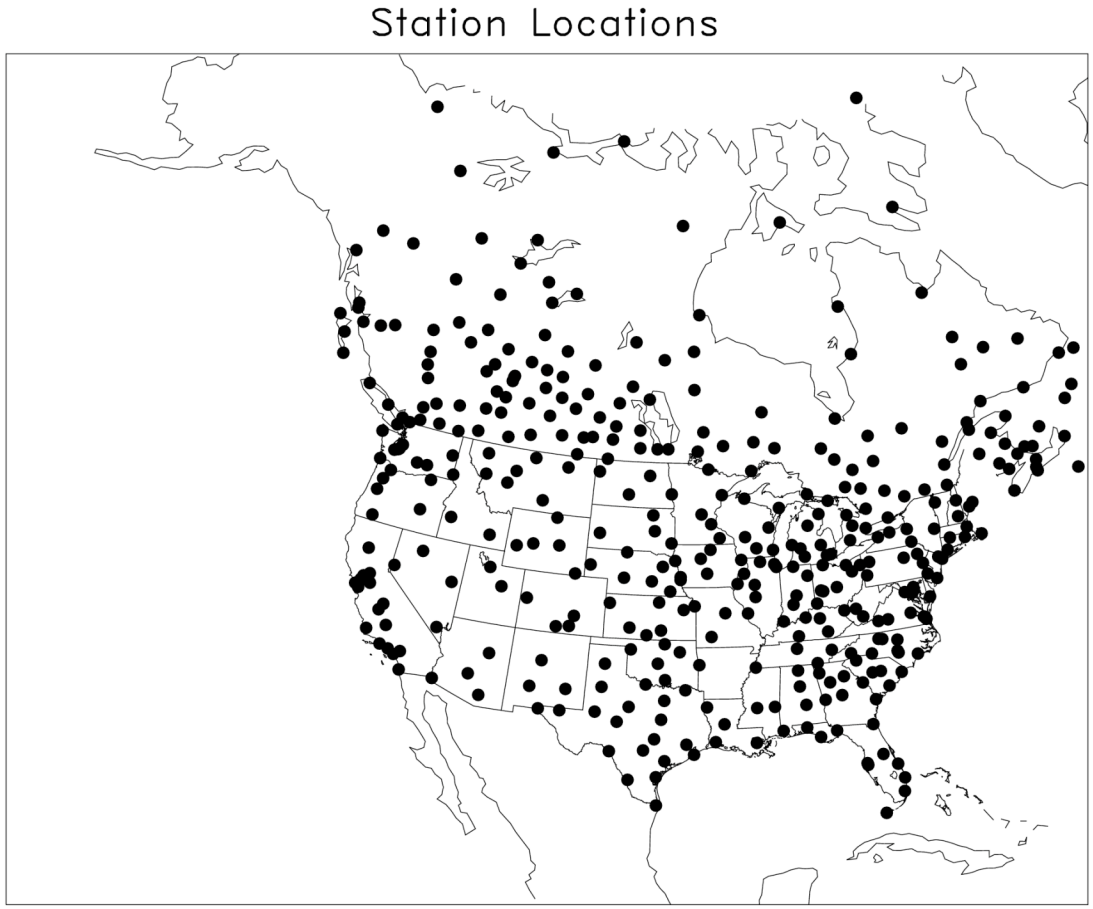
- How do probabilistic forecasts from the old GFS, with calibration, compare to the new ECMWF without?

- Are multi-decadal reforecasts really necessary? Given the computational expense of computing them, are much smaller training data sets adequate for probabilistic forecast calibration?

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 ×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.

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.

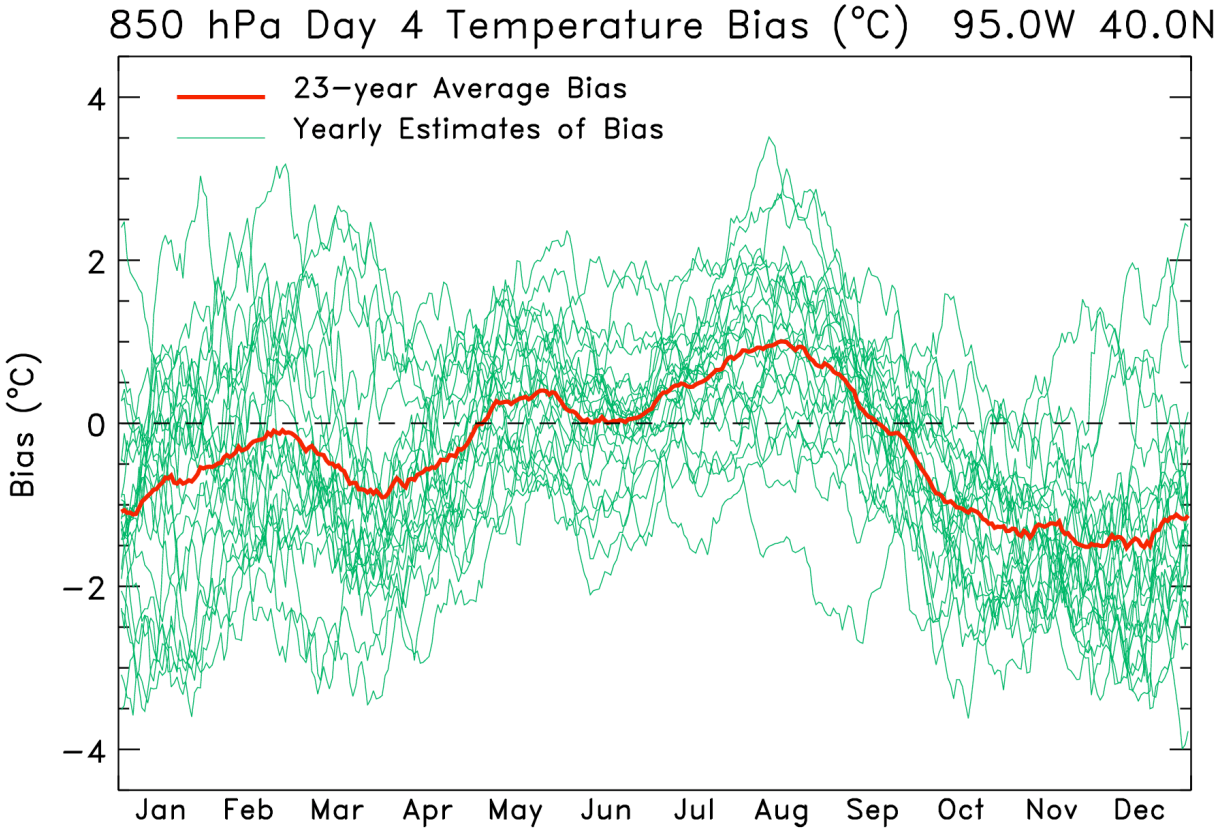
Calibration procedure: "NGR"
"Non-homogeneous Gaussian Regression"

- **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.

Inter-annual variability of forecast bias

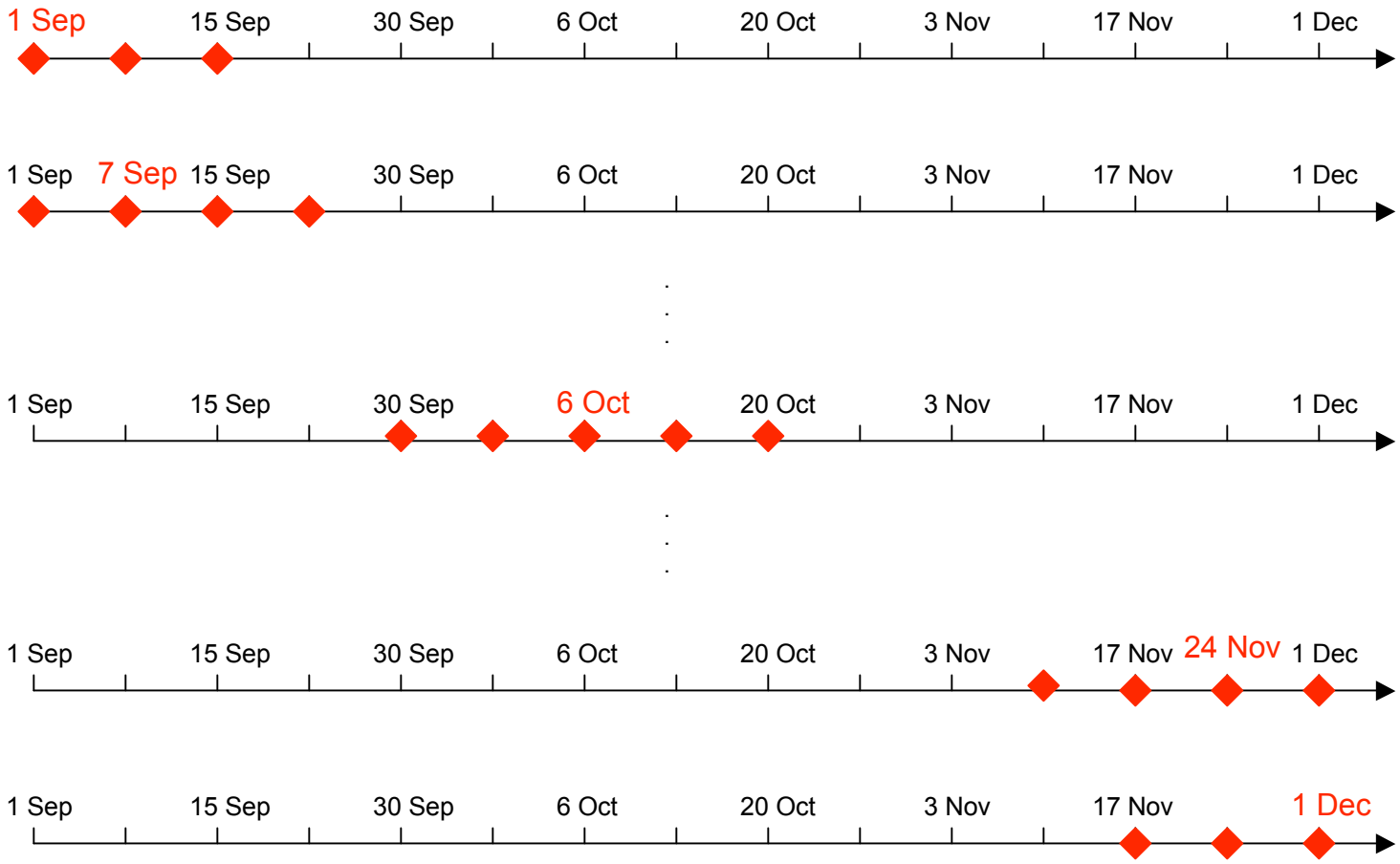


Red curve shows bias averaged over 23 years of data (bias = mean F-O in running 61-day window)

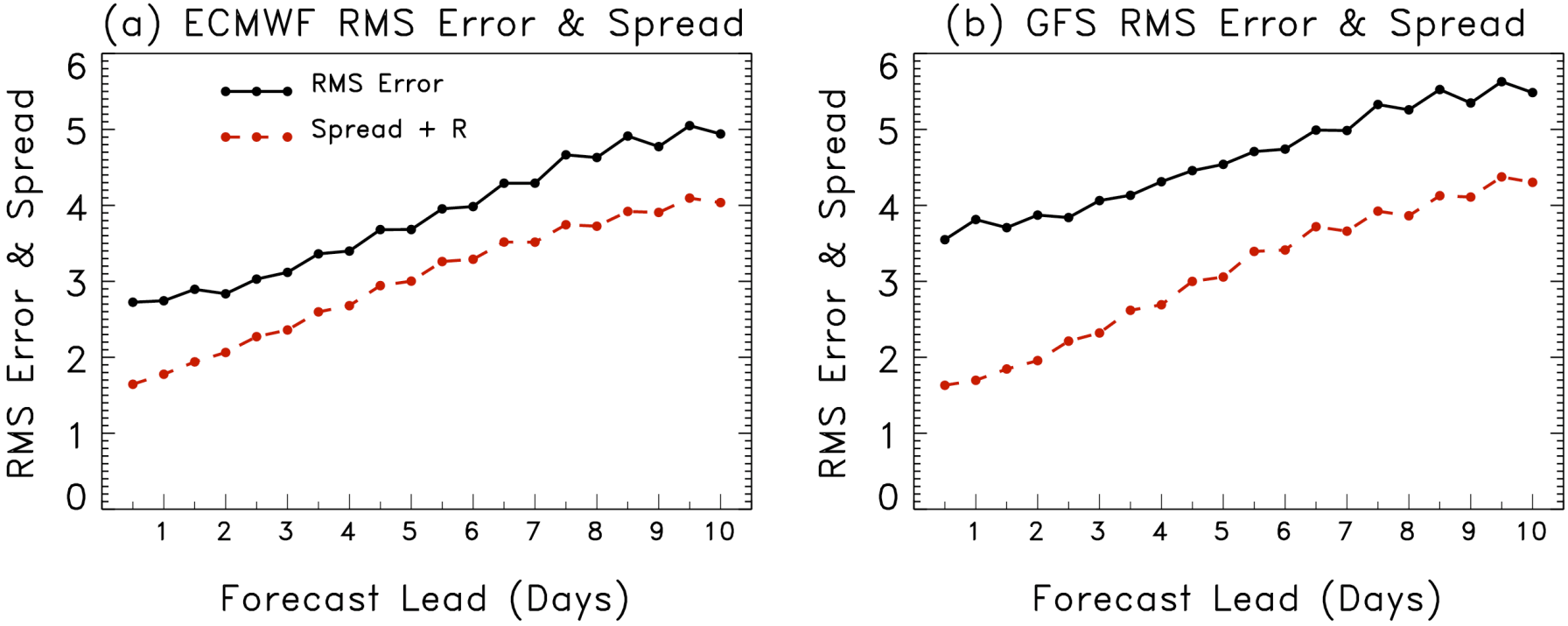
Green curves show 23 individual yearly running-mean bias estimates

Note large inter-annual variability of bias.

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

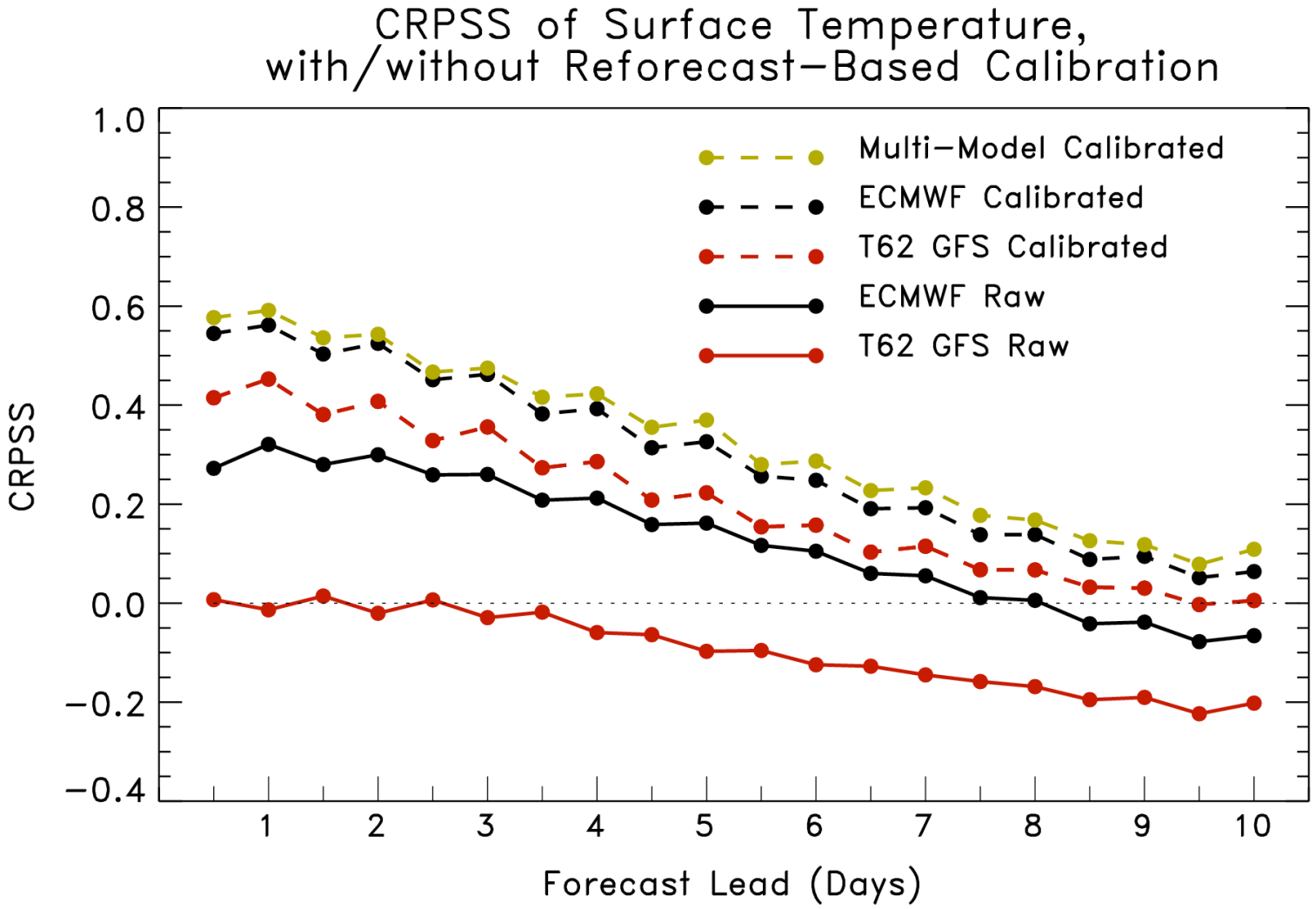


Forecast spread and error



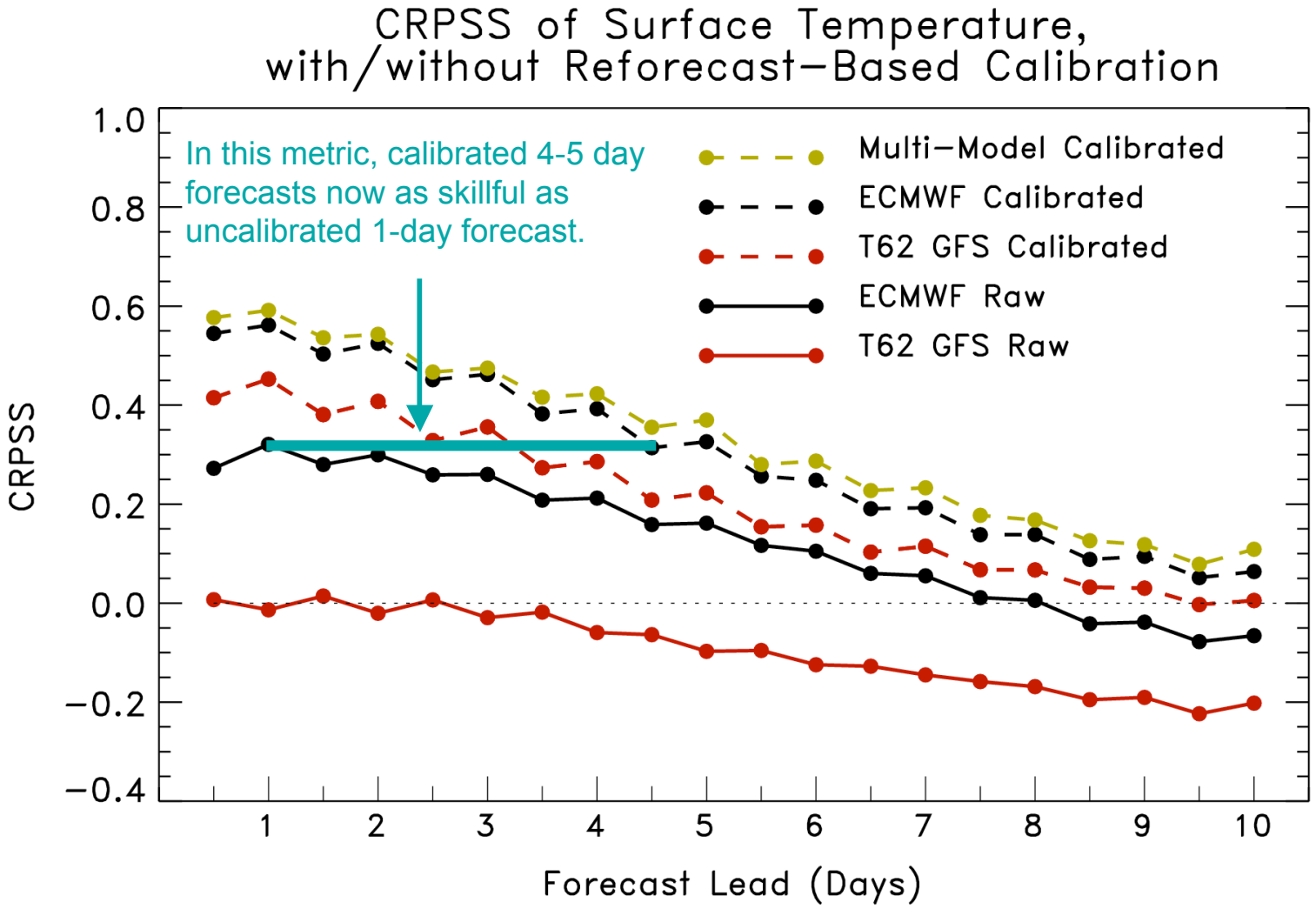
For both systems, with 2-m temperature, there is a deficiency of spread. This is much worse for GFS than ECMWF.

ECMWF, raw and post-processed



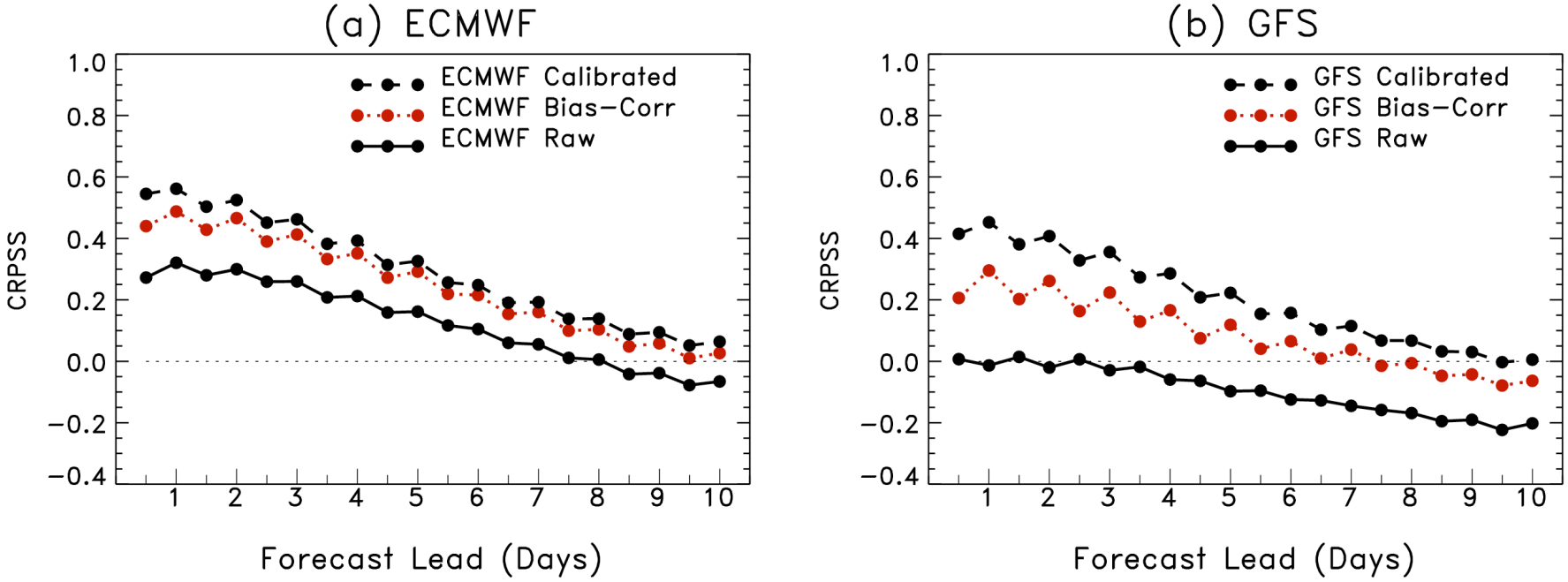
Note: 5th and 95th %ile confidence intervals very small, 0.02 or less

ECMWF, raw and post-processed



Note: 5th and 95th %ile confidence intervals very small, 0.02 or less

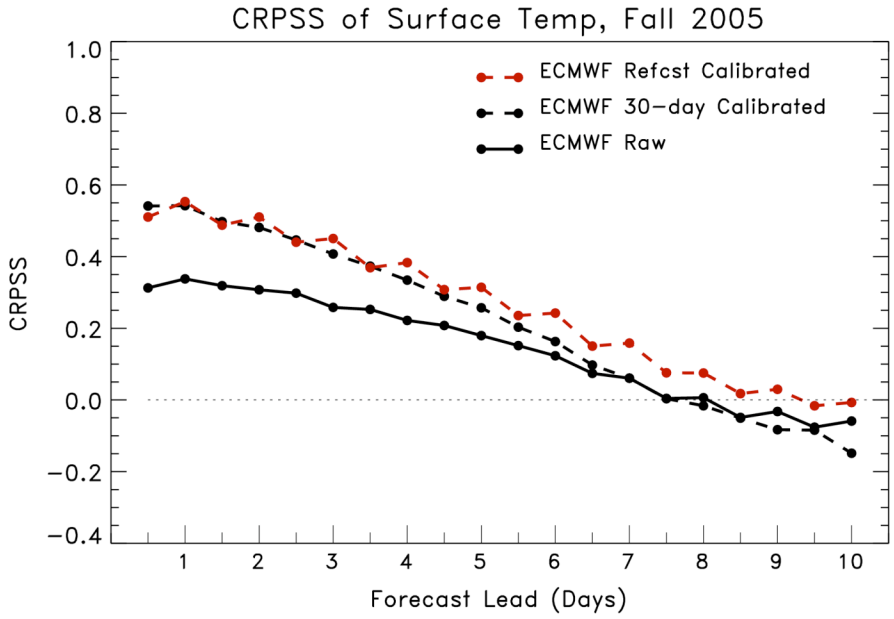
How much from simple bias correction?



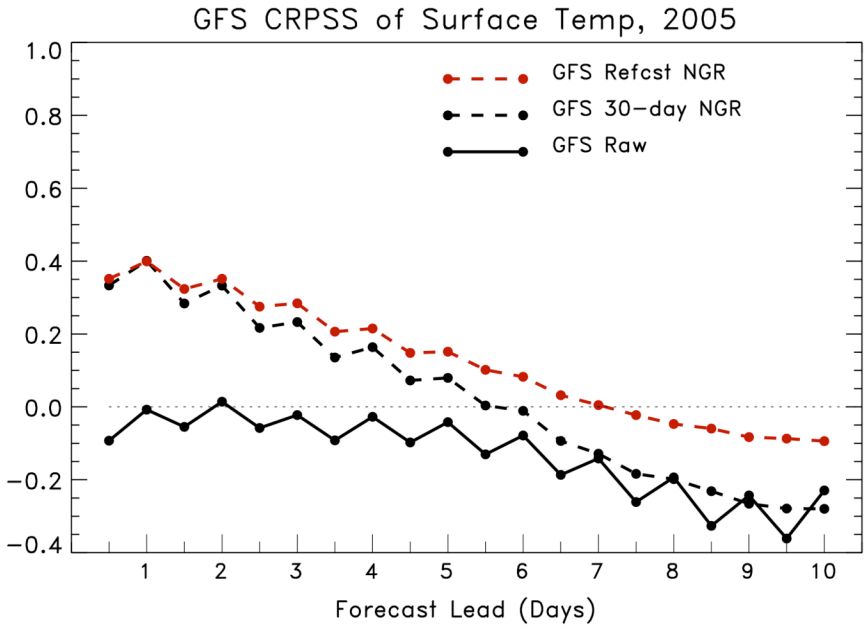
~ 60 percent of total improvement at short leads, 70 percent at longer leads.

How much from short training data sets?

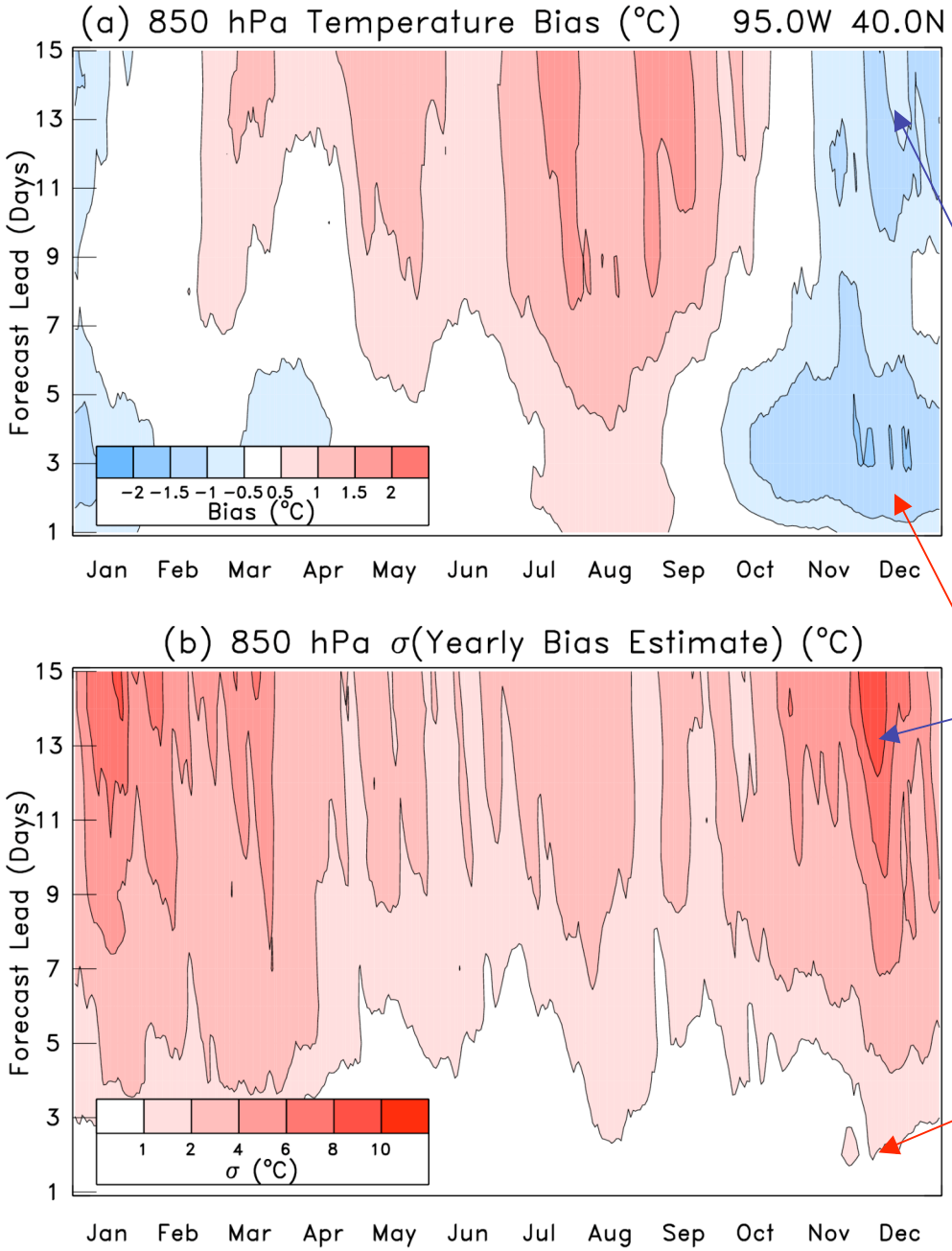
ECMWF



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 T2m; perhaps with, say, soil moisture as additional predictor, reforecast calibration would improve relative to 30-day.



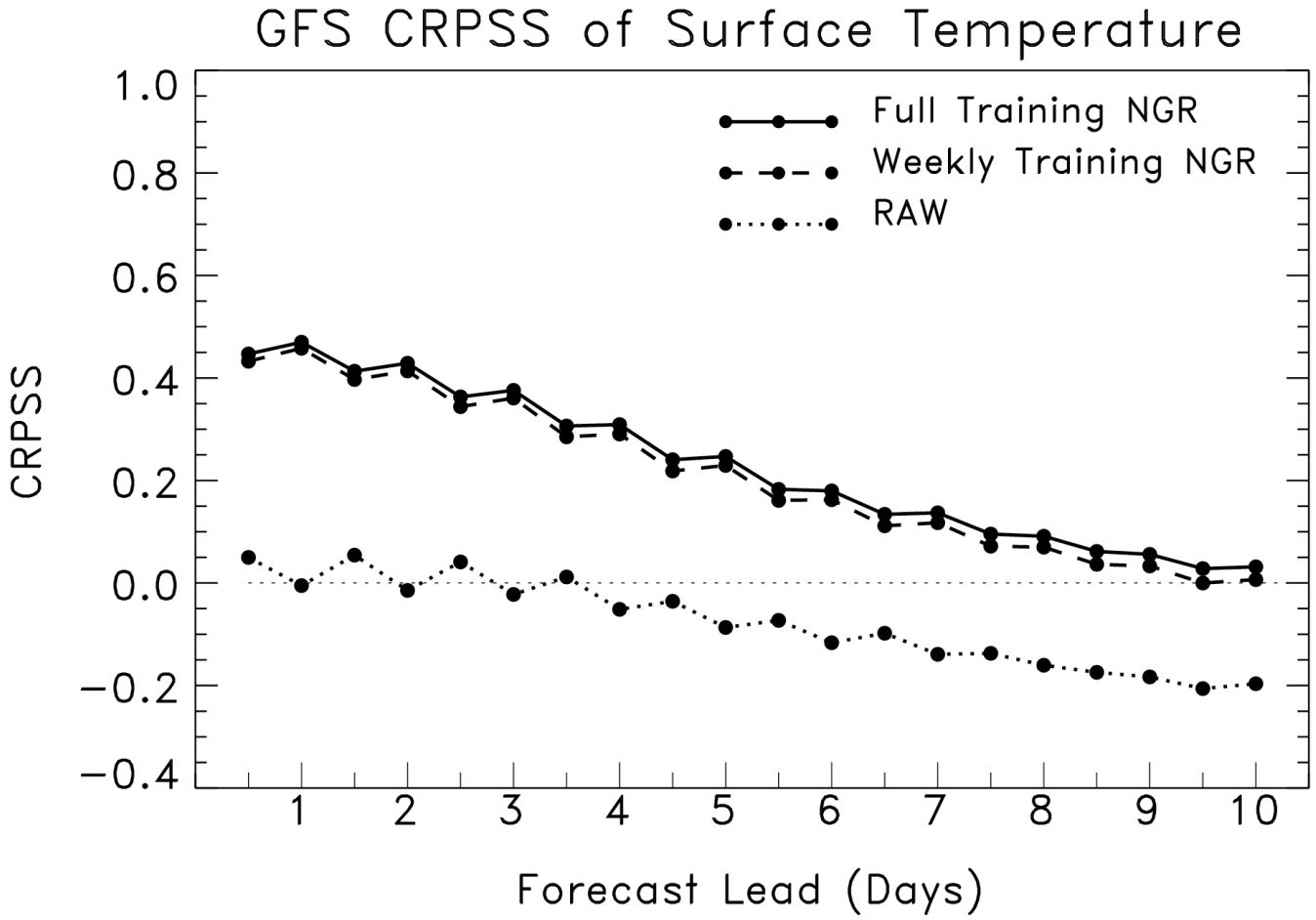
When are long reforecast data sets necessary, and when are they not?

Example: bias correction.

Here, **large training data set required**; bias is small relative to its yearly variability.

Here, **small training data set adequate**; bias comparable or greater than its yearly variability.

How much from long GFS training data set?



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

Precipitation calibration

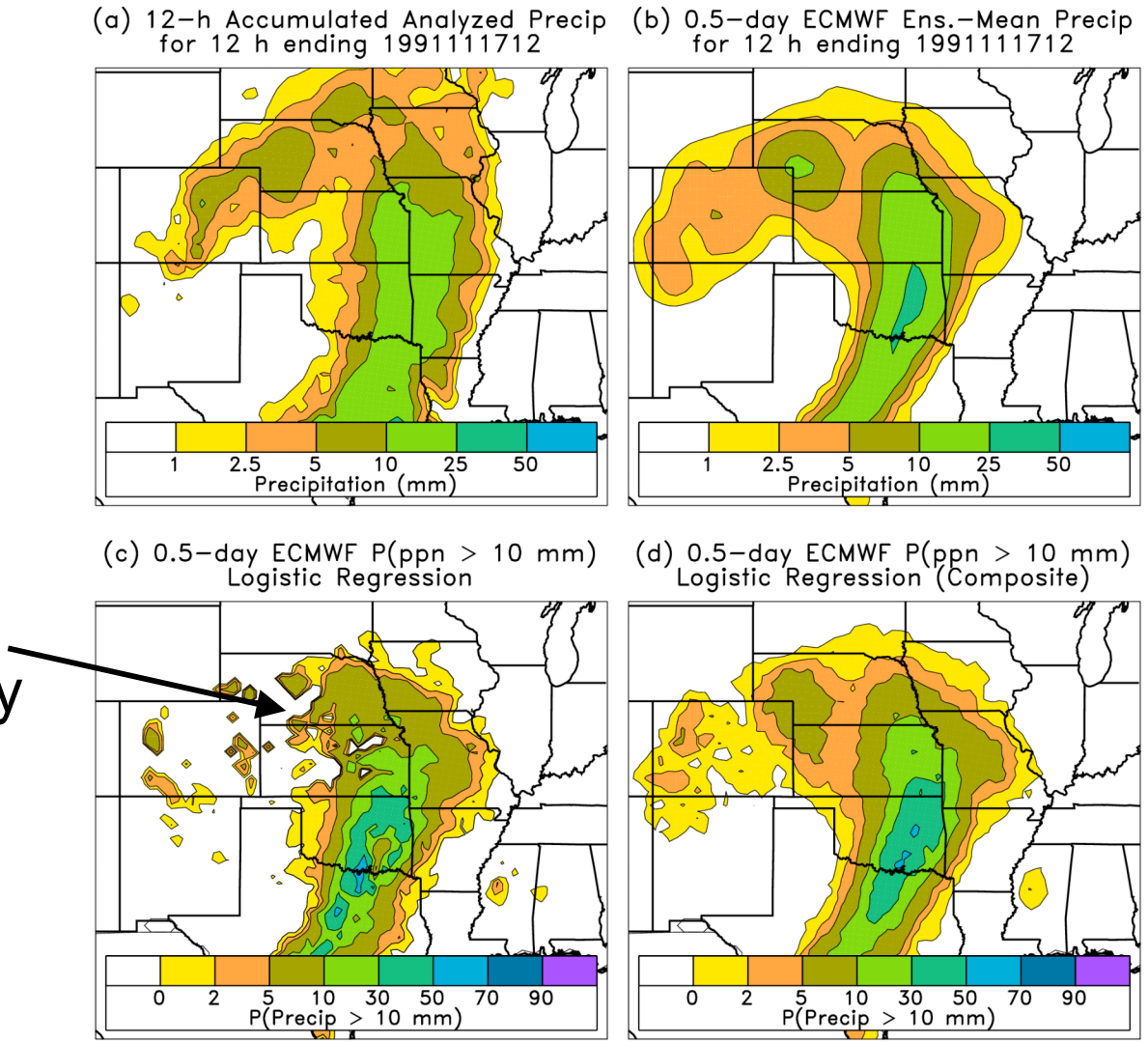
- NARR CONUS 12-hourly data used for training, verification. ~32 km grid spacing
- Logistic regression for calibration here

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

- 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.



When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

Make sure location A is not too close to location B

(2) When observed CDF at A and B are very similar.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar.

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

Need lots of samples. Luckily, ~28 year NARR provides them.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

(2) When observed CDF at A and B are very similar

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

Judging this would be tough with ECMWF forecasts. Only 14 weeks*20 years, not a large sample for non-normally distributed data. Can be fooled by rare events.

When is it proper to use training data at location B to supplement regression analysis at location A?

(1) When location B's errors are independent of location A's errors.

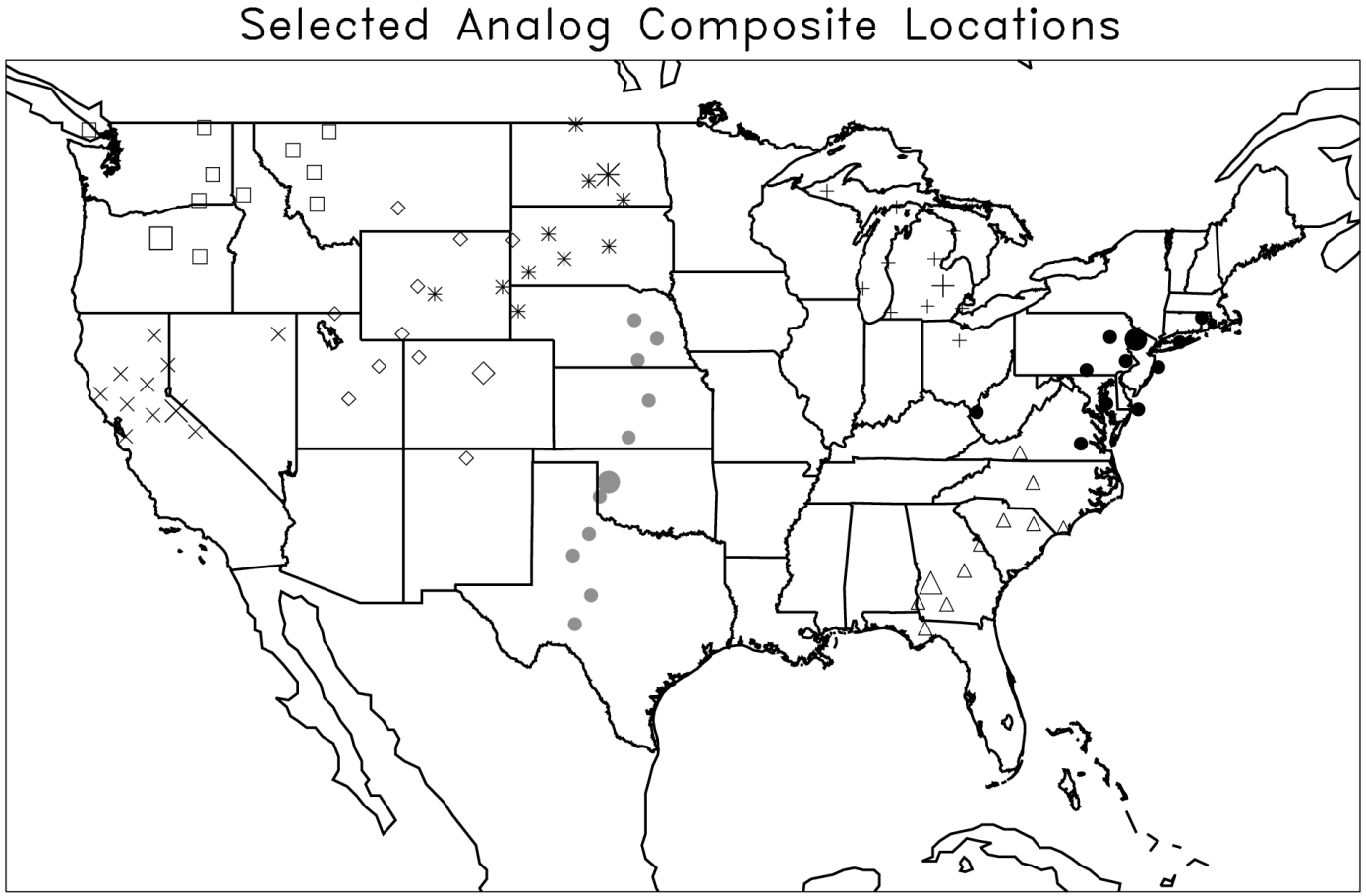
(2) When observed CDF at A and B are very similar

(3) When forecast CDF at A and B are very similar.

(4) When corr(forecast, observed) at A and B are similar.

Tricky to compute in dry regions, where overwhelming bulk of the samples are zero's.

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



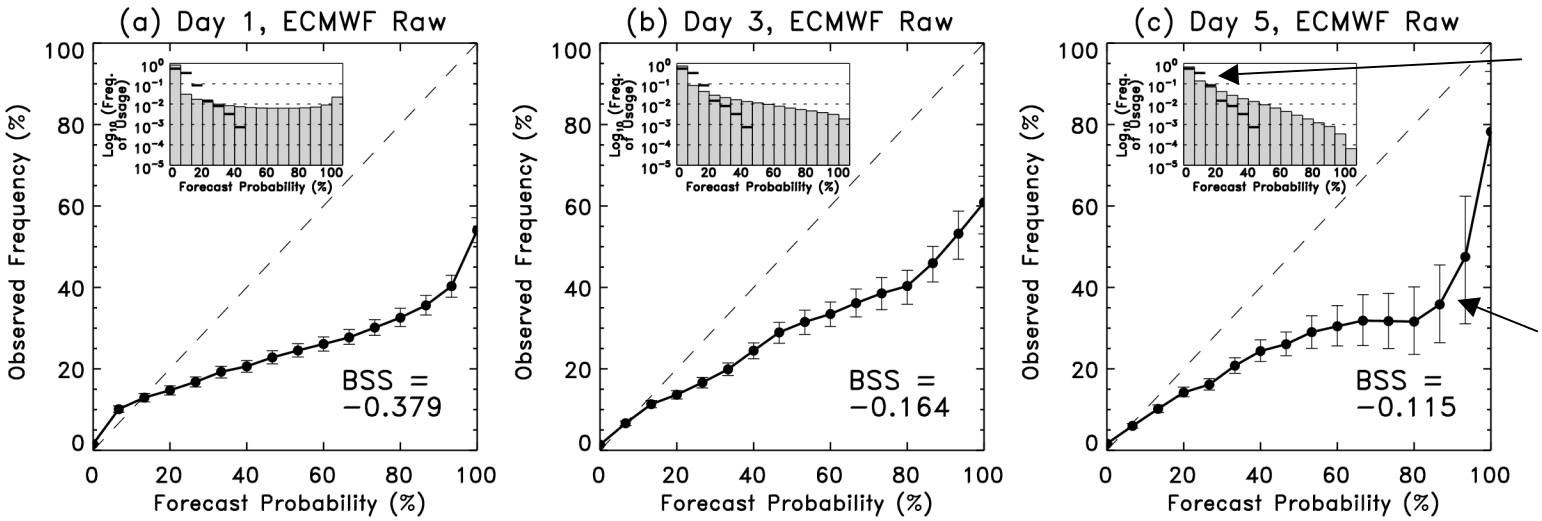
Big symbol: grid point where we do regression

Small symbols: analog locations with similar climatologies

Training data sets tested

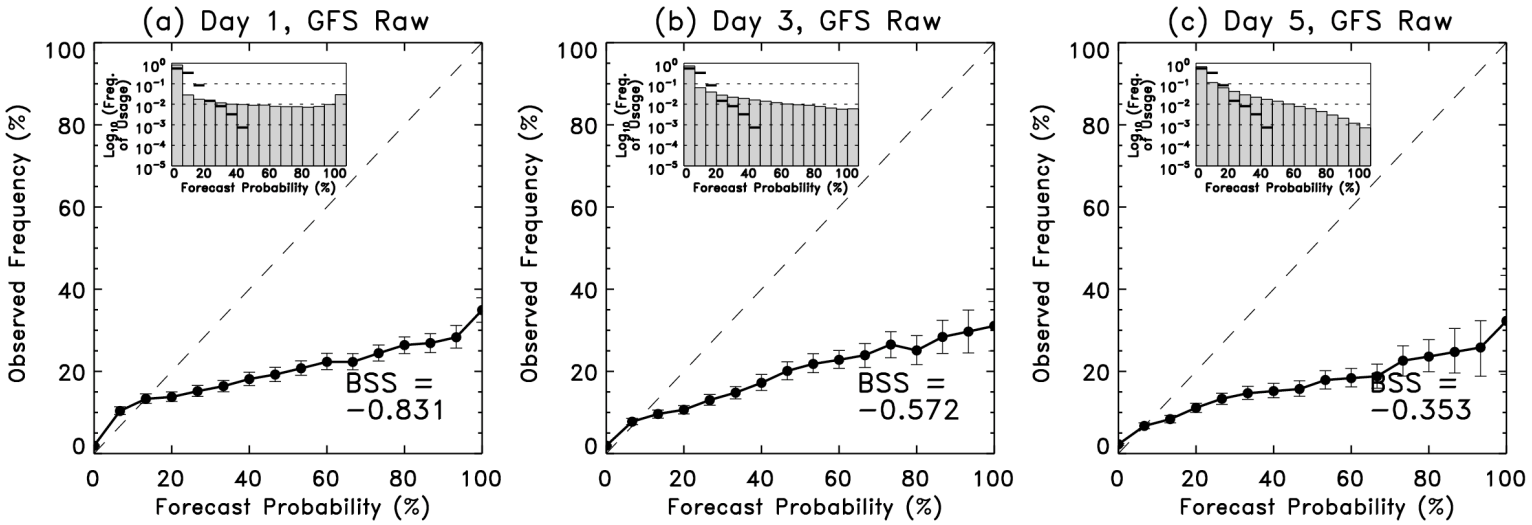
- "Weekly" - use 1x weekly, 20-year reforecasts for training data. Sep-Dec cases all thrown together. X-validated.
- "30-day" - for 2005 only, where forecasts available every day, train using the prior available 30 days.
- "Full" (GFS only) - use 25 years of daily reforecasts. X-validated.

5-mm reliability diagrams, raw ensembles



horizontal lines indicate distribution of climatology

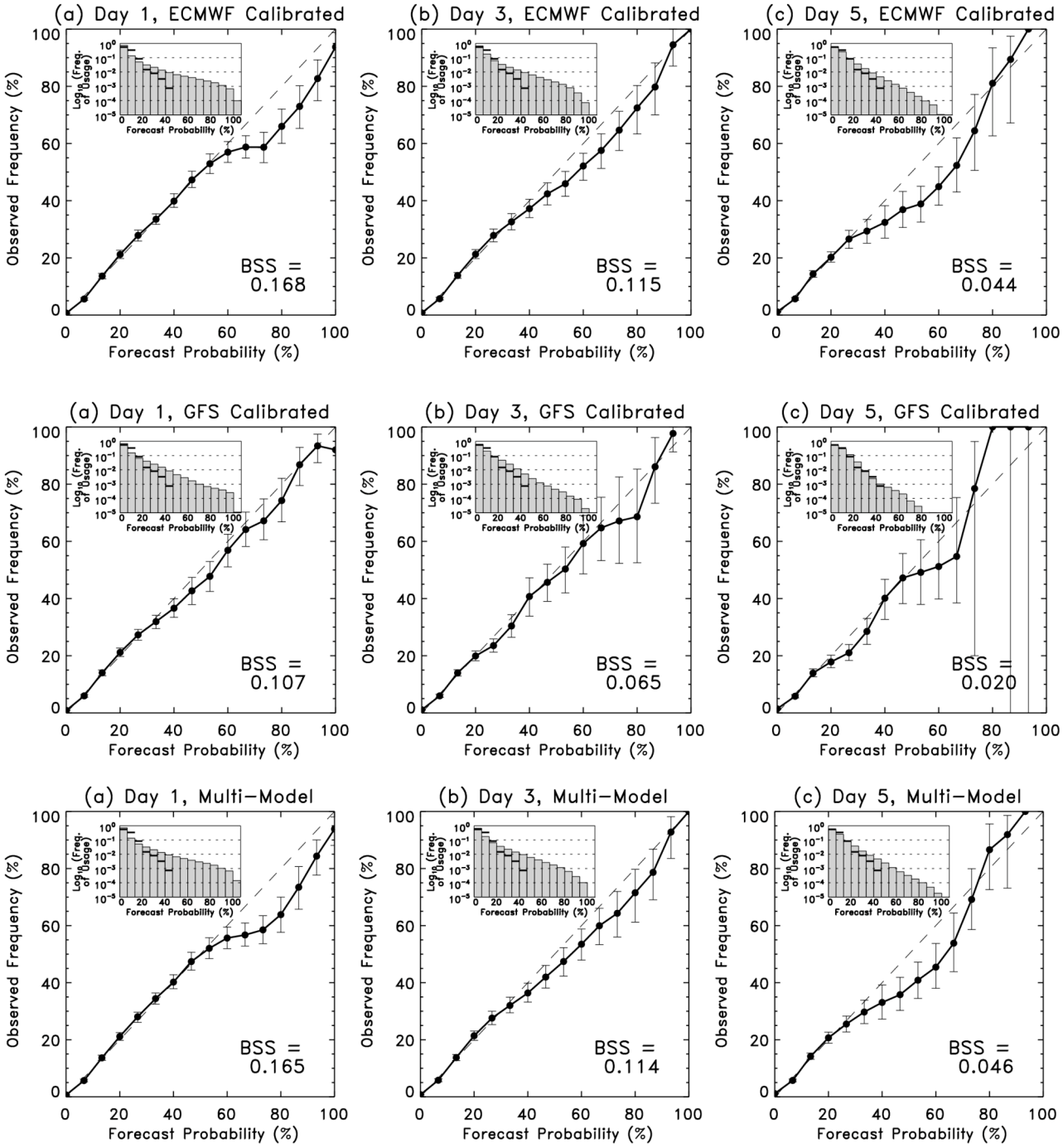
error bars from block bootstrap



Raw forecasts have poor skill in this strict BSS

5-mm reliability diagrams, calibrated

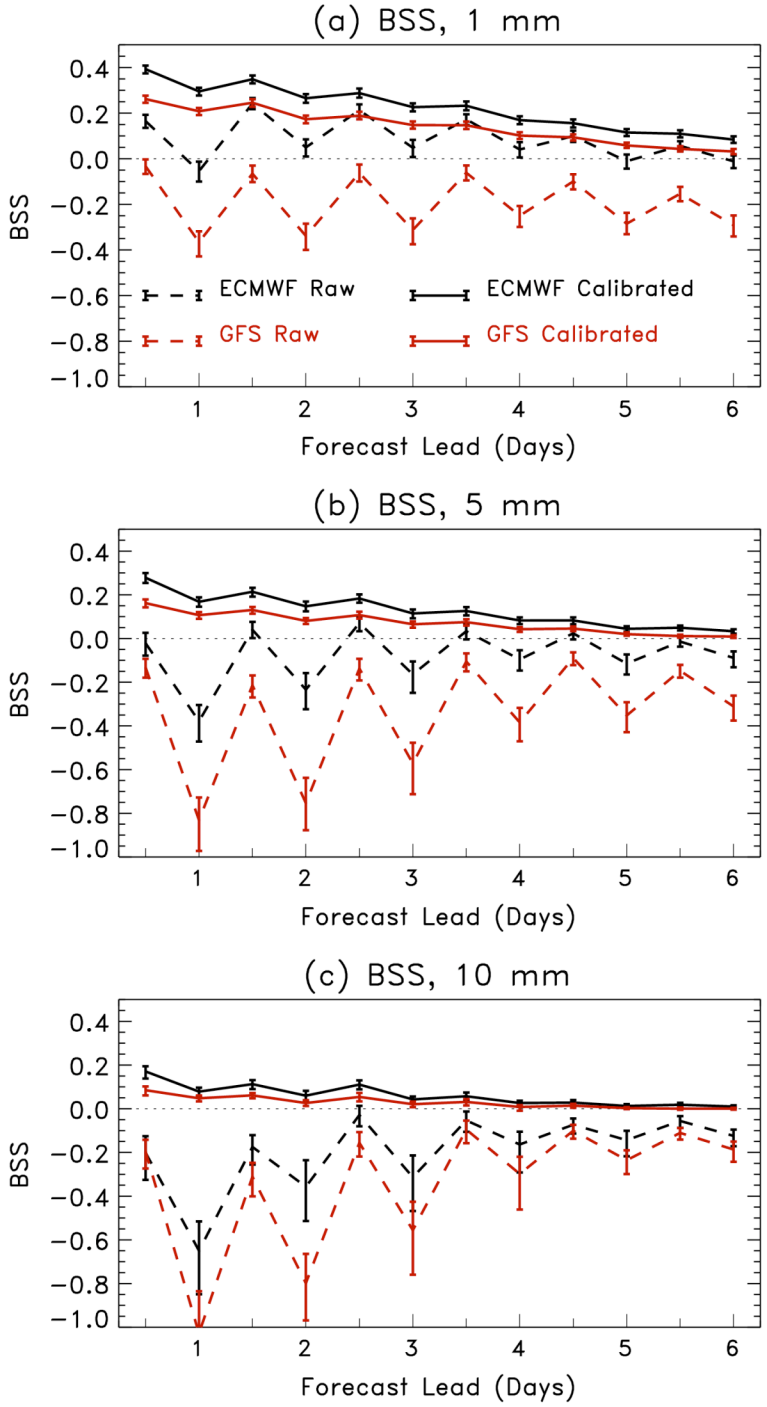
In some respects GFS forecasts look more calibrated but the frequency of usage histograms show ECMWF sharper and thus more skillful.



Brier Skill Scores

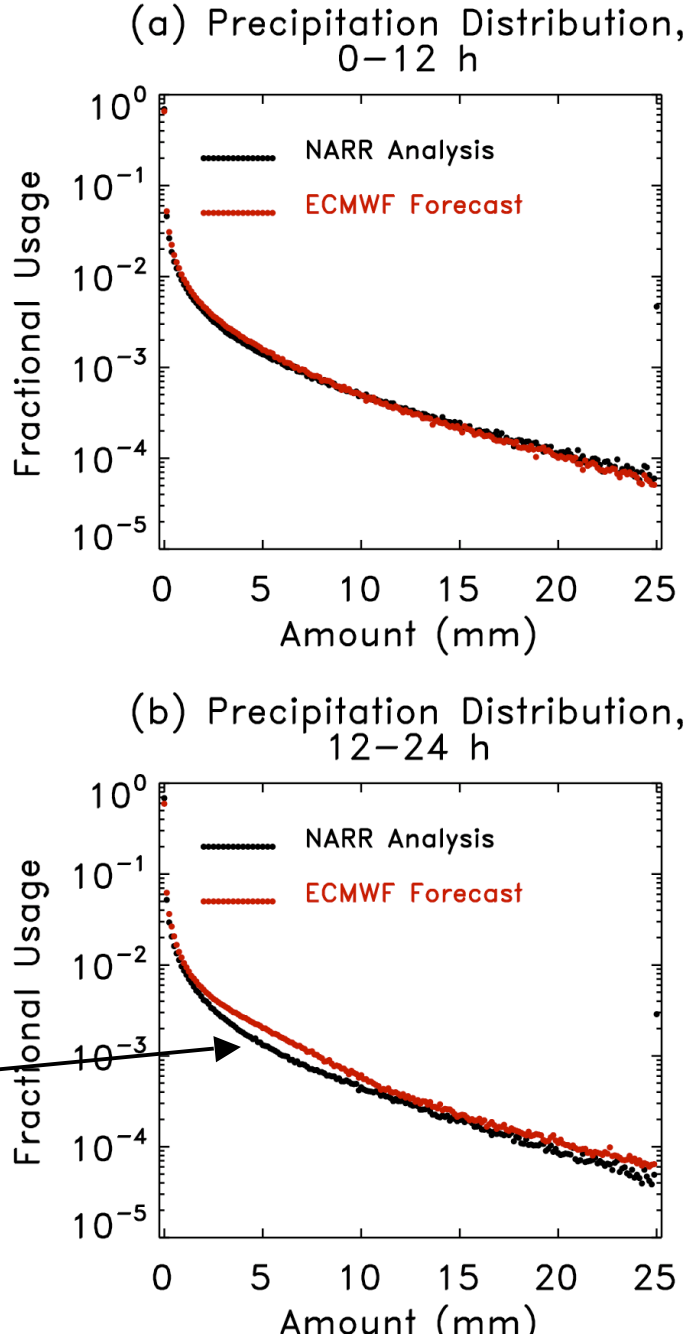
Notes:

(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

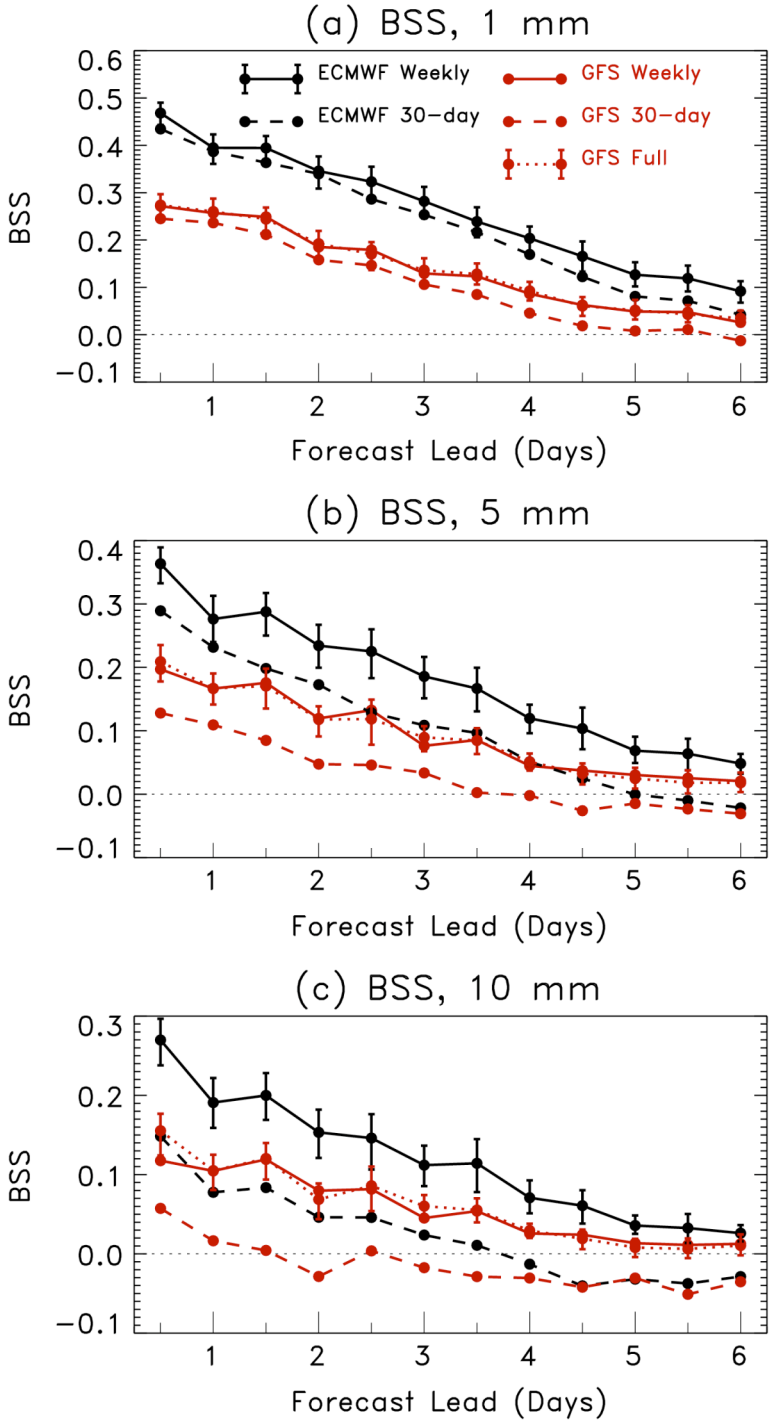


Why are 12Z - 00Z forecasts less skillful?

Over-forecast bias in models during daytime relative to NARR



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

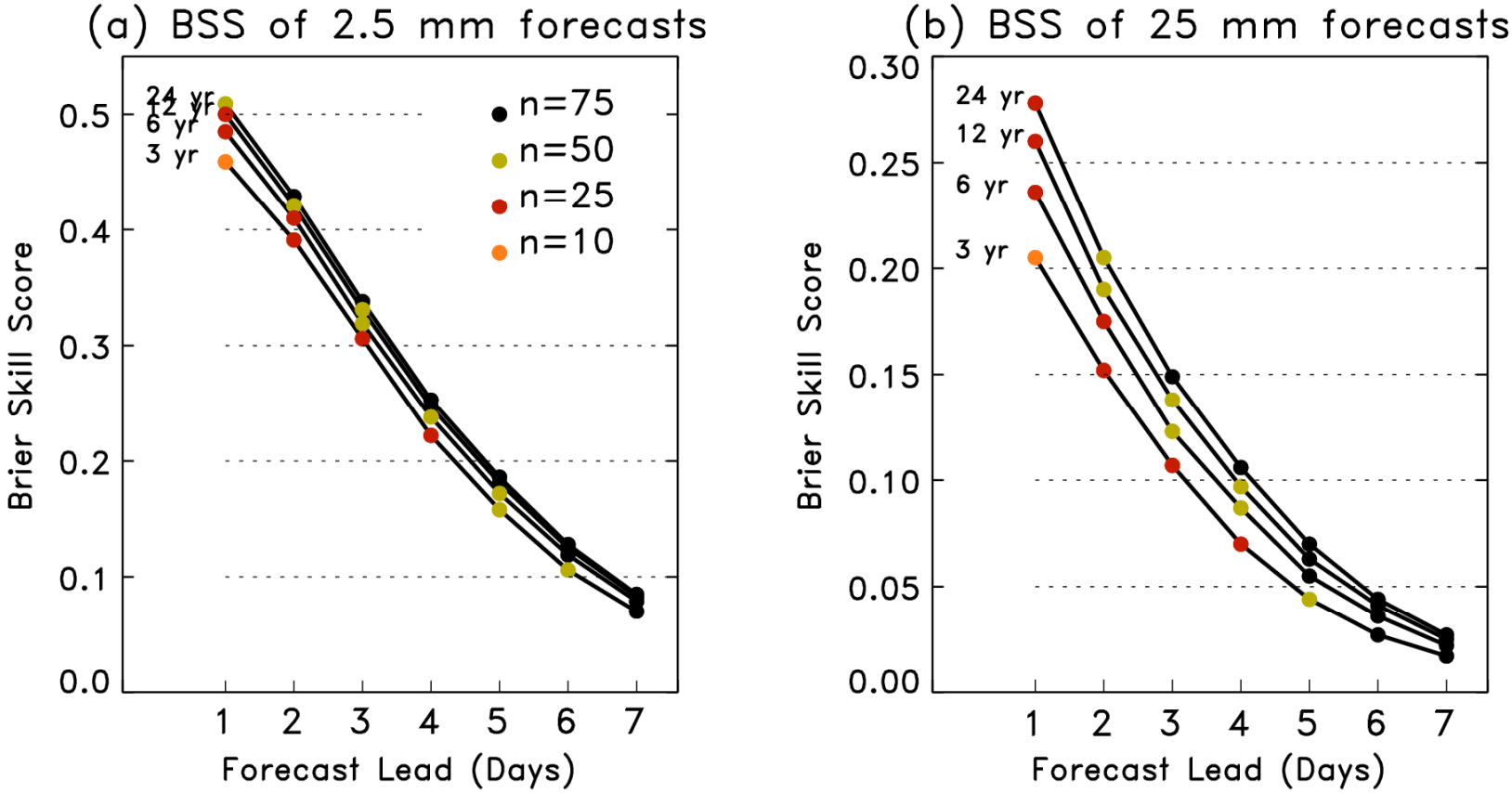


Notes:

(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.

Effect of training sample size: previous results with GFS



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

Adapting reforecasting ideas to calibration of SREF heavy precipitation events

- Q: where are you going to get past initial conditions from:
  - Good: re-run current operational analysis system
  - Worse: use reanalysis from some other model (possibly different initial condition biases
- Q: which subset of cases to run?
  - Good: where forecast indicated heavy precipitation.
  - Bad: where observed indicated heavy precipitation.

Conclusions

- Calibration important, especially for sensible-weather like temperature and precipitation
- Many fairly good calibration techniques, a few that can be problematic.
- Reforecasts shown to aid in calibration of forecasts for a wide variety of applications
- 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.
    - Skill when trained with daily data not much larger than when trained with weekly data (preliminary result, more testing needed).

Are operational centers heading toward reforecasting?

- **NCEP**: tentative plans for 1-member real-time reforecast.
- **ECMWF**: once-weekly, real-time 5-member reforecast starting mid 2008.
- **RPN Canada**: planning ~5-year reforecast data set, delayed by budget and staffing issues.

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*., in press. 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*., in press. Available at http://tinyurl.com/38jgkv

- (and many reforecast references therein)

Acknowledgments

(1) Renate Hagedorn, ECMWF. Valuable colleague

(2) ECMWF, for providing data.

(3) Jeff Whitaker, Gary Bates, Xue Wei (NOAA)