Comparison of GFS/EnKF ensemble precipitation forecasts for Jul-Oct 2010 over CONUS relative to NCEP, ECMWF, UKMO, and CMC operational ensembles

Tom Hamill
NOAA ESRL, Physical Sciences Division
tom.hamill@noaa.gov
(303) 497-3060

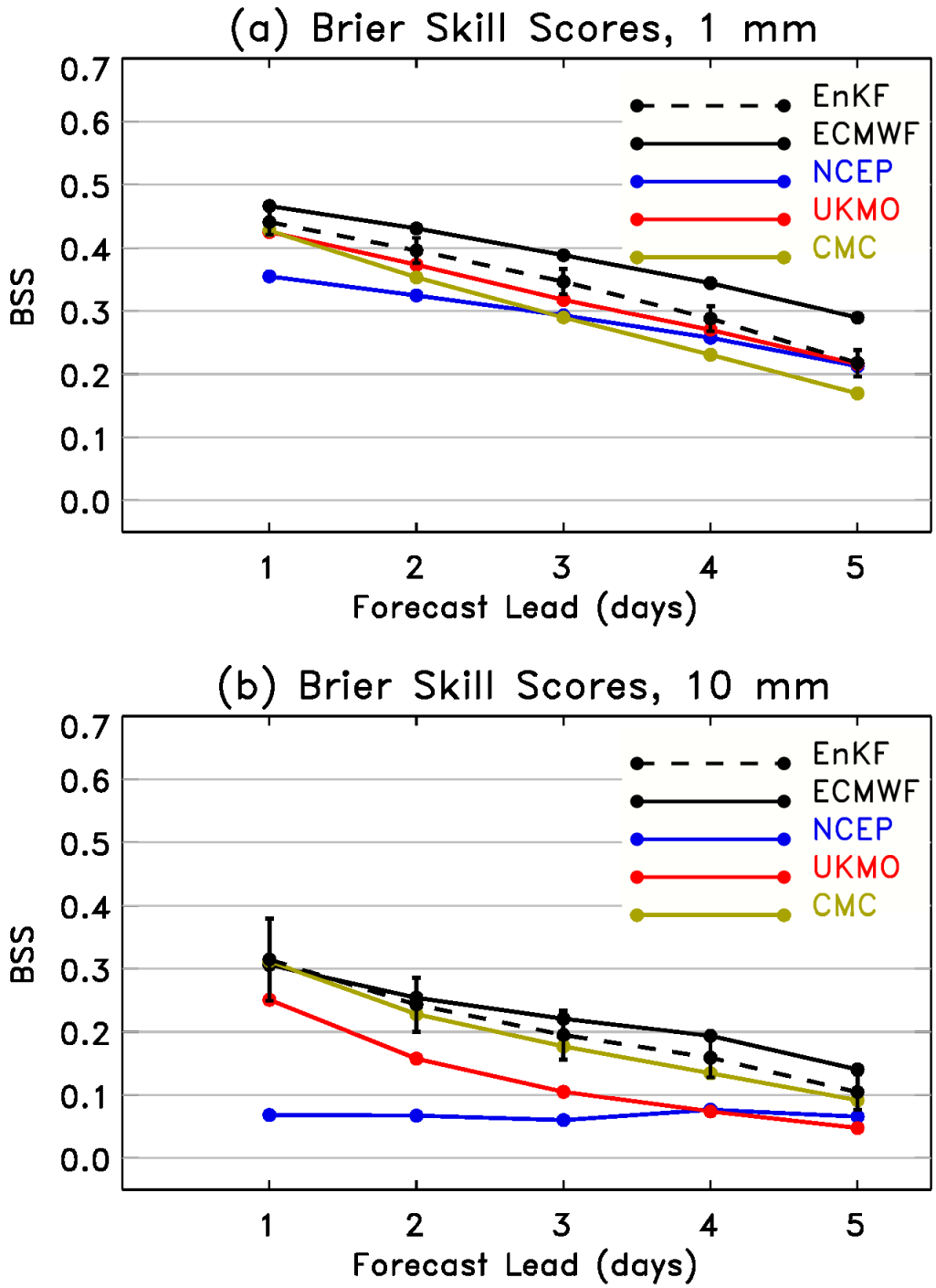
(thanks to Jeff Whitaker for support)

Main points

- We examined precipitation forecasts from T254 GFS/EnKF and other 20-member operational ensembles (NCEP, ECMWF, CMC, UKMO) run during Jul-Oct 2010.

- Validated against EMC-derived "CCPA" 1-degree analysis of daily precipitation, which uses regression correction of Stage-IV to CPC analysis (Luo et al. 2010).

- Key results
  - Better (consistently better) Brier skill scores in EnKF with new GFS relative to NCEP operational.
  - Less sharp (less binary) probabilistic forecasts.
  - Better equitable threat scores in EnKF with new GFS, too.



One teaser: improved probabilistic skill scores

- EnKF + T254 new GFS narrows the difference between NCEP and ECMWF, especially at higher amounts.
- Confidence interval 2.5% to 97.5% based on paired t-test between EnKF and NCEP, including correction for sample size due to autocorrelation.

Precipitation verification

- Use EMC "CCPA" dataset of Stage-IV precipitation, regression-corrected to CPC analysis over CONUS, and upscaled to 1-degree. Described in Luo et al. (2010).
- Verify only where 1-degree box is within CONUS.
- Verify 24-h accumulations.
- Verify forecasts initialized from both 00Z and 12Z.

GFS/EnKF configuration

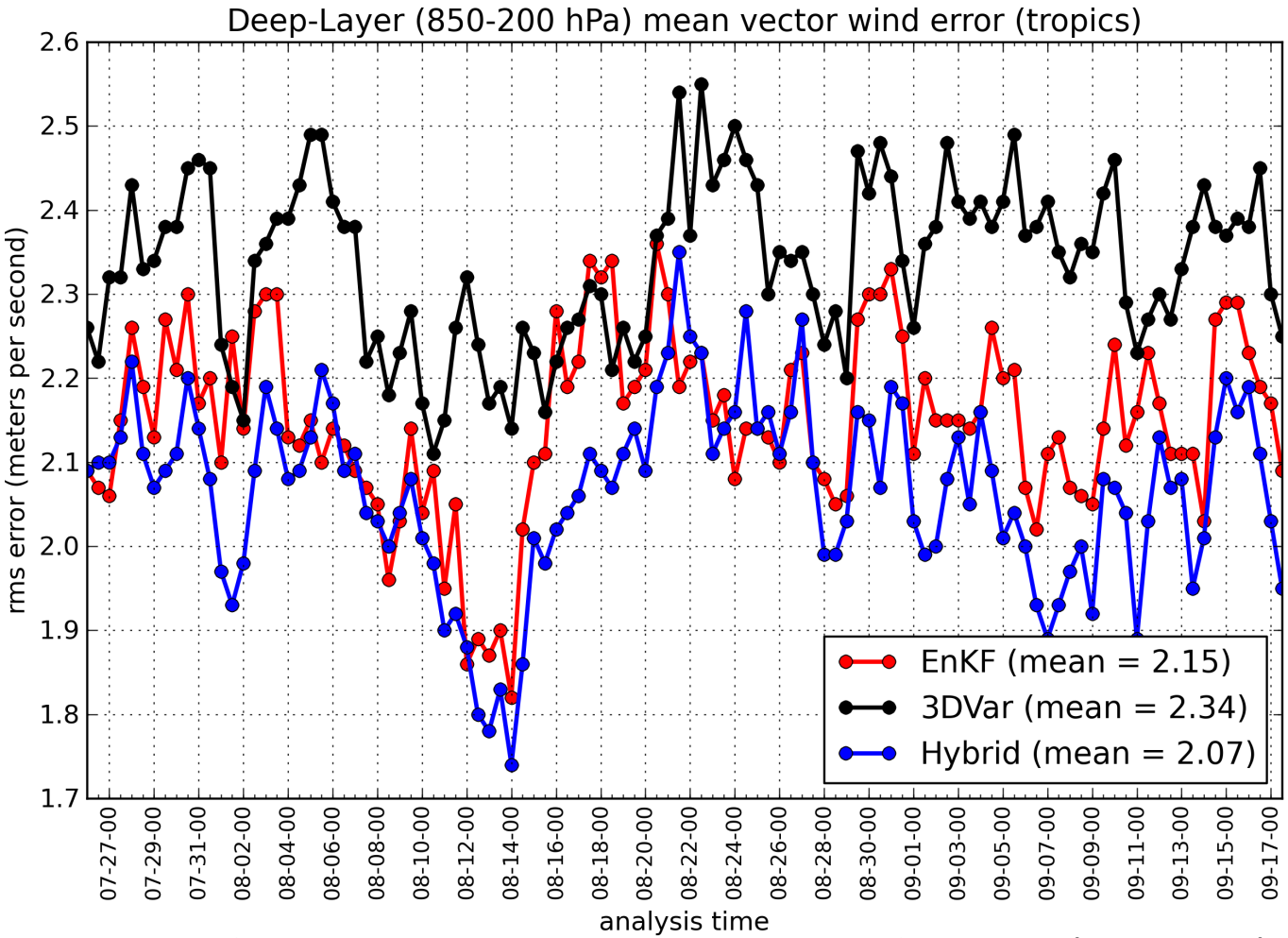
- 2010 version 9.01 of GFS; T254L64, 80 members cycled for EnKF assimilation
  - Assimilate all observations used in GSI + TCVitals sea-level pressure observation.
  - Run in quasi-realtime during hurricane season; leveraging these forecasts.
- Hurricane verification discussed at **http://tinyurl.com/4kzamb3** , AMS Annual '11
- Thereafter, 20 perturbed member forecasts + control computed to 5 days lead.
- The EnKF here approximates that which will be used in GSI-EnKF hybrid; we are working with data assimilation group members, e.g., Kleist, Derber, Parrish, aiming for April 2012 implementation.

Are improvements due to the EnKF?

- In following comparisons of NCEP and EnKF, there are several differences:
  - Version of GFS used. T254 EnKF uses the latest (9.01), operational T190 ensemble older version (version 8.0).
  - Model resolution (T254L64 for EnKF, T190L28 for NCEP ensemble).
  - Different control analysis around which perturbed initial conditions are centered. GSI for NCEP, EnKF mean for EnKF.
  - Different structures of the perturbed initial conditions.
- We don't know yet for sure which provides the greatest benefit; some results suggest EnKF mean may provide substantial benefit, others suggest GFS changes more important than data assimilation (next 2 slides).

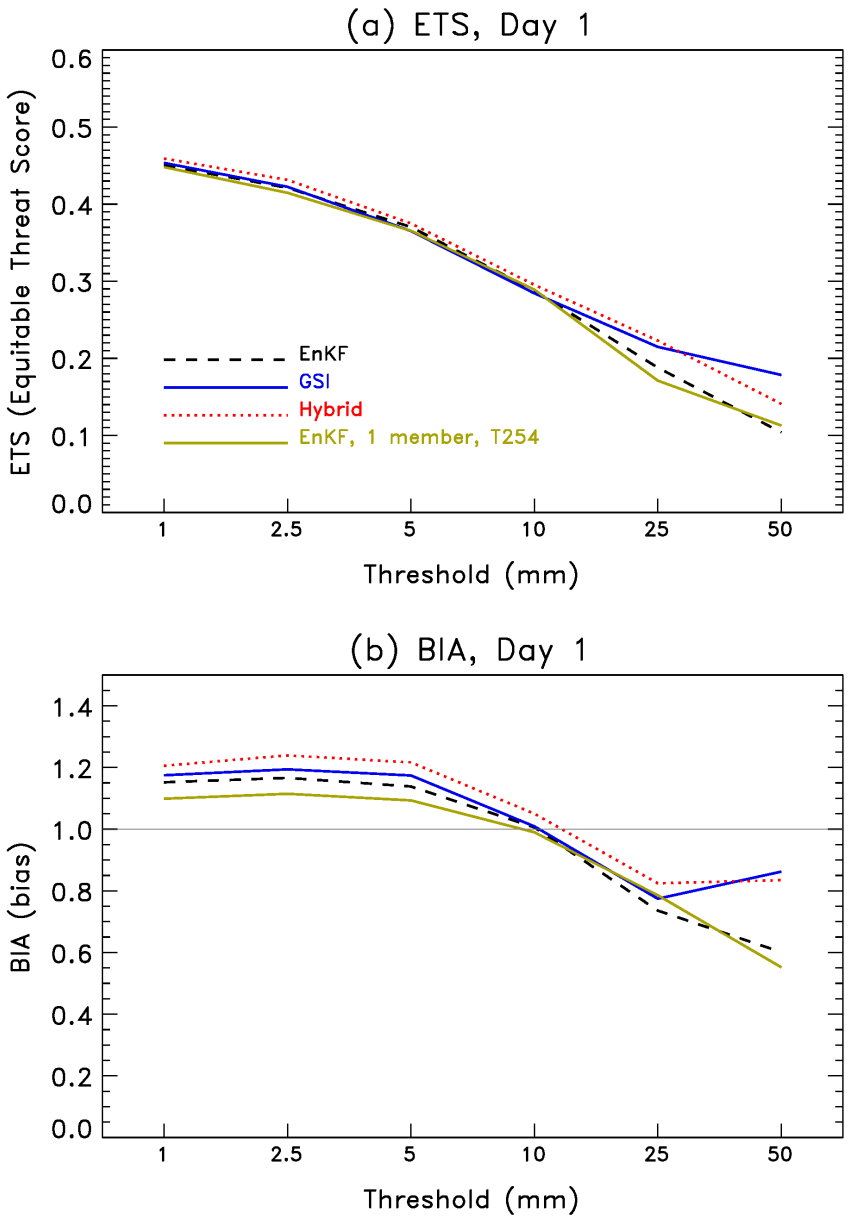
72-h (deterministic) wind errors vs. ECMWF analysis

c/o Jeff Whitaker & Daryl Kleist



Same version of GFS used in each deterministic forecast (T574L64).
If ECMWF analysis can be regarded as truth, hybrid and EnKF are consistently much lower RMS than GSI.

Threat scores of T574 deterministic forecasts



For these day+1 forecasts (similar results for days +3 and +5), we examine ETS and BIA for control runs with various data assimilation systems using the lastest GFS model. The GSI used a T574 forecast model and the operational GSI assimilation system. The EnKF took the mean analysis from the T254 EnKF system and initialized T574 control forecasts. The hybrid used the T254 EnKF forecasts in the T574 GSI hybrid system. Forecasts here are from July through mid-September 2010.

There appears to be little difference in threat scores. We do not yet have a parallel run of the full ensemble system (a T254 GSI+ETR with the same version of the GFS) to compare ensemble products.

Other ensemble systems (via TIGGE)

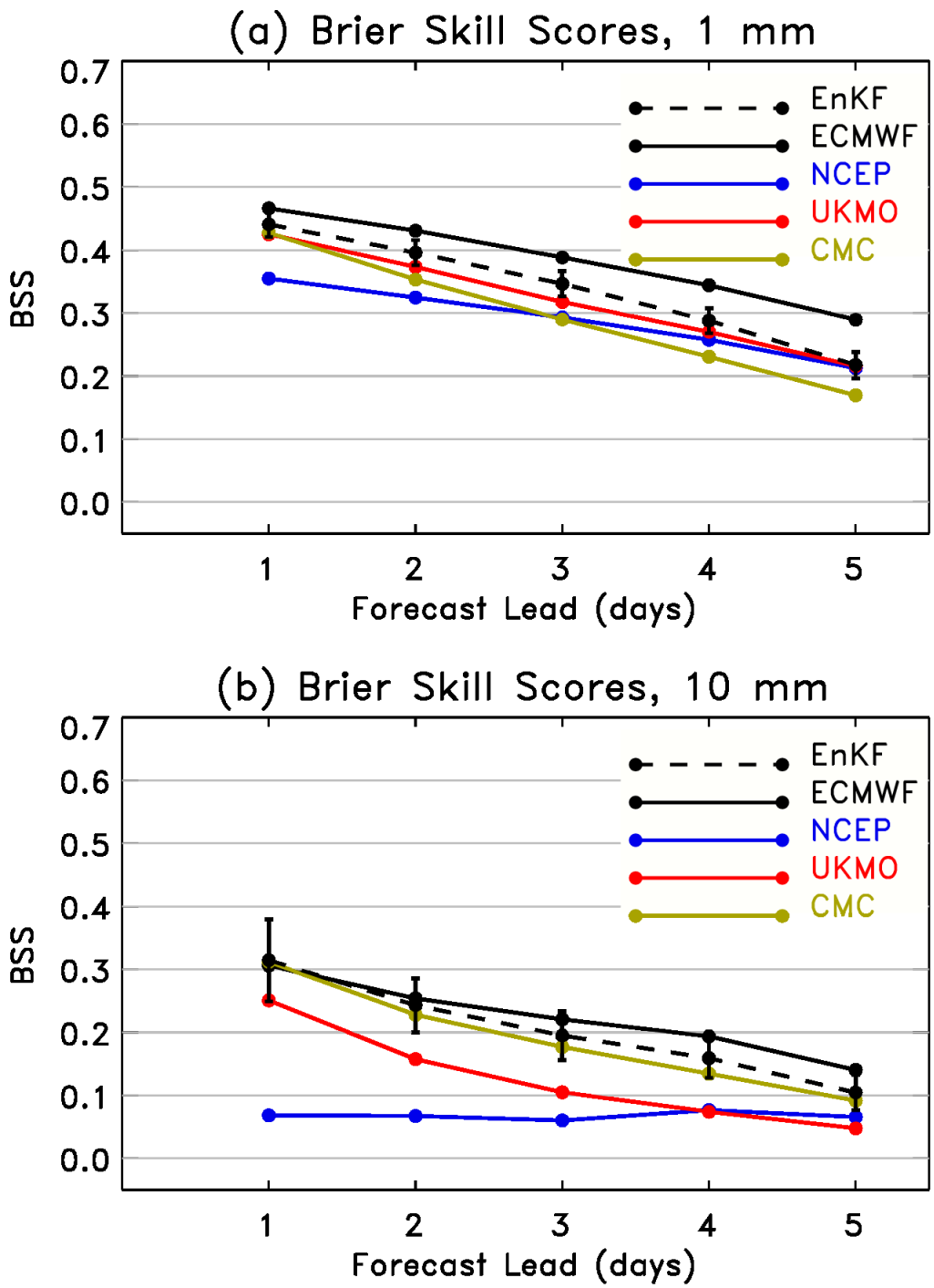
- NCEP operational, 20 members, T190L28. Uses older version of GFS
- ECMWF operational, first 20 members, T639L80
- UK Met Office, 20 members
- CMC, 20 members

Verification techniques

- CRPS, continuous ranked probability score
- Brier skill scores
- Reliability diagrams
- ETS (Equitable Threat Score) & Bias
- Sharpness plots

- Details in supplementary slides, here.

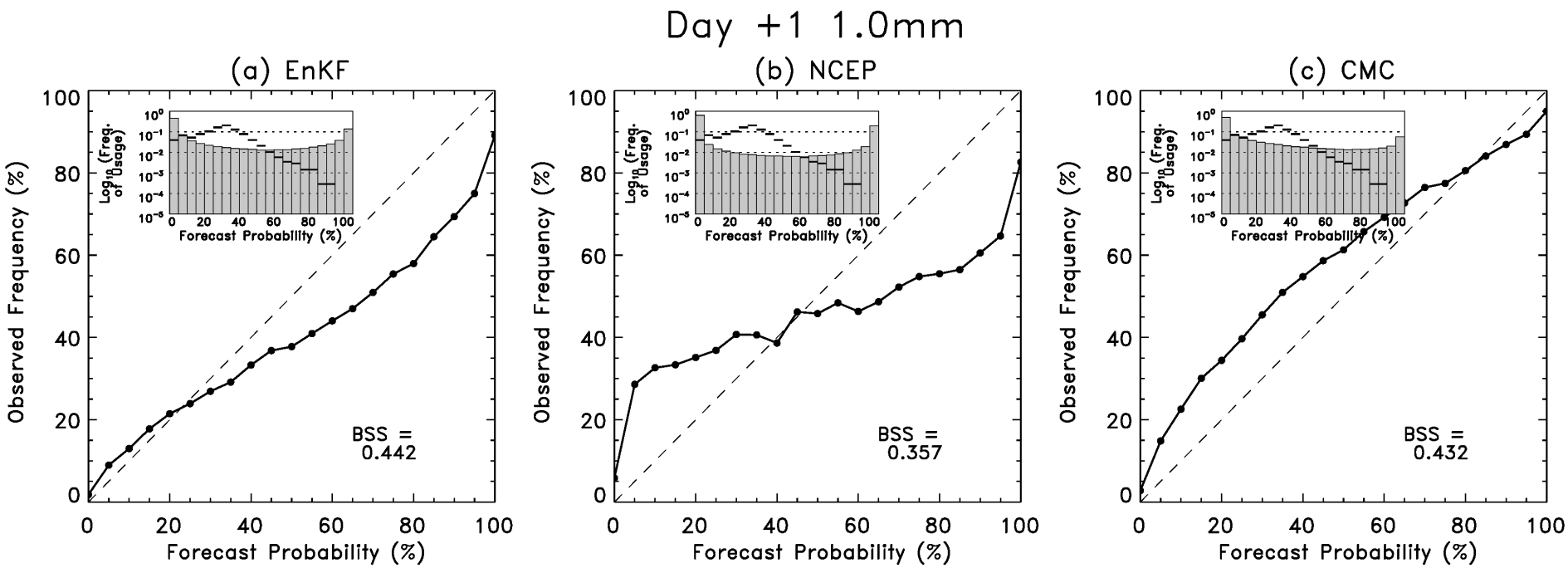
Results



Brier skill scores

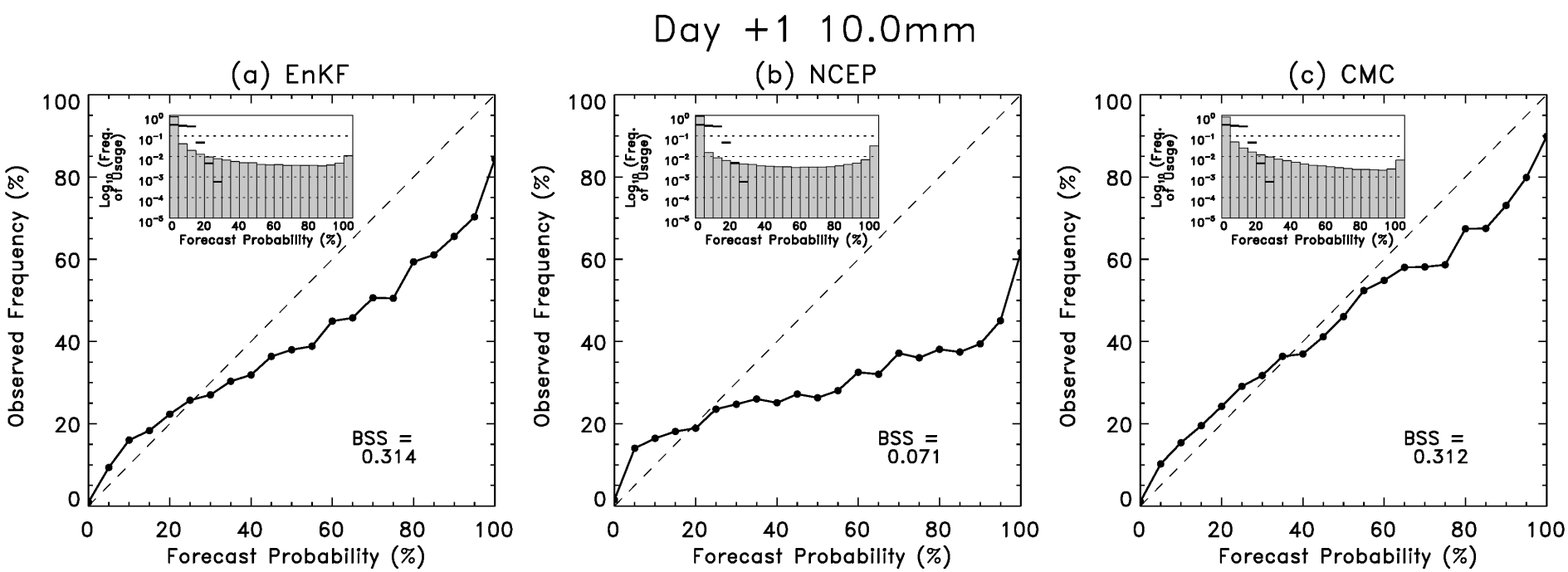
- EnKF + T254 + new GFS narrows the difference between NCEP and ECMWF, especially at higher amounts.
- Confidence interval 2.5% to 97.5% based on paired t-test between EnKF and NCEP, including correction for sample size due to autocorrelation.

1-mm reliability, day +1



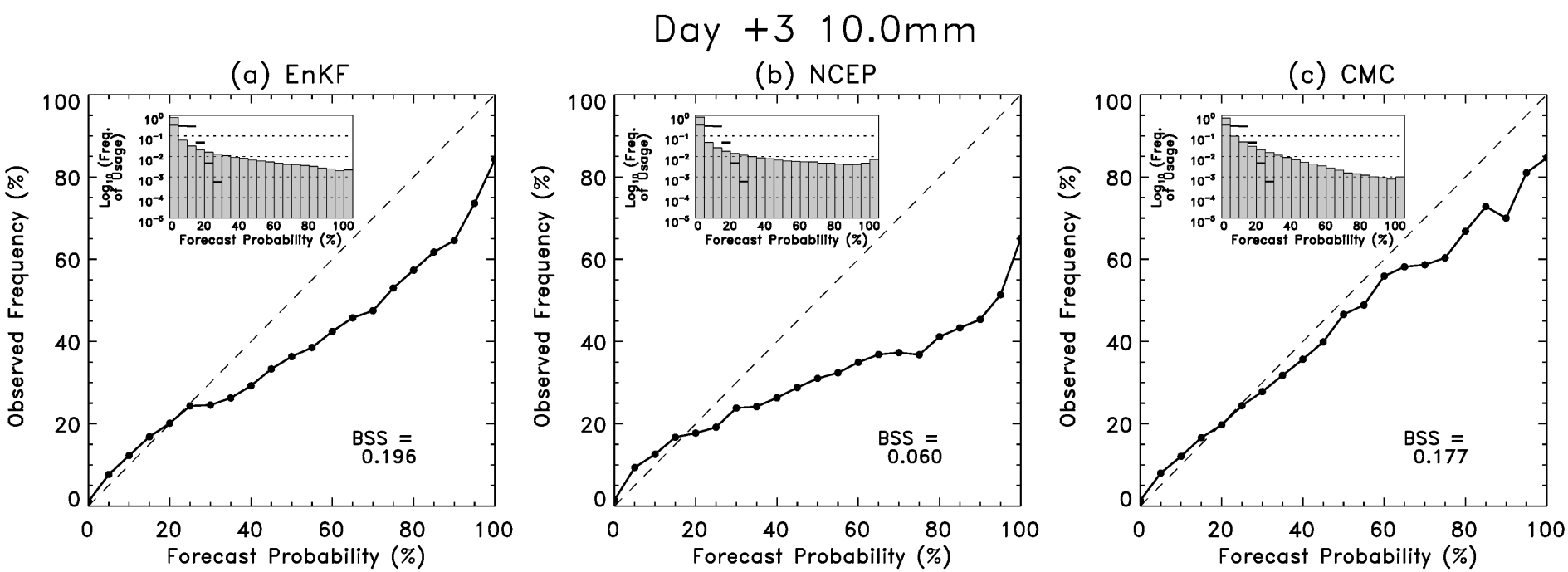
Solid lines on inset frequency of usage histogram are the frequency of usage of the climatological probabilities across the CONUS and over the 4-month period.

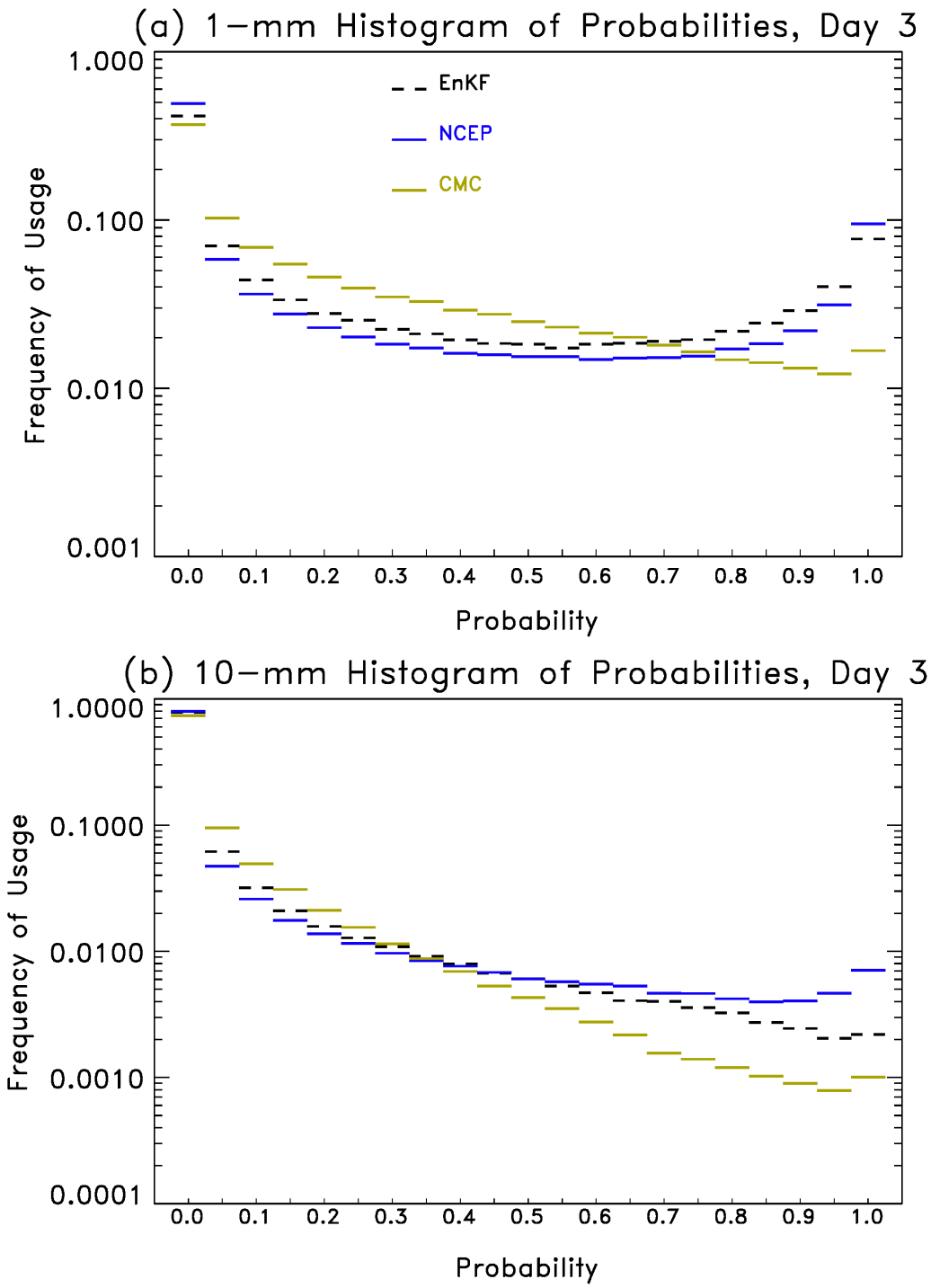
10-mm reliability, day +1



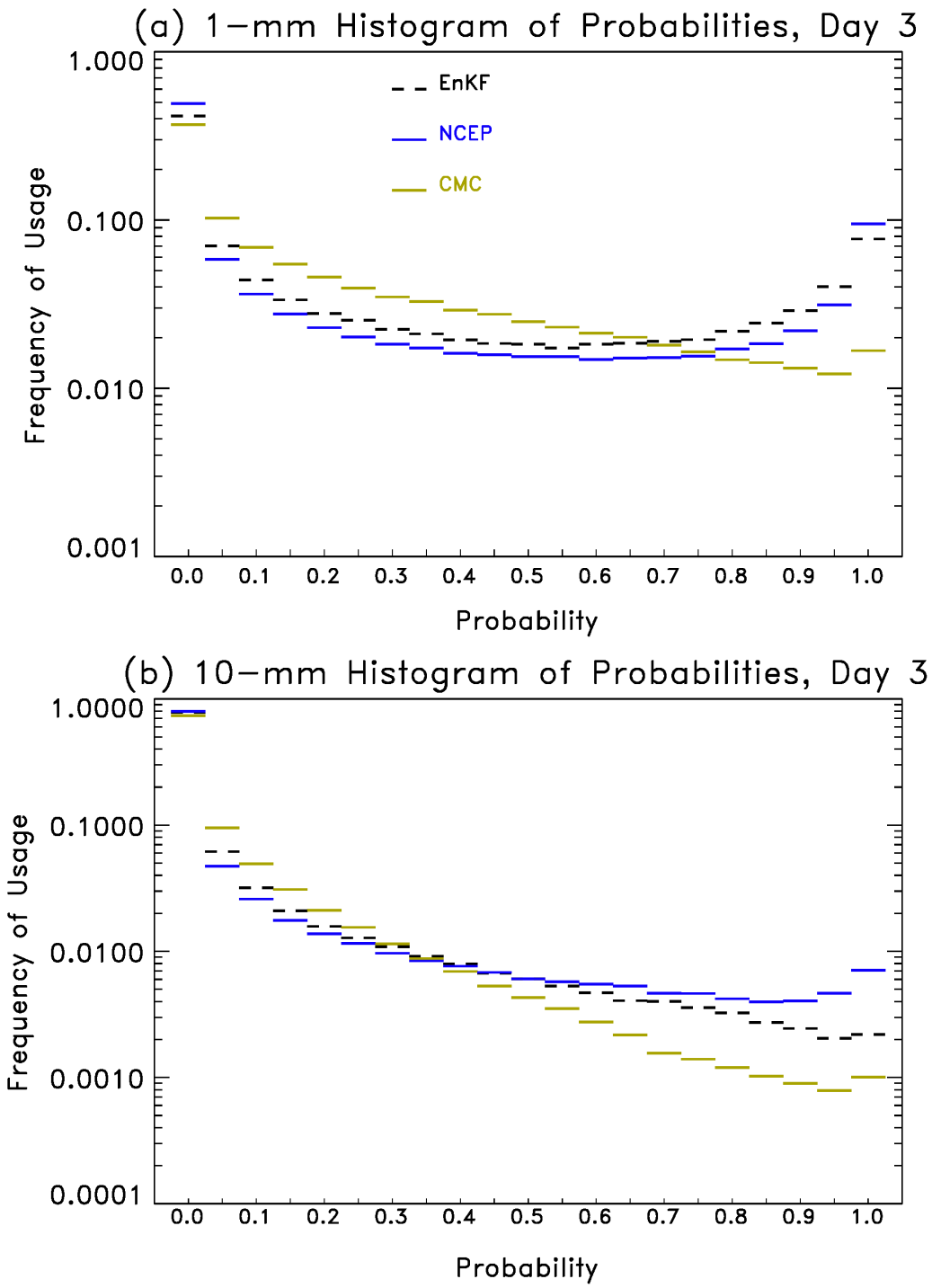
(in between) (sharpest) (most reliable)

10-mm reliability, day +3

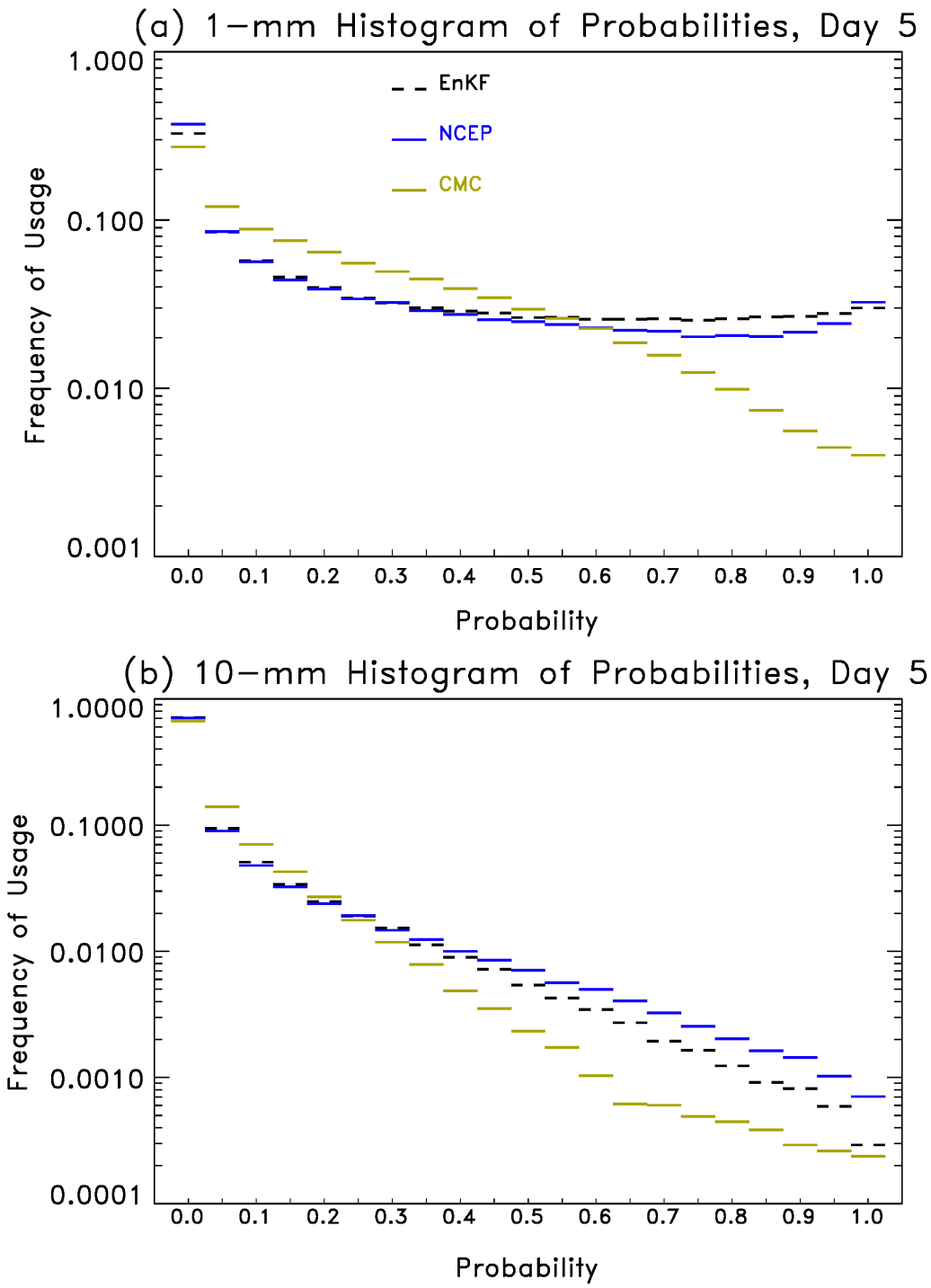




Sharpness of the probabilistic precip. forecasts, day +1

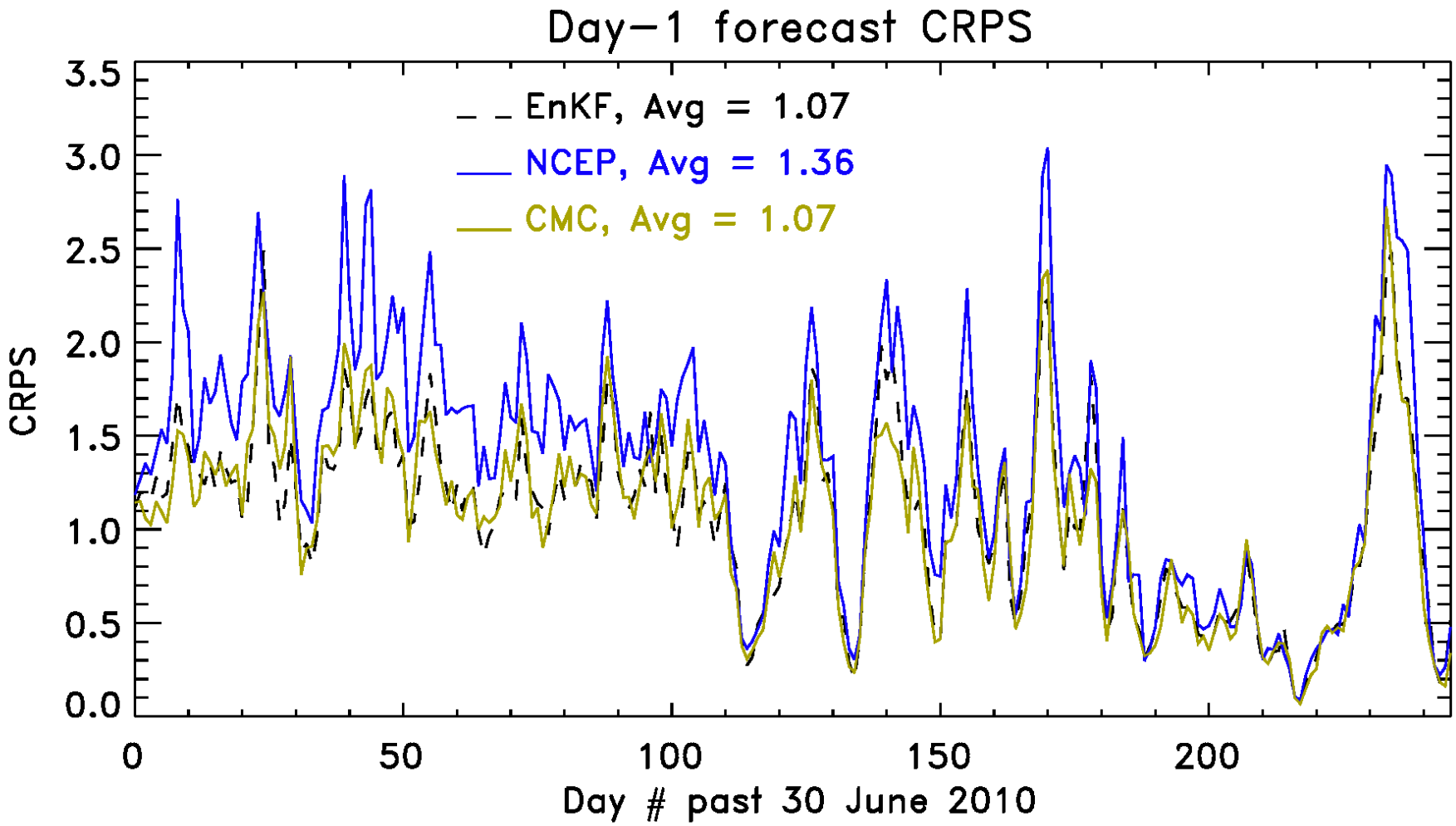


Sharpness of the probabilistic precip. forecasts, day +3



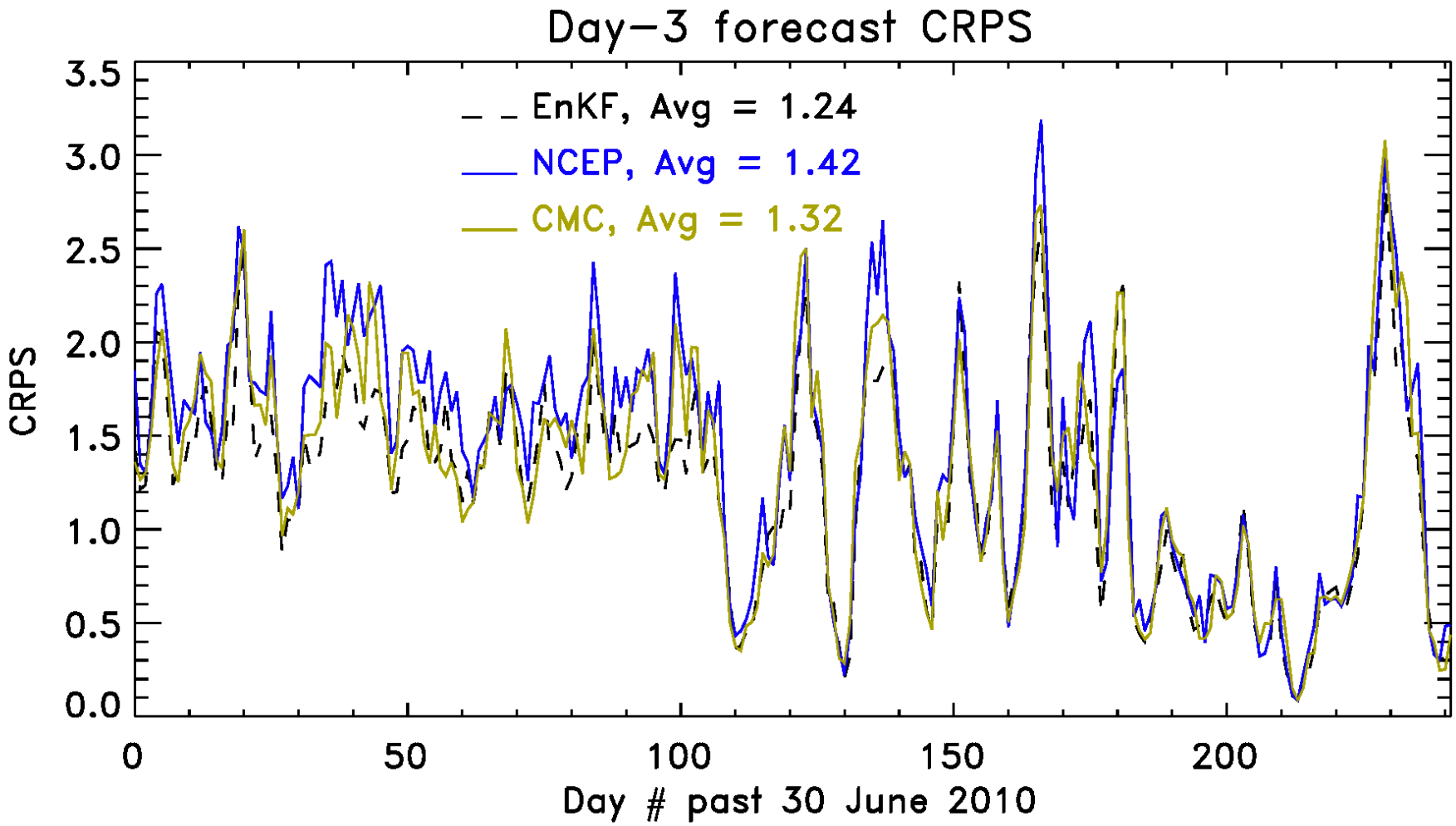
Sharpness of the probabilistic precip. forecasts, day +5

Continuous ranked probability scores



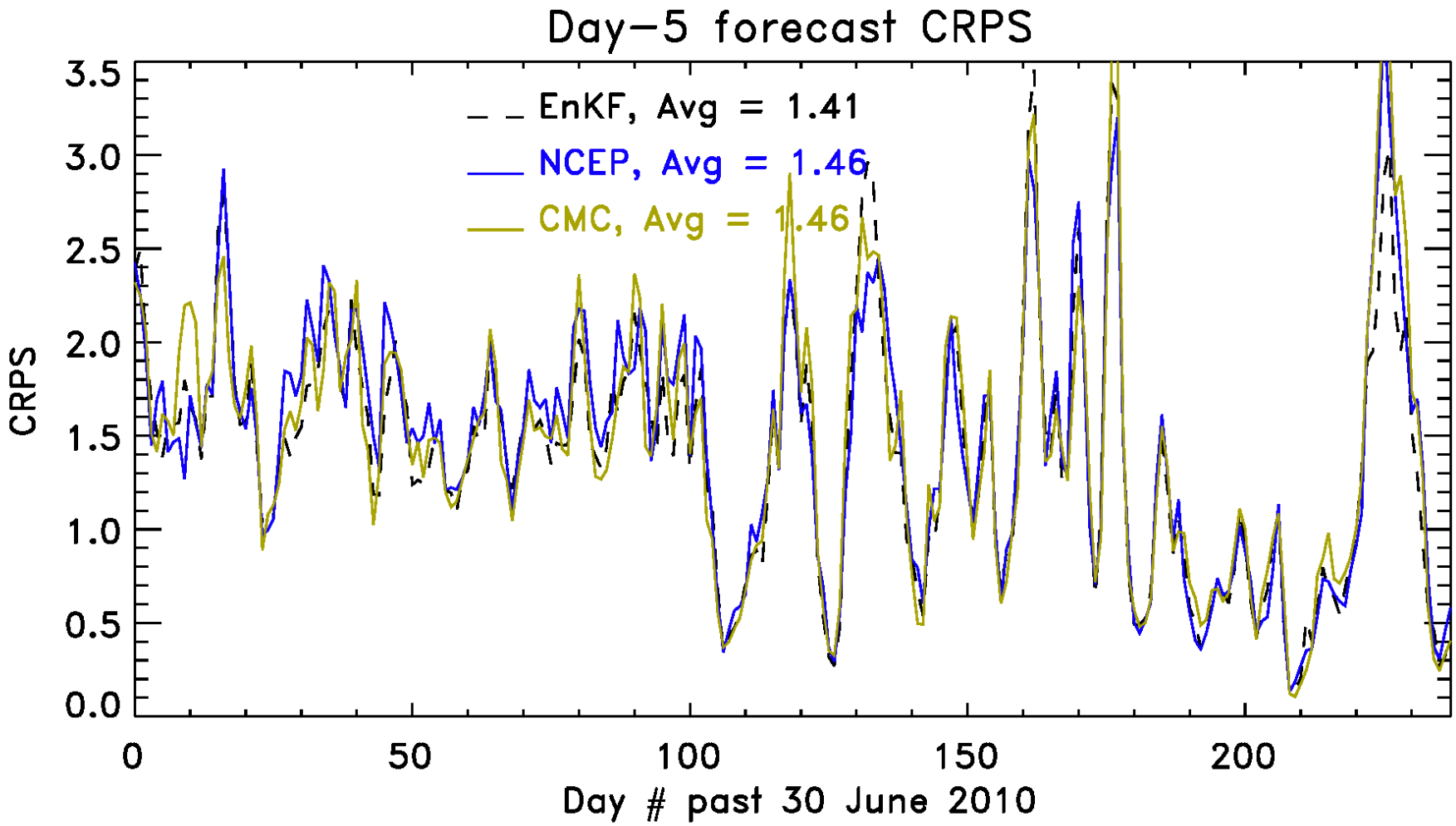
This measures the integrated squared difference between the forecast and observed cumulative distribution functions. **Lower is better**. Main point here is that the time series shows that the T254 EnKF +new GFS is rather consistently lower in error than the NCEP system. Would see the same in a time series of BSS.

Continuous ranked probability scores



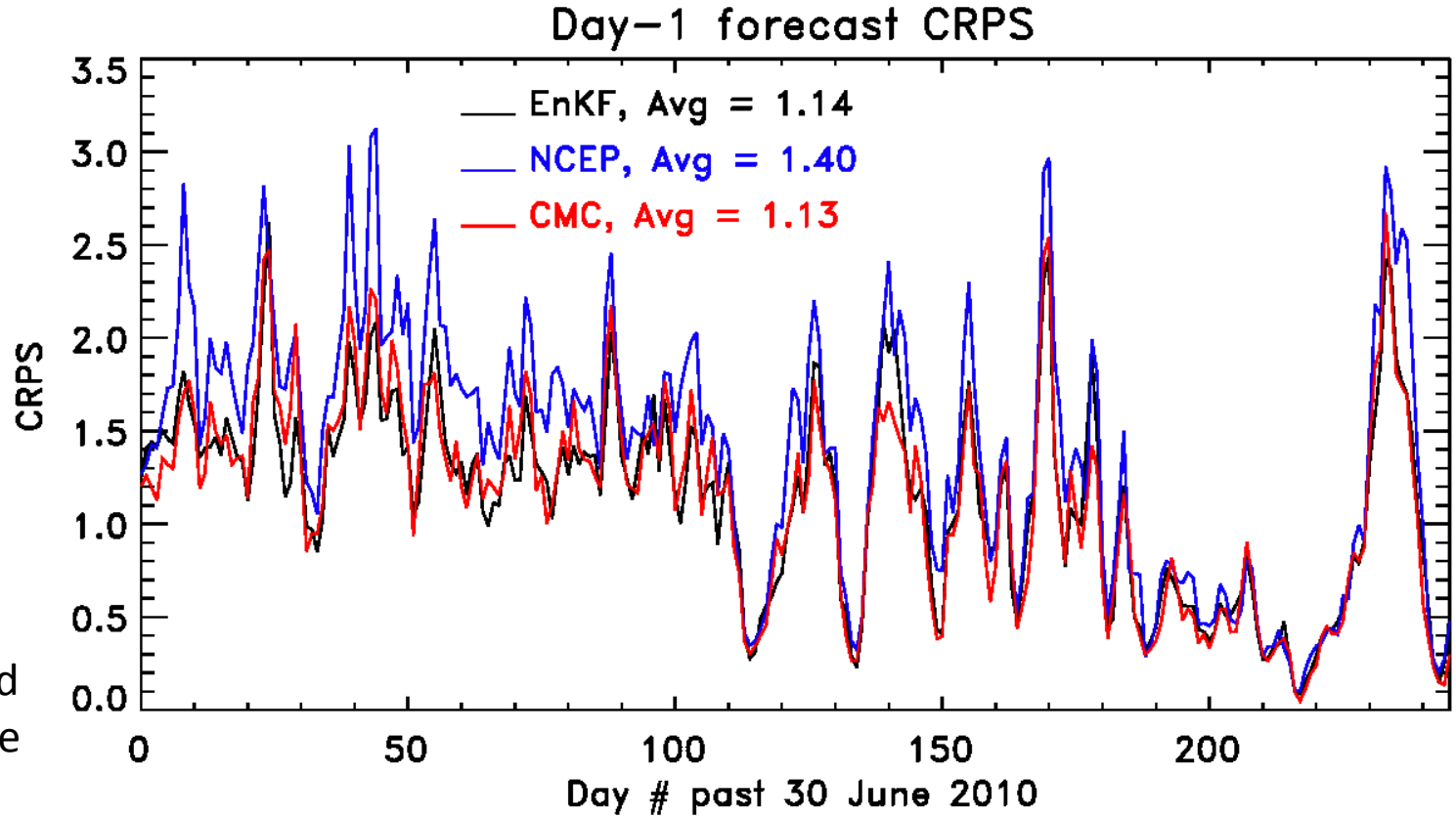
T254 EnKF + new GFS preserves relative advantage over NCEP at day +3 lead.

Continuous ranked probability scores



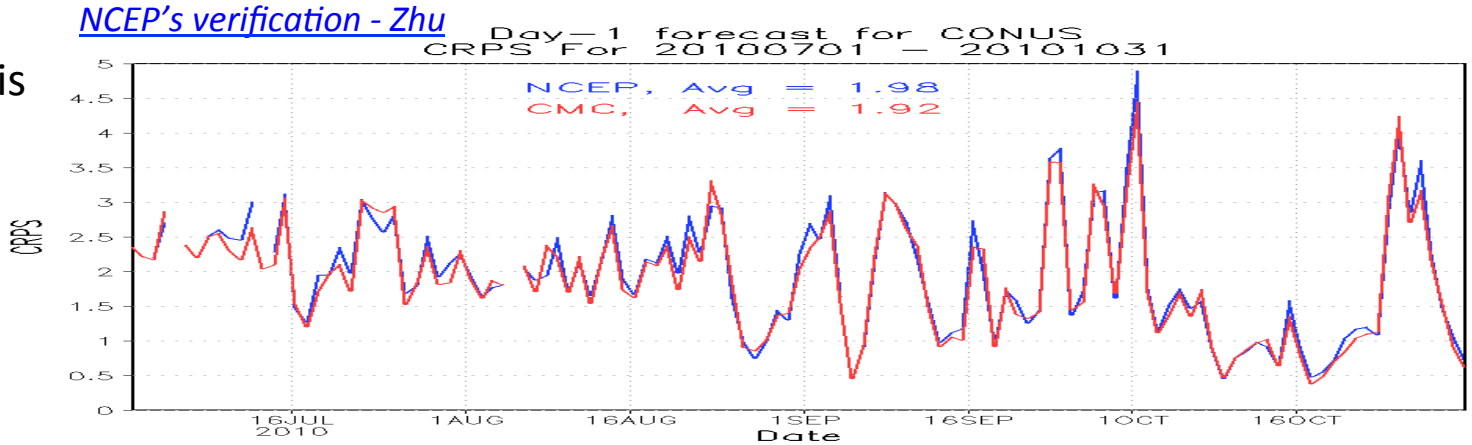
Model forecasts are all more similar as predictability of daily precipitation is marginal.

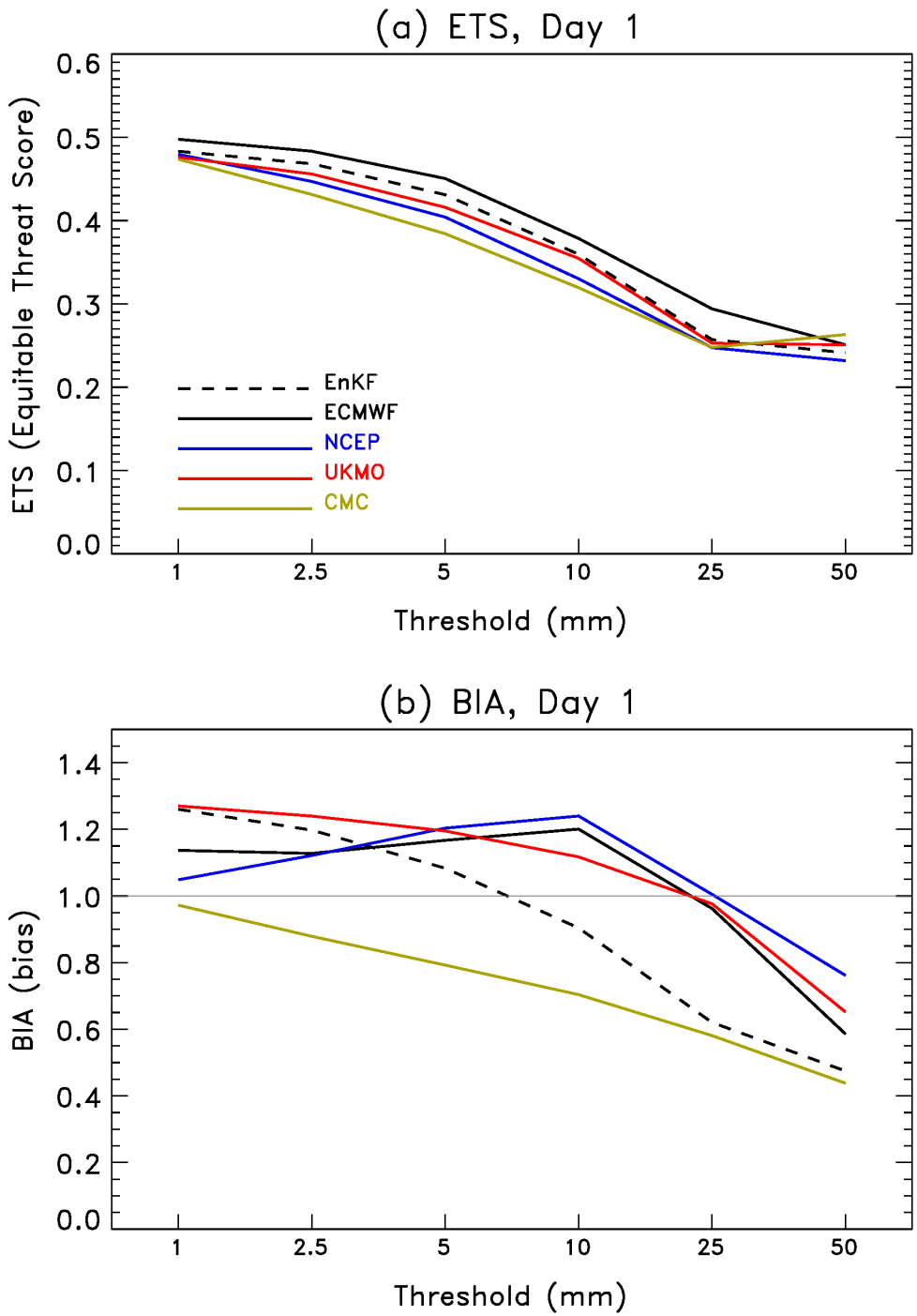
Comparison w. Yuejian Zhu's time series



Differences:
(1) NCEP only 12Z forecasts
(2) NCEP verified against rain gage observations, mine here with Stage–IV analysis
(3) Code used -- mine at end of presentation

NCEP's verification - Zhu

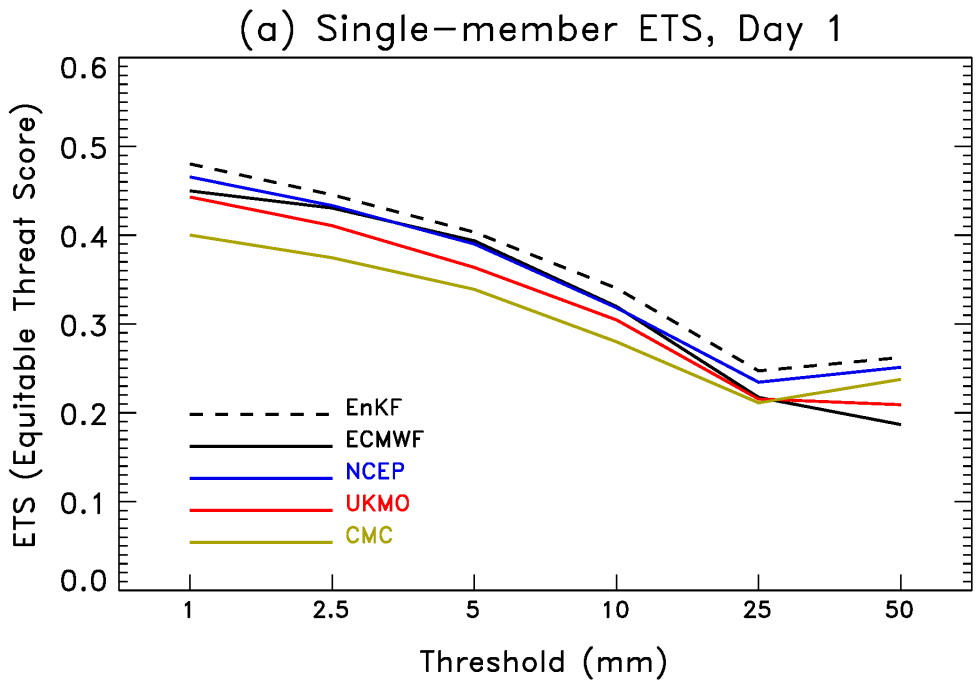


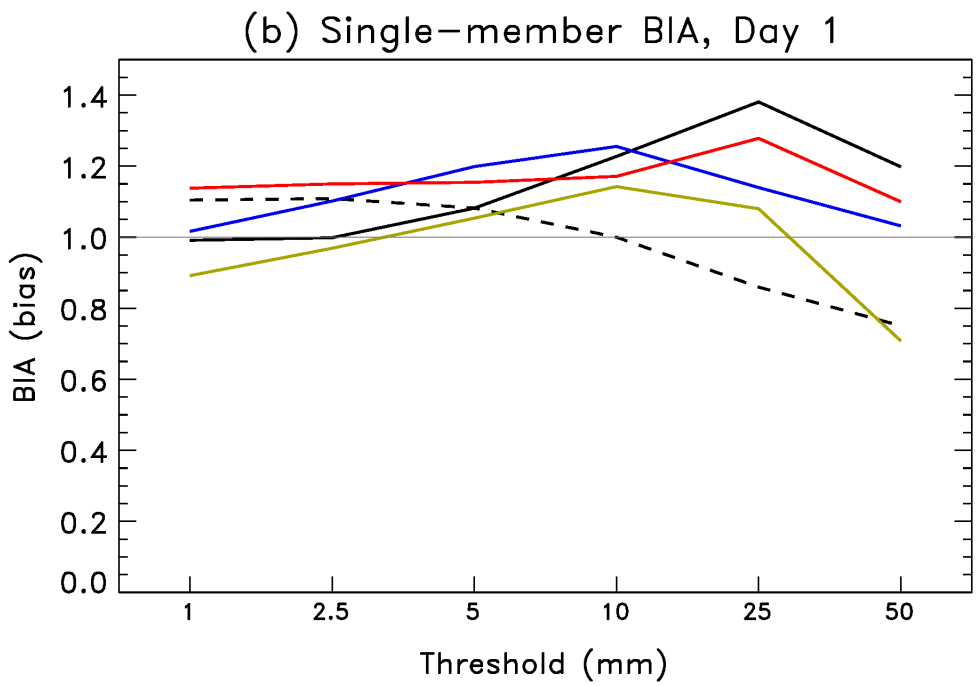


Equitable Threat Score and Bias

ETS of the 20-member ensemble-mean forecast. Again, improvement with EnKF + new GFS.

CMC forecasts less skillful because their forecasts are less sharp (see later slides).

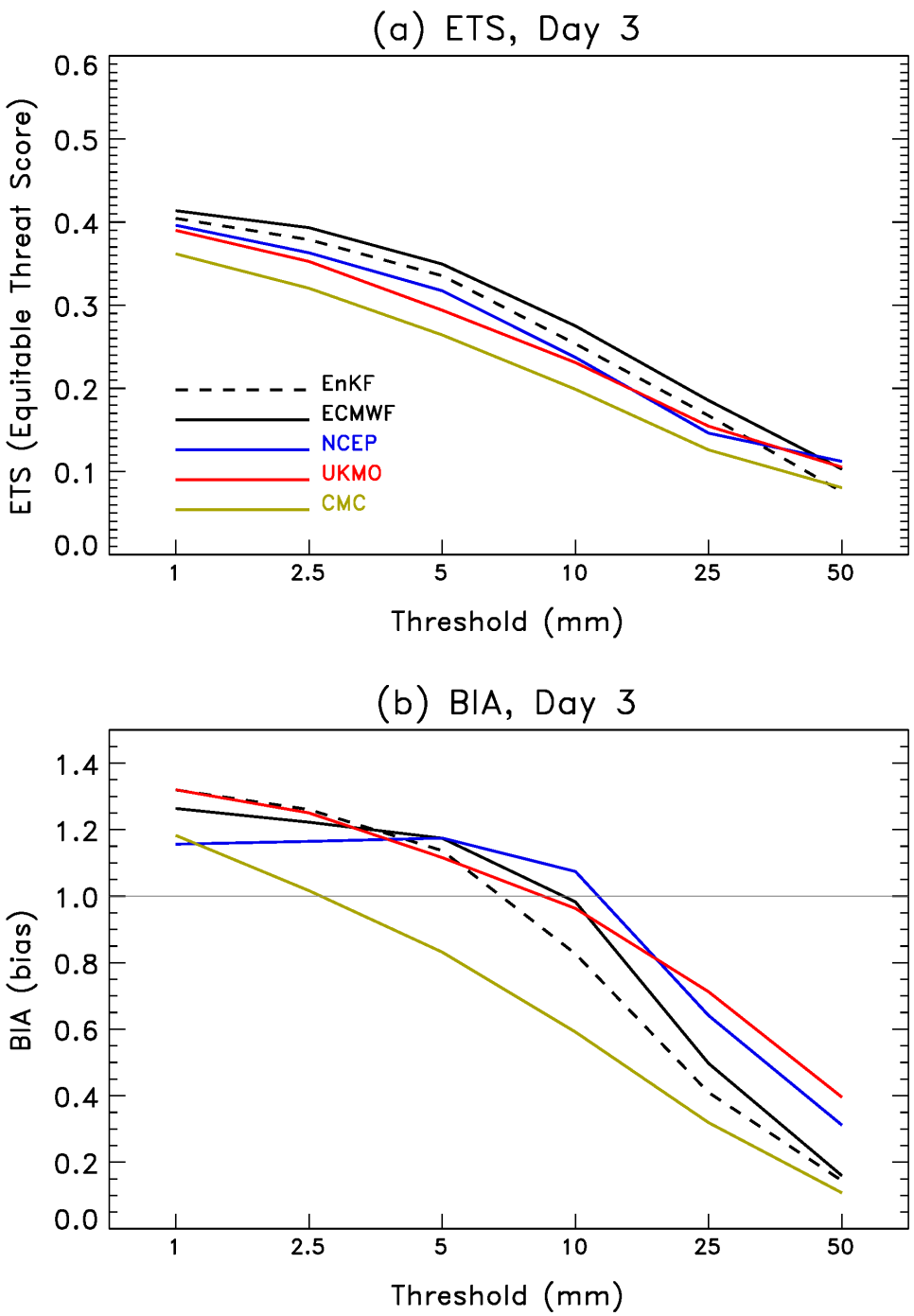




ETS of single member forecast

ETS and BIA will be affected by the ensemble averaging process, e.g., smearing out and reducing the amplitude of the precipitation maxima. Here's stats for one member only.

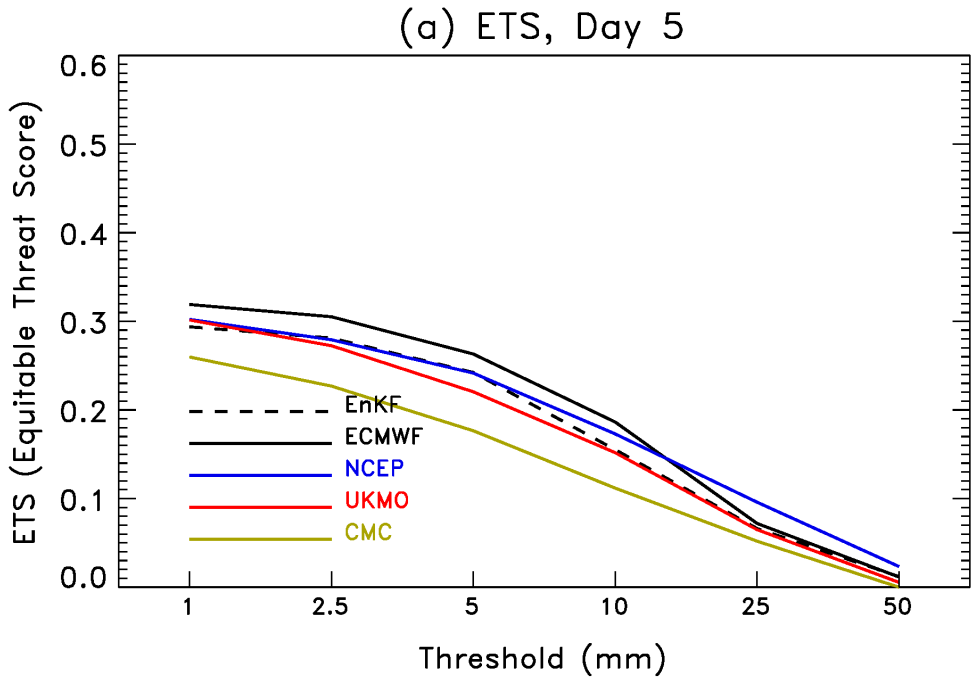
EnKF with new GFS at the top of the pack (but worse than ECMWF in probabilistic metrics)?! Our ensemble system doesn't simulate model error in any fashion, e.g., no stochastic backscatter.

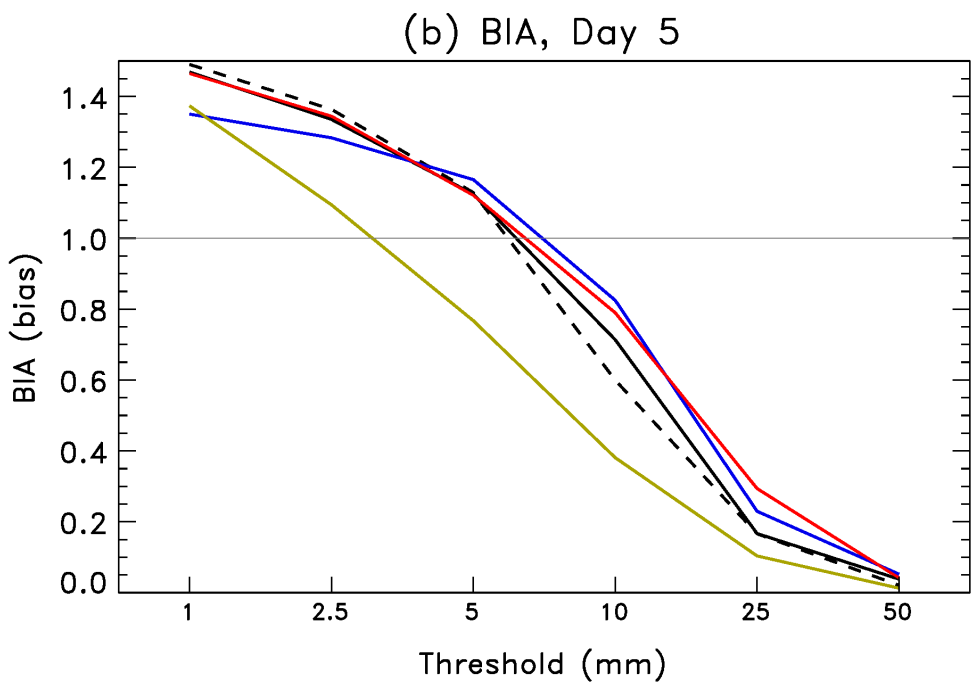


Equitable Threat Score and Bias

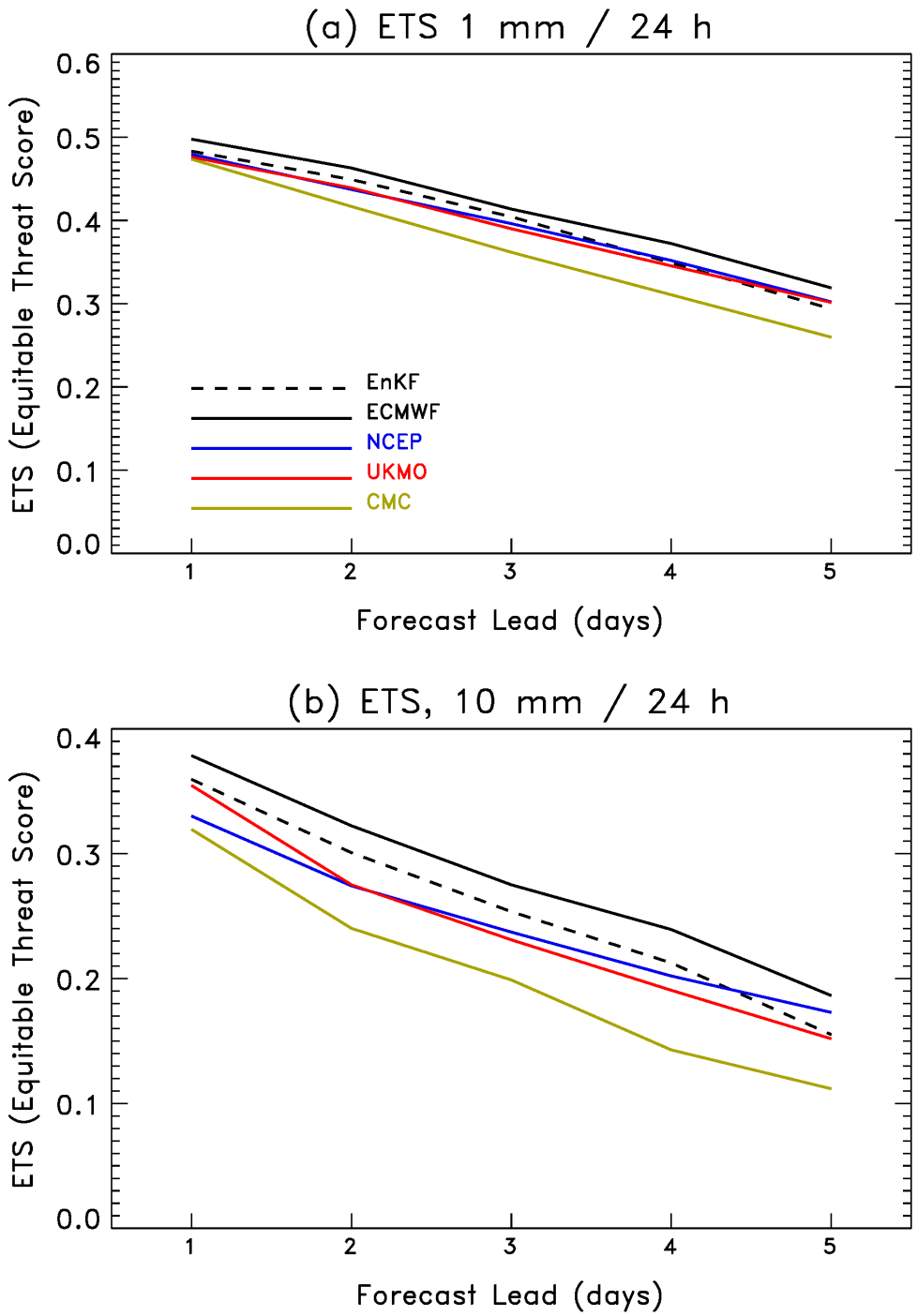
A similar story for the day +3 forecasts.

Equitable Threat Score and Bias

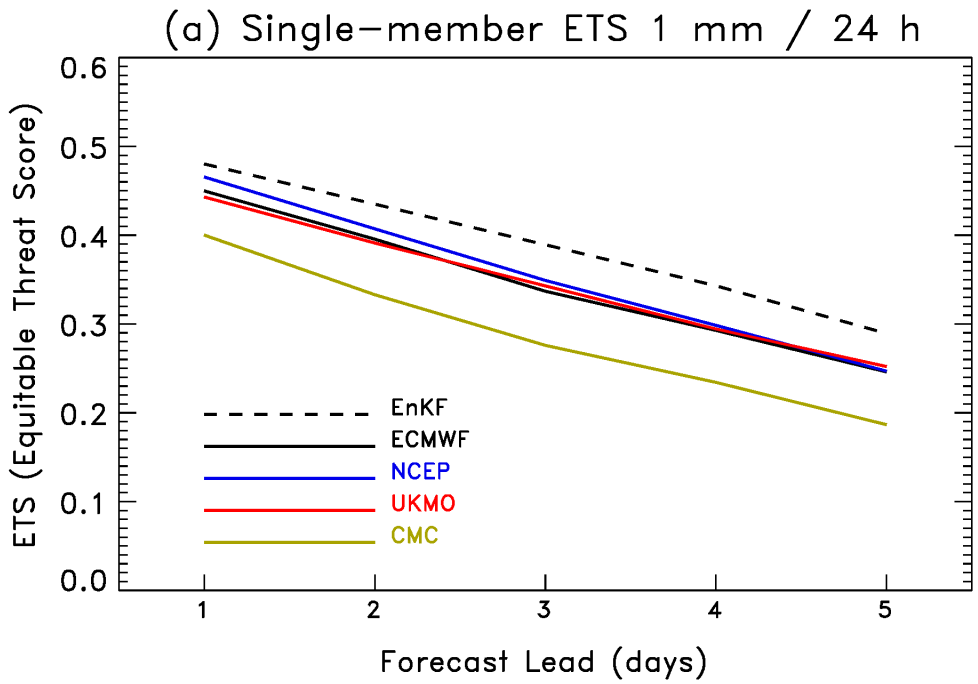


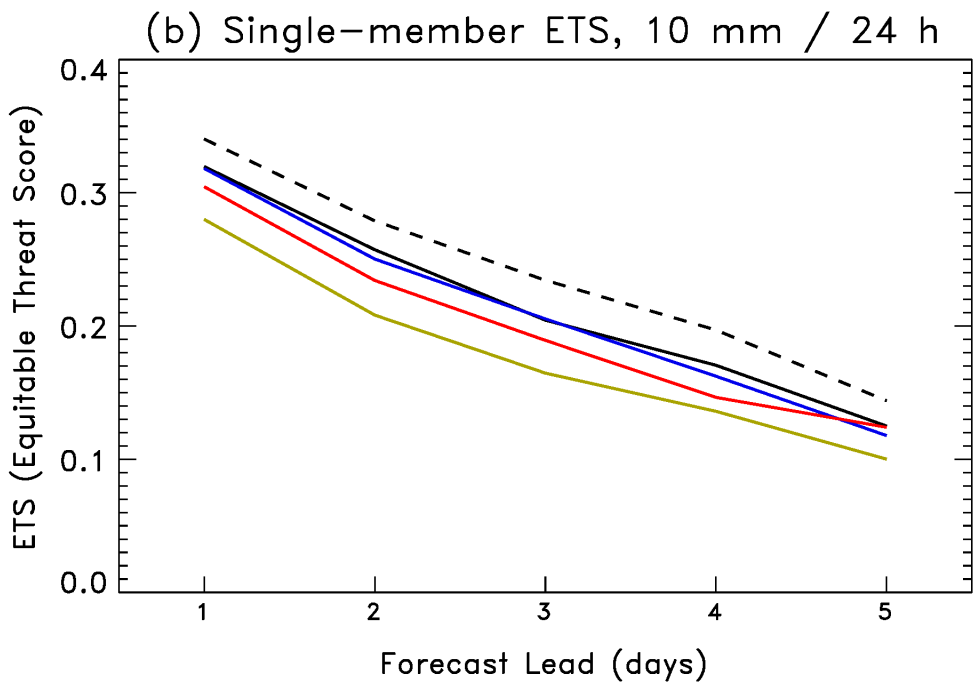


Now the advantage of the EnKF + new GFS has largely disappeared relative to NCEP in this norm.



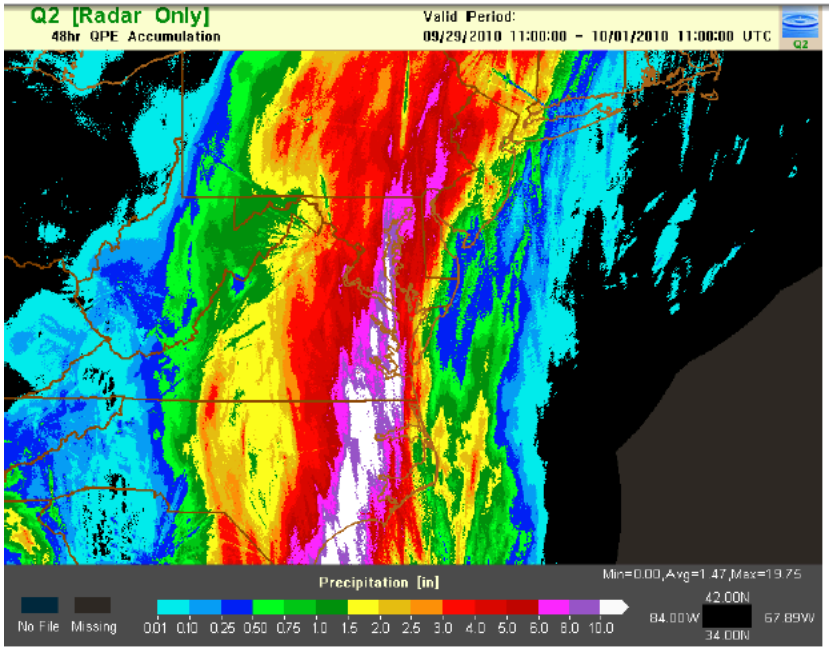
ETS as a function of lead time





ETS as a function of lead time, single member of ensemble.

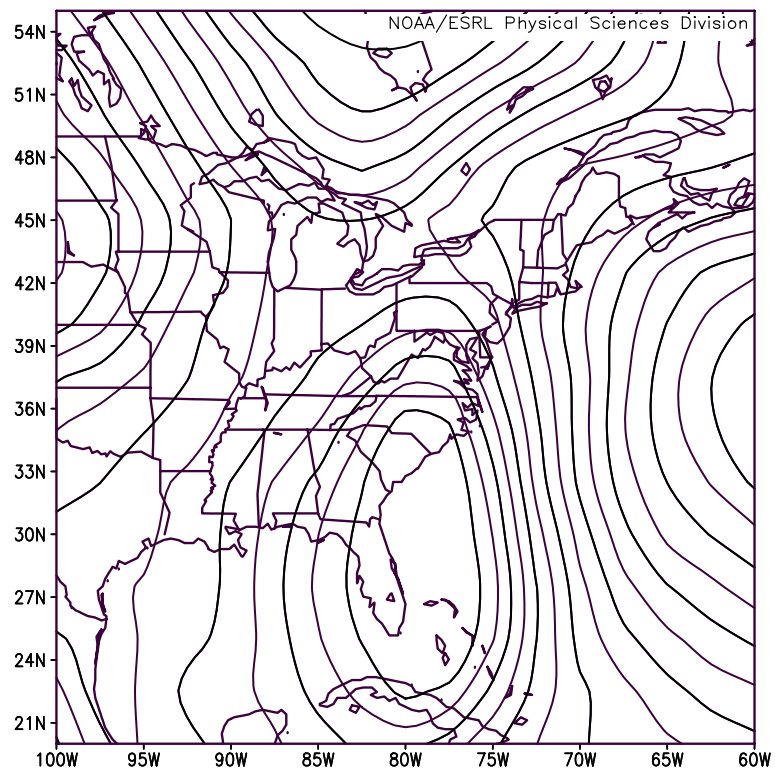
Case study: Extreme rainfall from remnants of TS Nicole, 29-30 September 2010



24-h accumulated precipitation valid 1100 UTC 1 October 2011

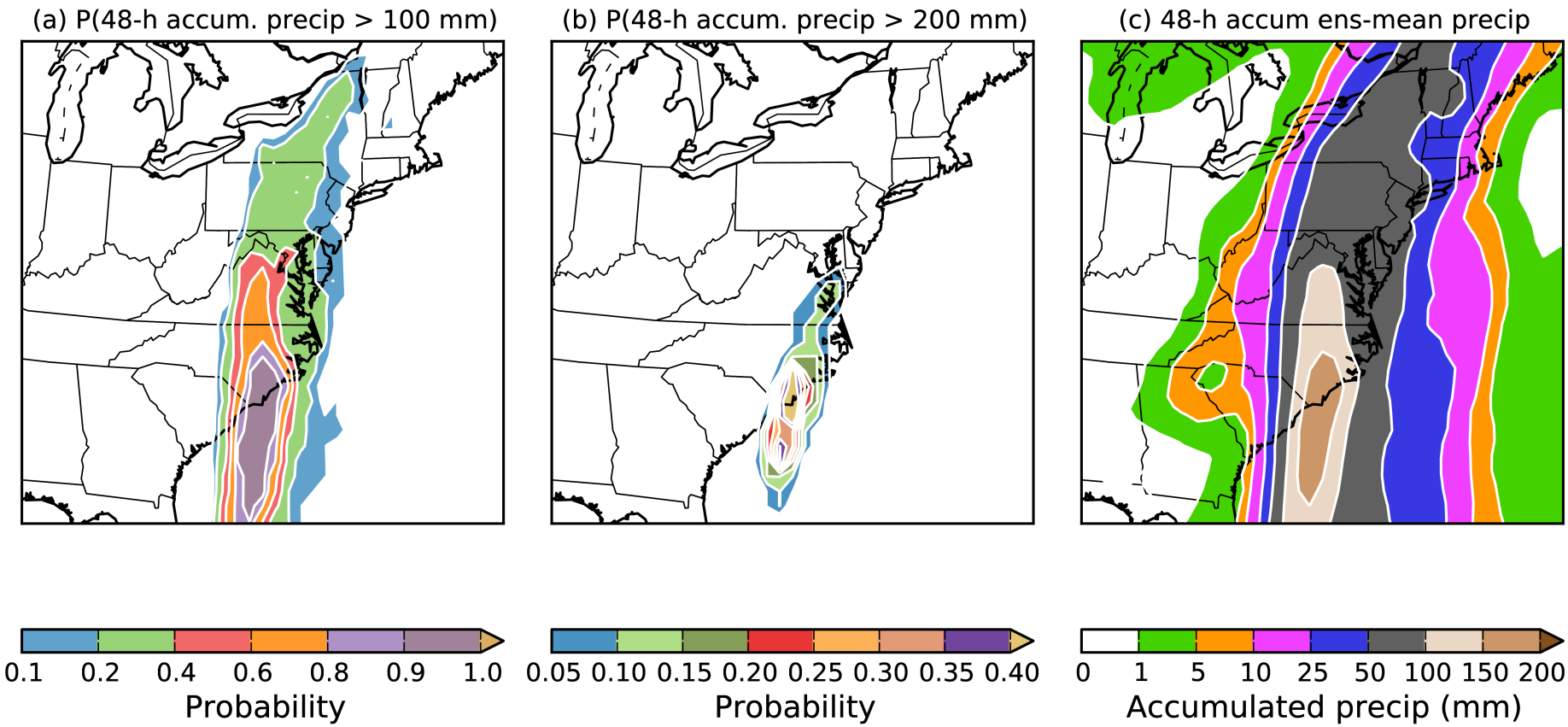
Massive and widespread East-coast rain event, linked in part to moisture advected ahead of remnants of tropical storm Nicole.

precipitation analysis, c/o Rich Grumm, NWS/WFO, State College PA



Probability and ensemble mean, GFS/EnKF

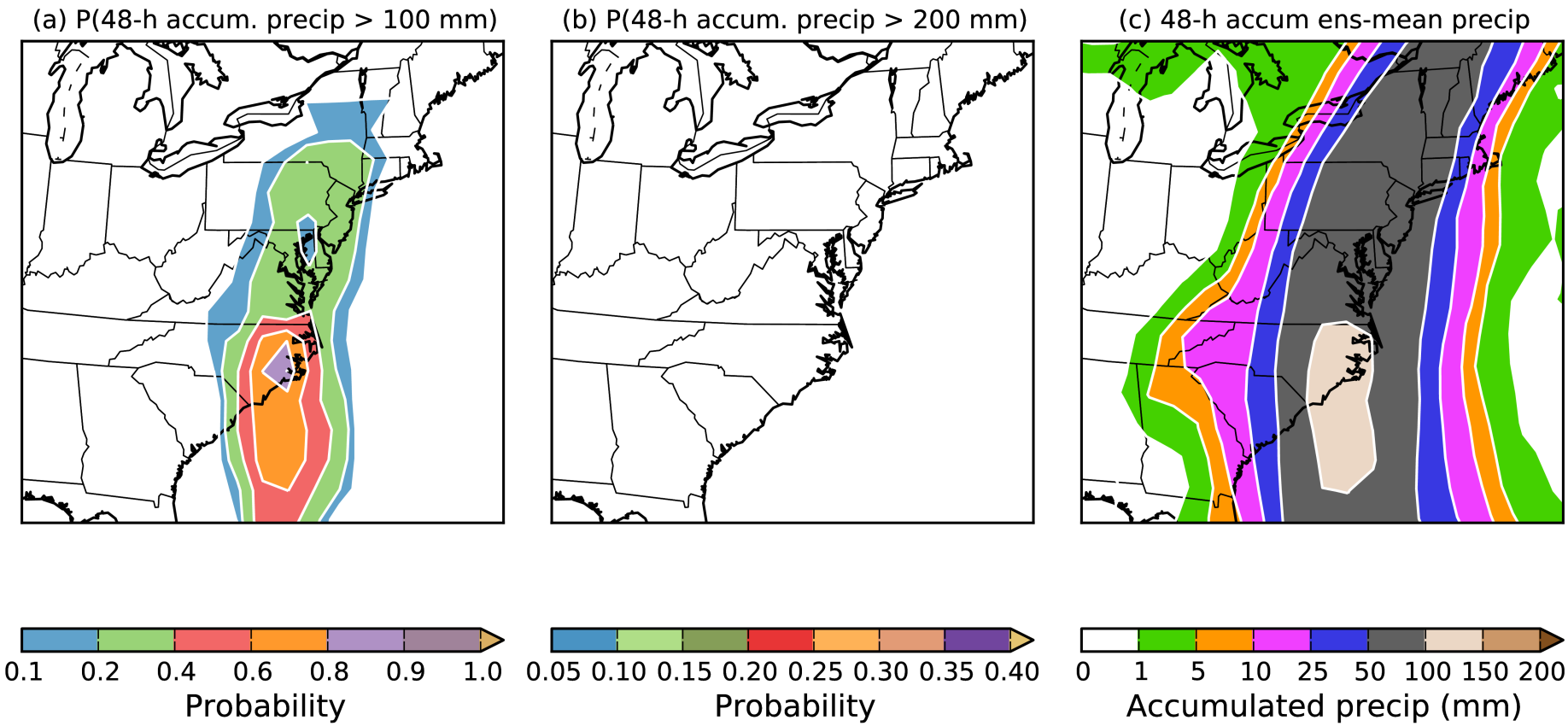
T254 GFS/EnKF 12-60 hour forecast from 2010092900



Axis of high probability for 100 mm too far west, but axis for 200 mm better. Ensemble-mean amounts in excess of 150 mm.

Probability and ensemble mean, NCEP opnl

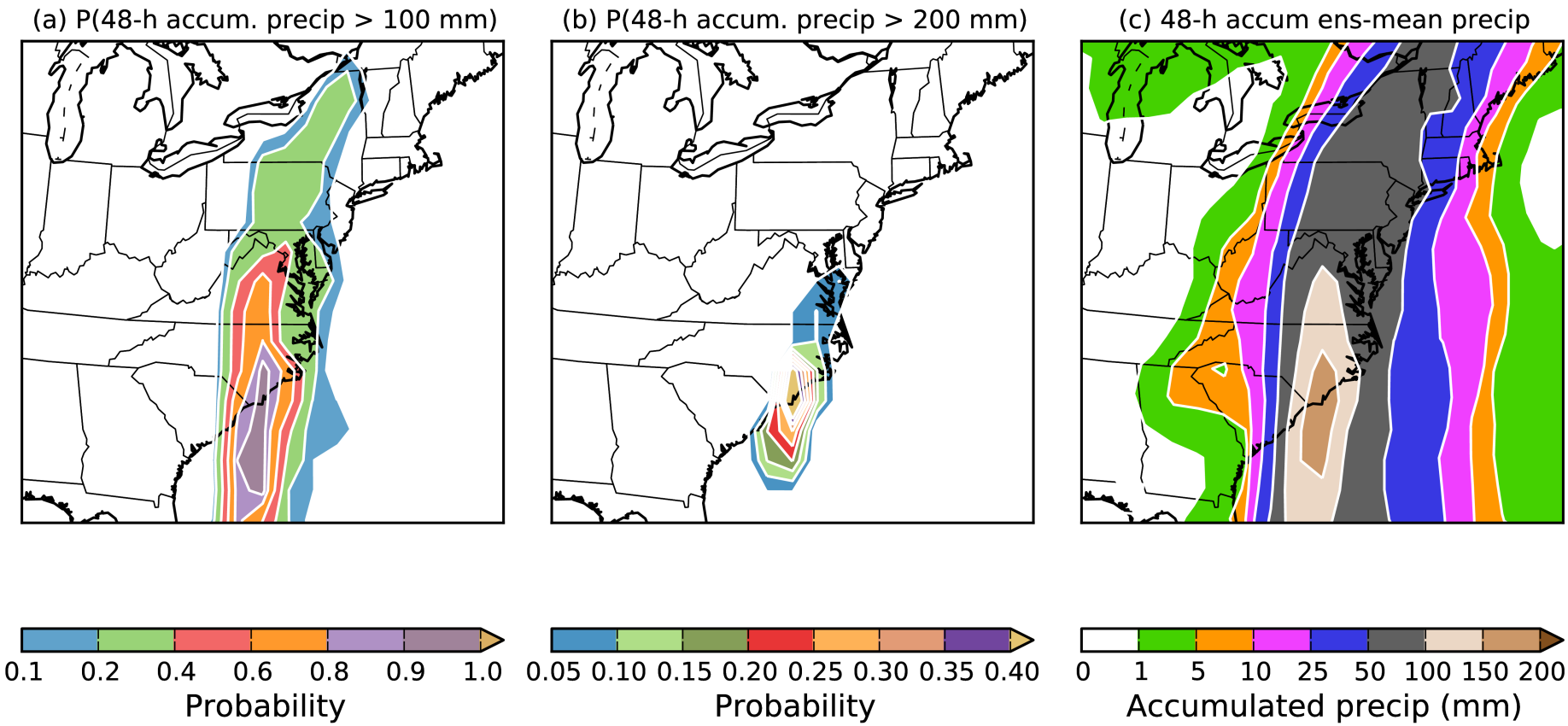
T190 Operational GFS, GSI-ETR 12-60 hour forecast from 2010092900



Better job with axis of high probability for 100 mm relative to GFS/EnKF on previous slide, but no probabilities in excess of 200 mm. Are the lesser amounts relative to GFS/EnKF a function of the coarser grid size (here, 1-degree grid vs. 0.5-degree for GFS-EnKF)?

Probability and ensemble mean, GFS/EnKF, degraded to 1-degree grid of operational

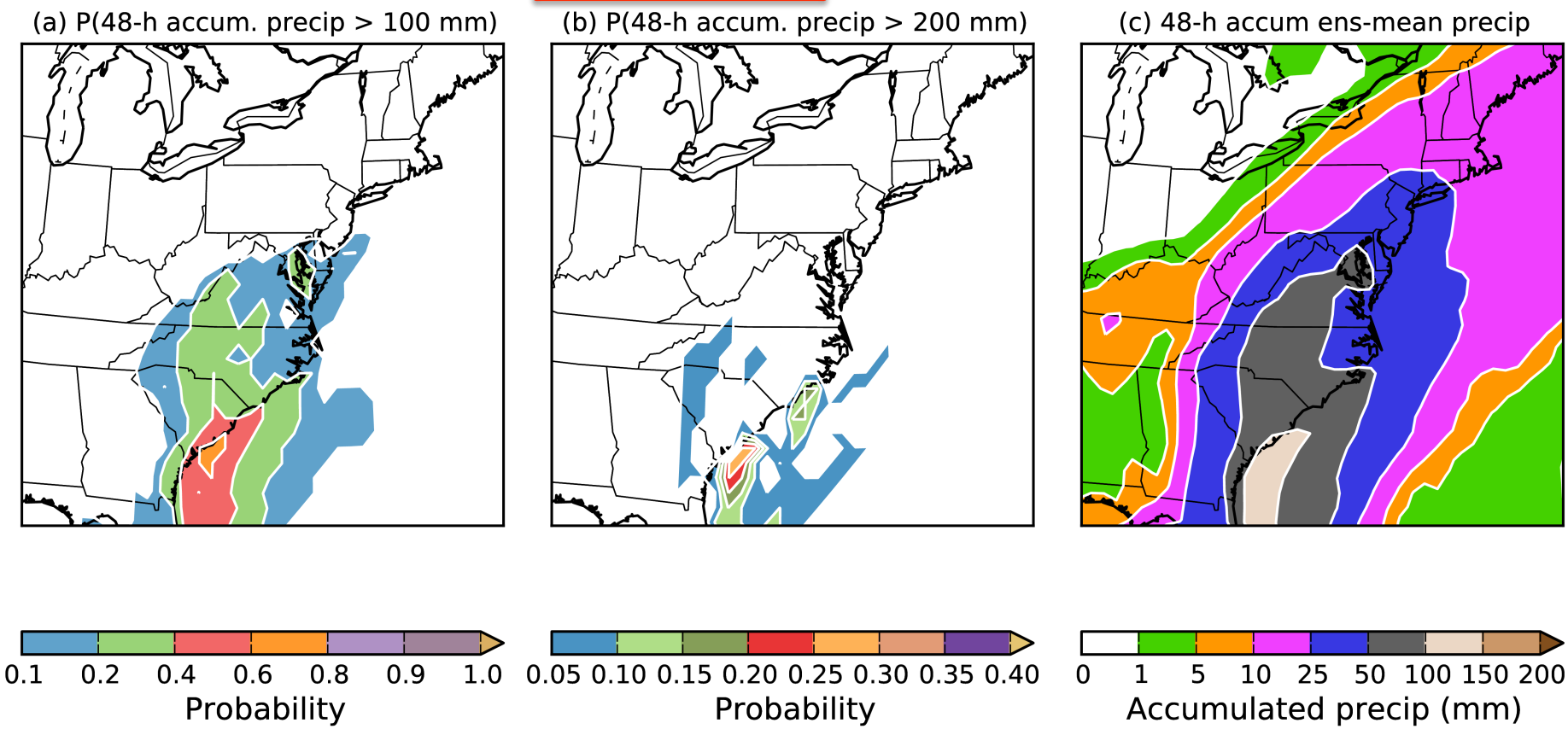
T254 GFS/EnKF 12-60 hour forecast from 2010092900



Axis of high probability for 100 mm too far west, but axis for 200 mm better. Ensemble-mean amounts in excess of 150 mm.

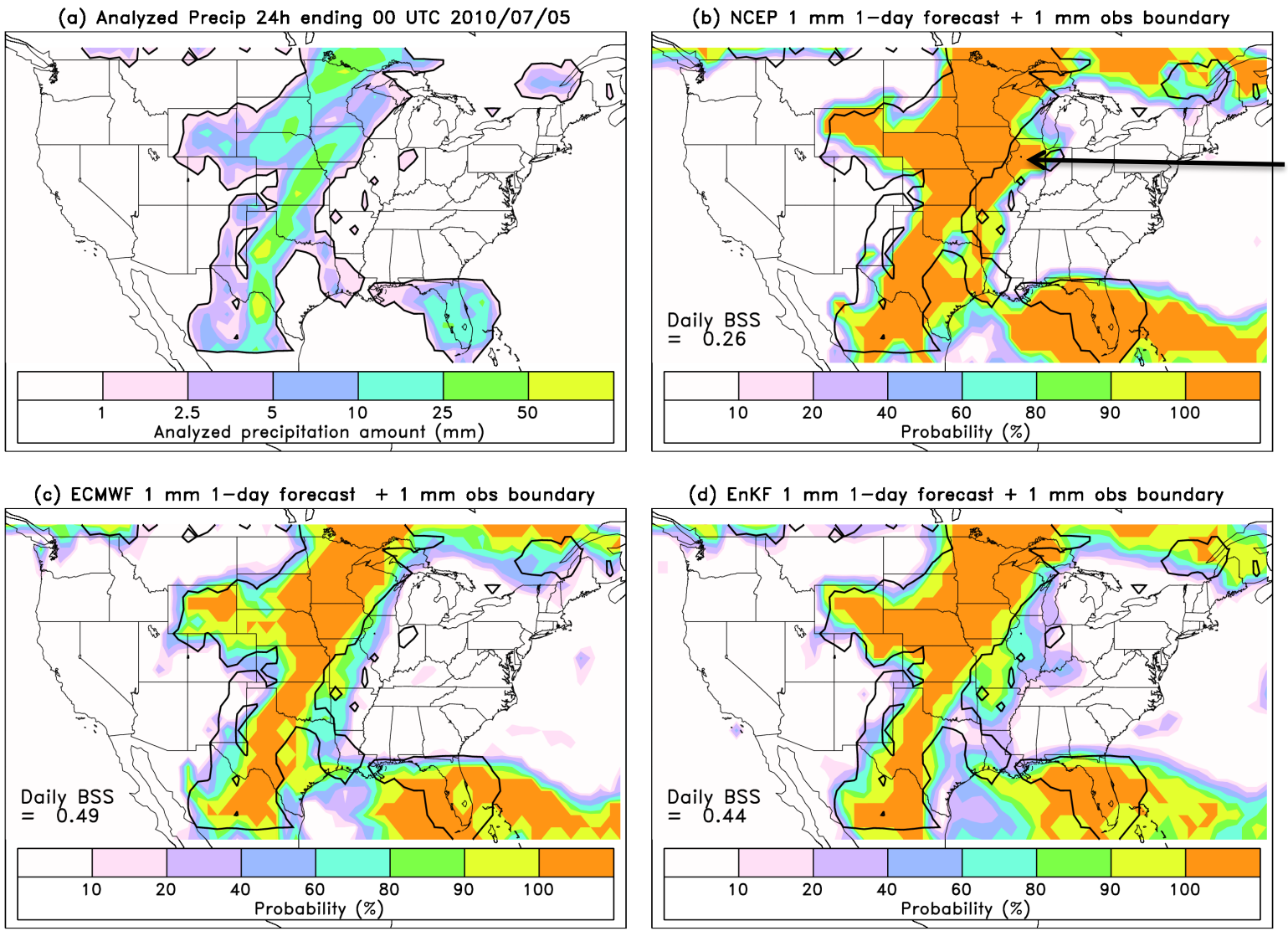
Probability and ensemble mean, GFS/EnKF

T254 GFS/EnKF 72-120 hour forecast from 2010092612



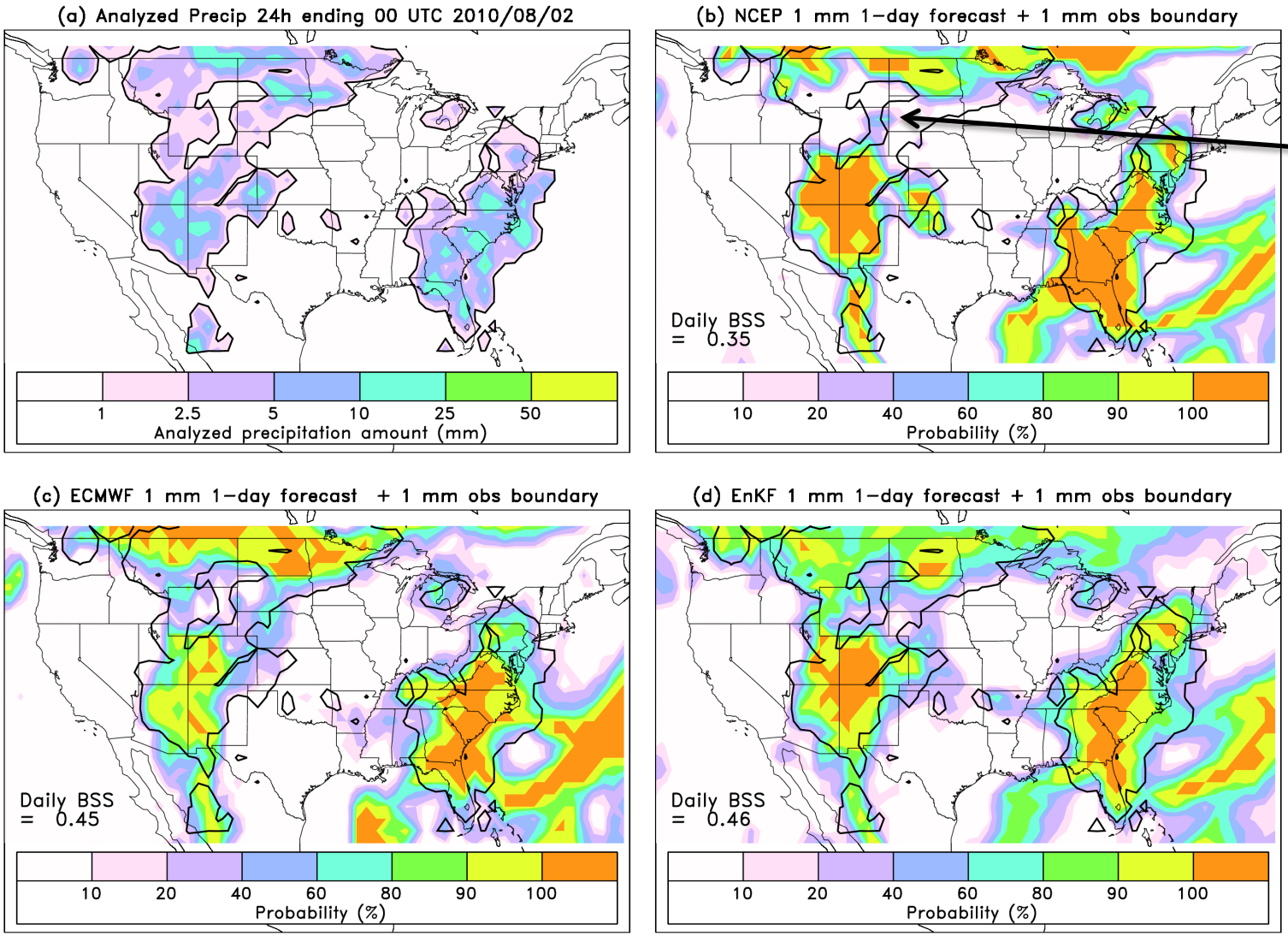
System appears to be a bit slow at this lead, with precip max too far south and west.
Ensemble mean is indicating widespread significant rain along I-95 corridor.

Representative day-1 forecast of 1-mm probabilities from NCEP, ECMWF, EnKF



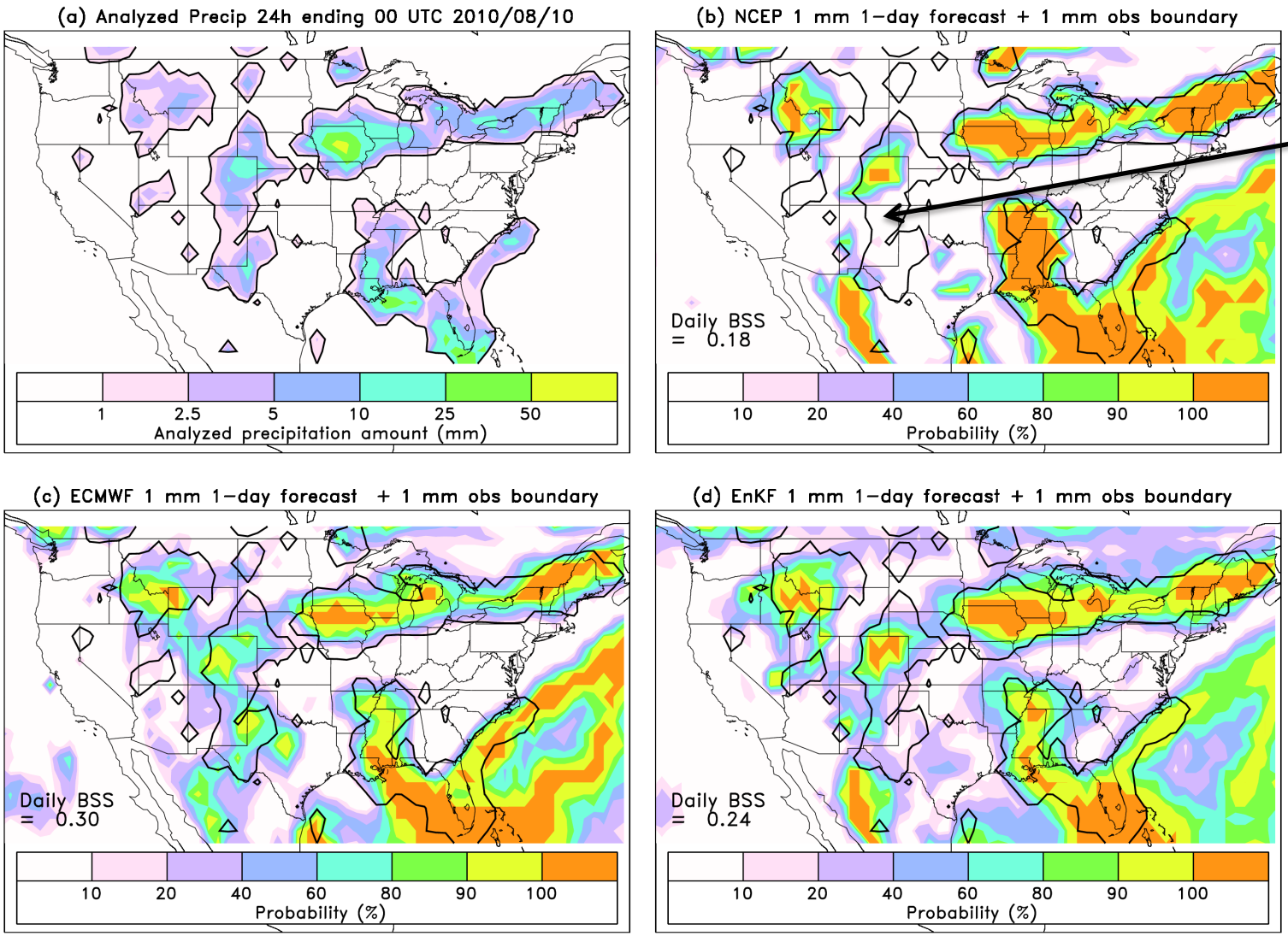
NCEP penalized for overforecast in central MO, IL

Representative day-1 forecast of 1-mm probabilities from NCEP, ECMWF, EnKF



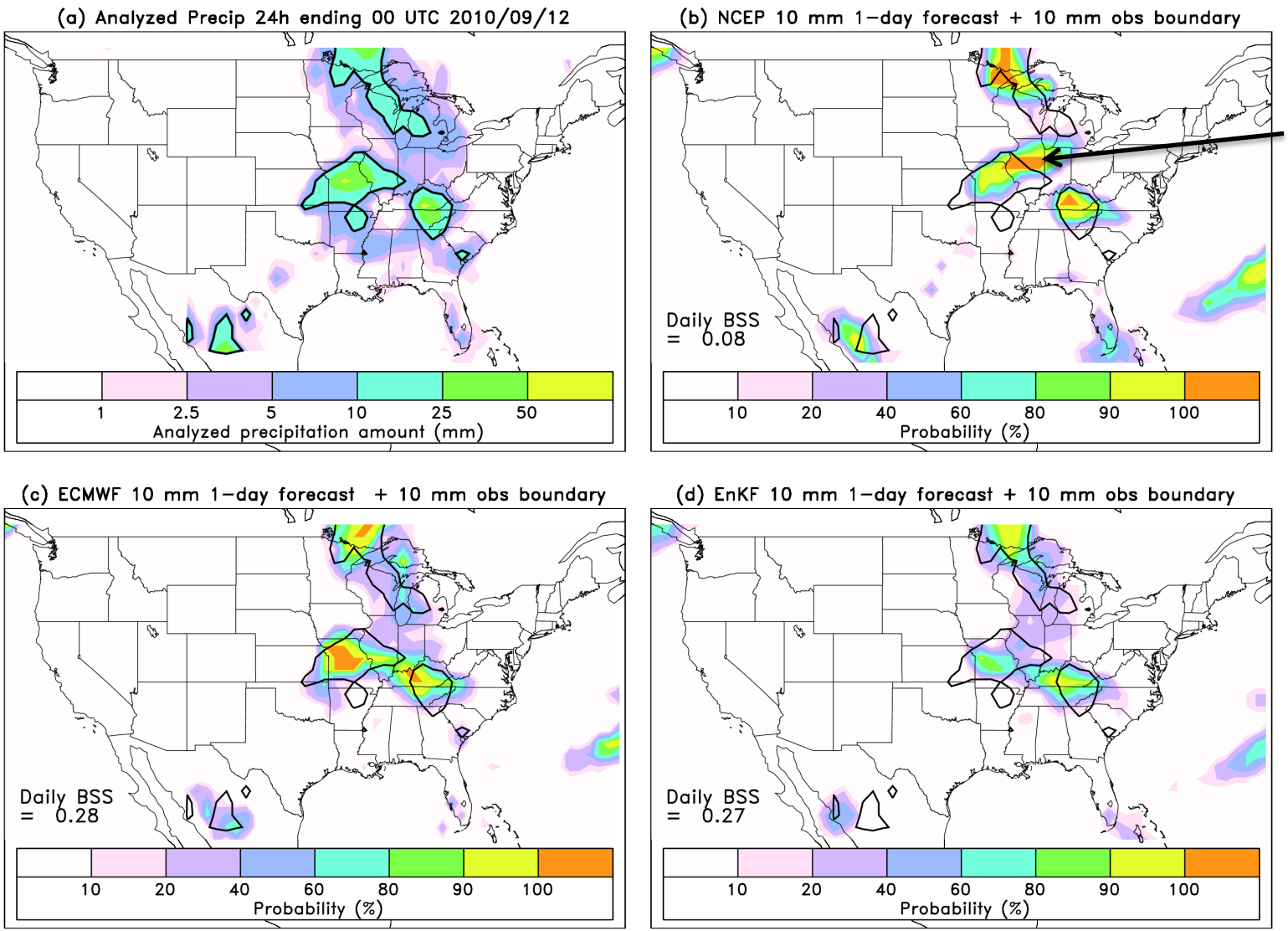
NCEP penalized for under-forecast in upper Great Plains, WY, MT

Representative day-1 forecast of 1-mm probabilities from NCEP, ECMWF, EnKF



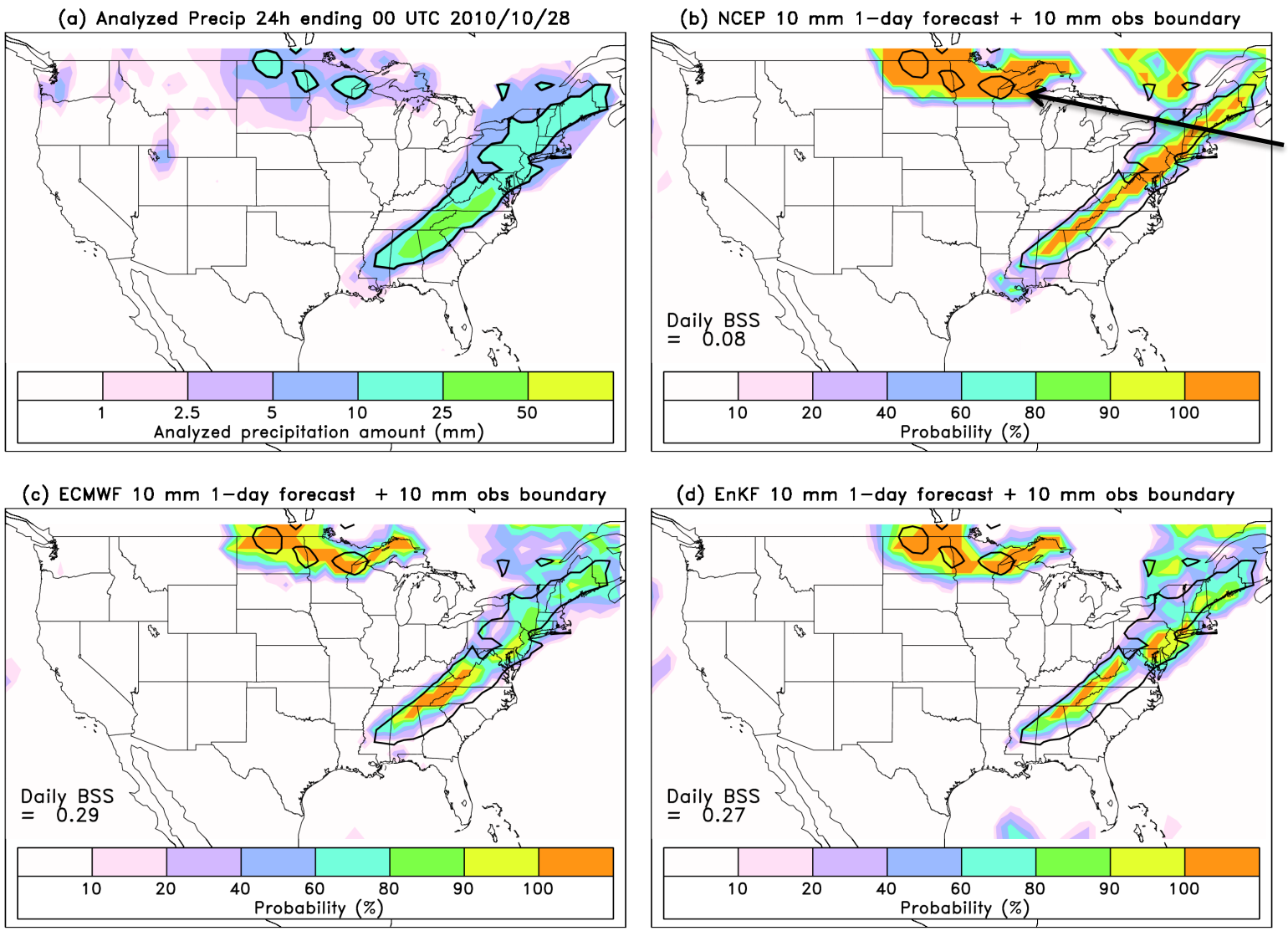
NCEP less filled in with higher probabilities in eastern Rockies

Representative day-1 forecast of 10-mm probabilities from NCEP, ECMWF, EnKF



NCEP more penalized for overforecast in central IL.

Representative day-1 forecast of 10-mm probabilities from NCEP, ECMWF, EnKF



NCEP overforecasts the most in northern Great Plains.

Conclusions

- Improved performance of the EnKF + new GFS is consistent with the improved performance we've seen for, e.g., hurricane track, tropical winds, 500 hPa anomaly correlations.

- More evidence that the new GFS and the EnKF should benefit the hybrid system, in development.
  - suggestion that impact of improved precipitation forecasts here may be more due to new GFS than EnKF, however.

- Thanks to Daryl Kleist, John Derber, Dave Parrish for their work in developing EnKF & hybrid, Yuejian Zhu for discussions, Yan Luo for precip. data set.

Verification details

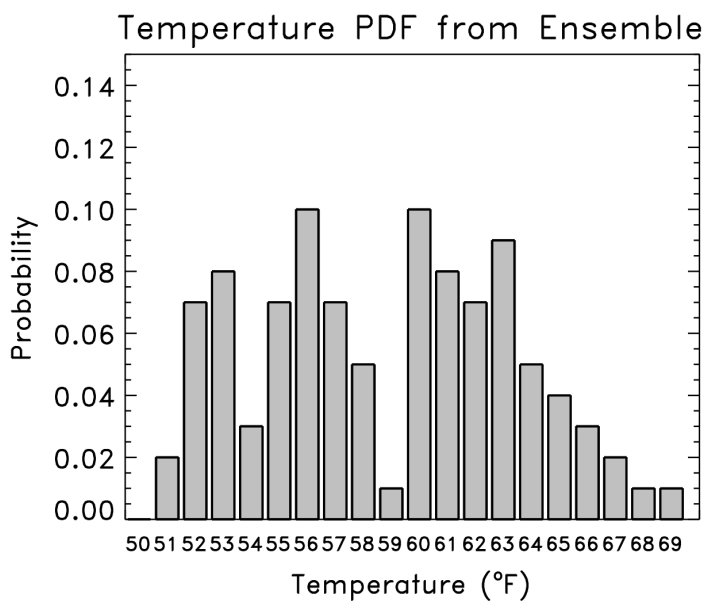
- CRPS
- Brier skill scores
- Reliability diagrams
- ETS & Bias
- Sharpness plots

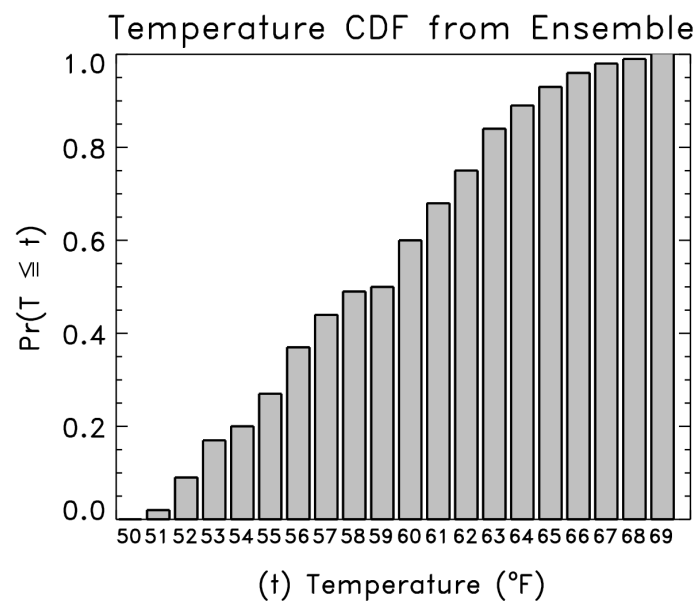
[go back]

Cumulative distribution function (CDF); used in CRPS

- $F^f(x) = \Pr\{X \leq x\}$

  where X is the random variable, x is some specified threshold.

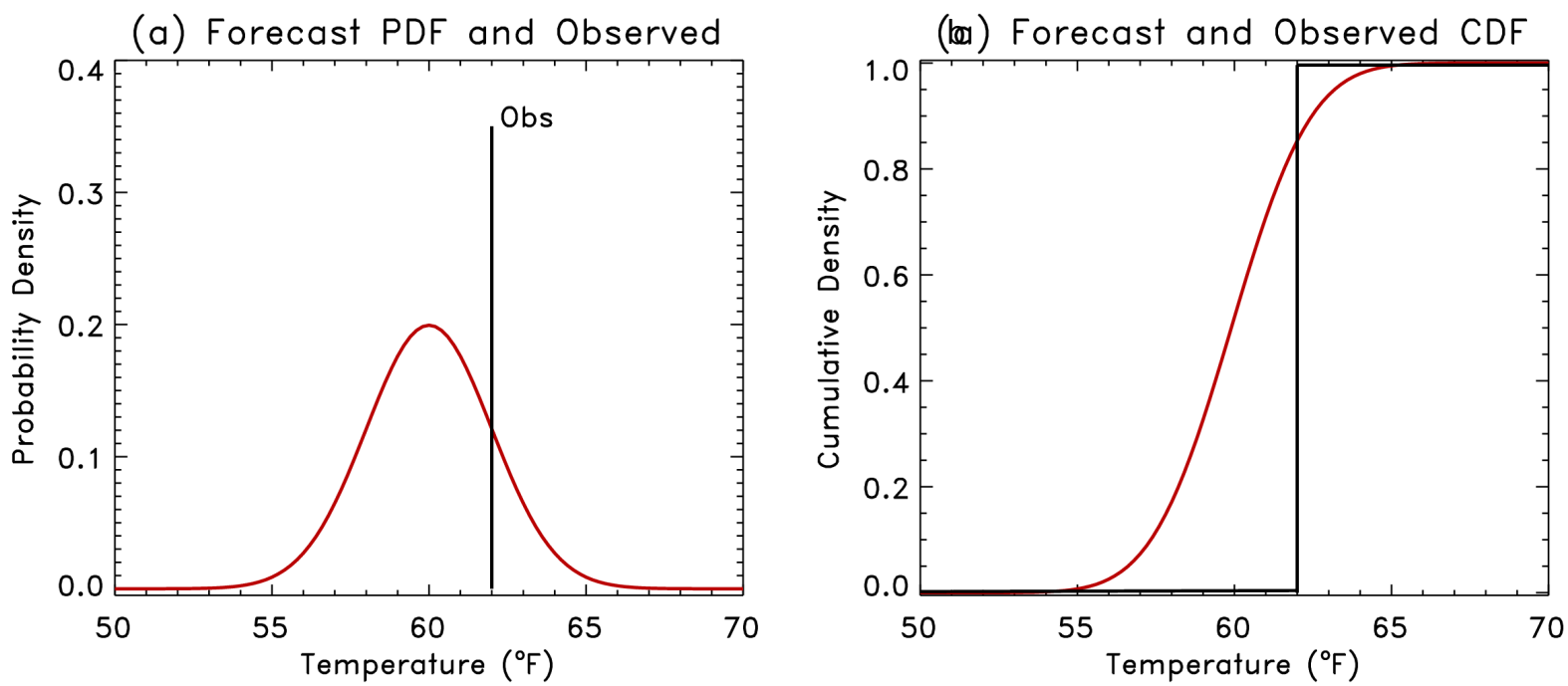




Continuous Ranked Probability Score

- Let $F_i^f(x)$ be the forecast probability CDF for the *i*th forecast case.
- Let $F_i^o(x)$ be the observed probability CDF (Heaviside function).

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{x=-\infty}^{x=-\infty}\left(F_i^f(x) - F_i^o(x)\right)^2 dx$$

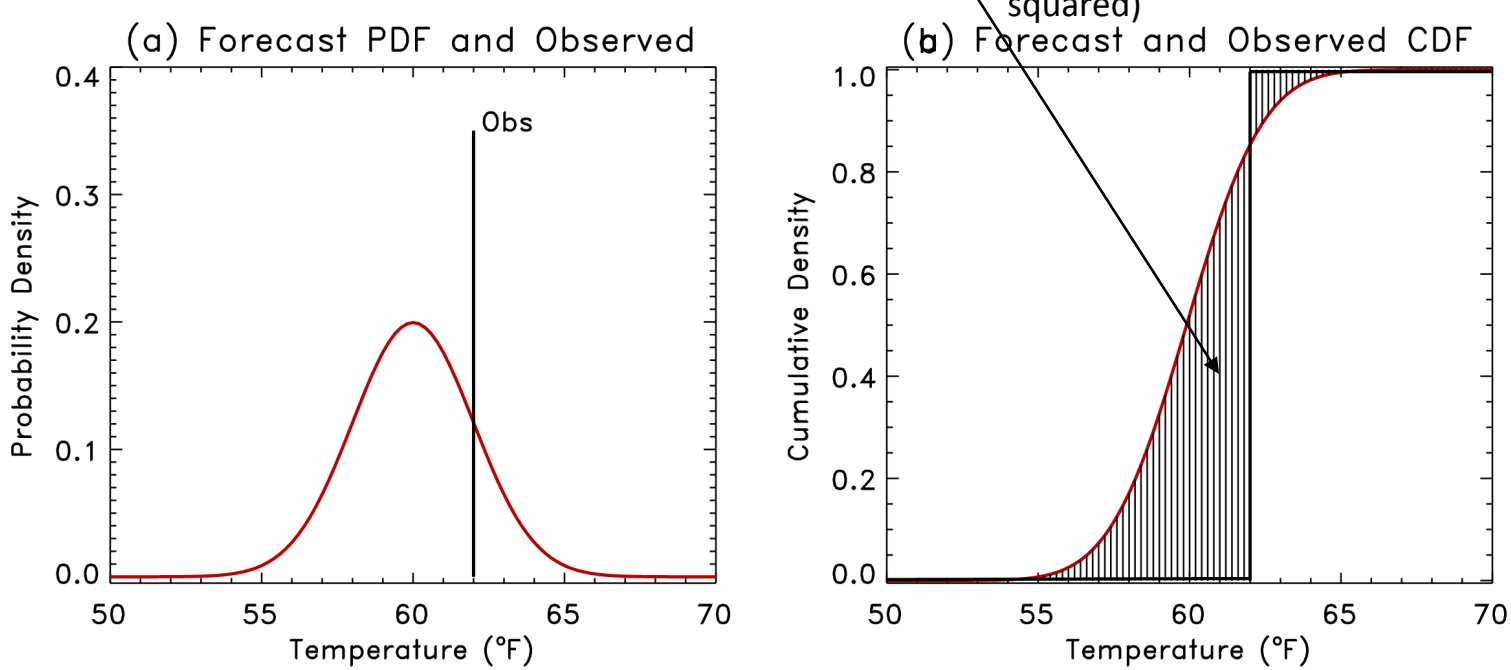


Continuous Ranked Probability Score

- Let $F_i^f(x)$ be the forecast probability CDF for the *i*th forecast case.
- Let $F_i^o(x)$ be the observed probability CDF (Heaviside function)*.

$$CRPS = \frac{1}{n}\sum_{i=1}^{n}\int_{x=-\infty}^{x=-\infty}\left(F_i^f(x) - F_i^o(x)\right)^2 dx$$

(difference squared)



* or incorporate obs error; see Candille and Talagrand, QJRMS,2007

Brier score and skill score

$$BS(f) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - o_i\right)^2$$

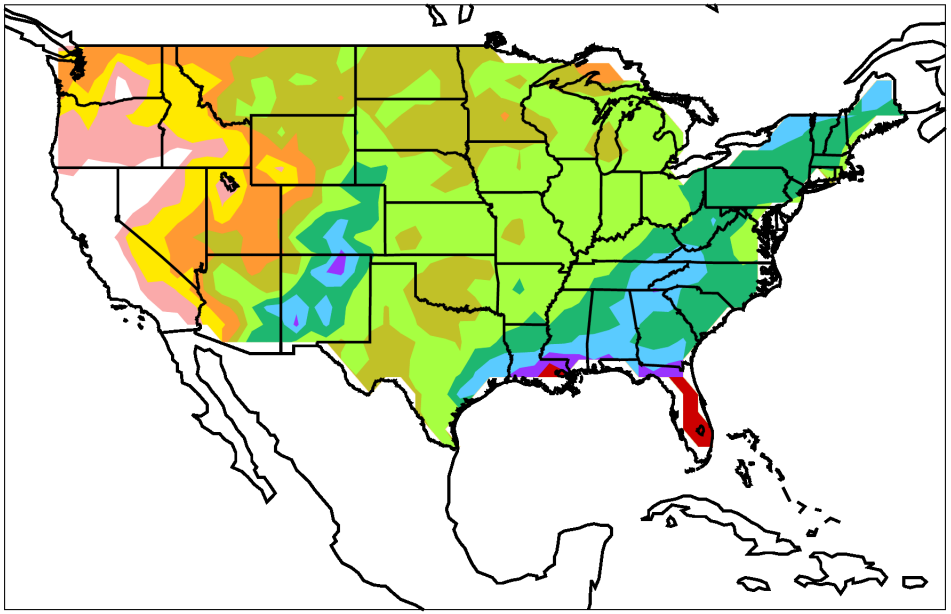
$$BSS = 1.0 - \frac{BS(f)}{BS(cl)}$$

My past research has discussed how this can over-estimate forecast skill (Hamill and Juras, October 2006 QJRMS). For purposes today, we'll ignore this effect, as we're more interested in relative scores of various models than in the absolute scores.

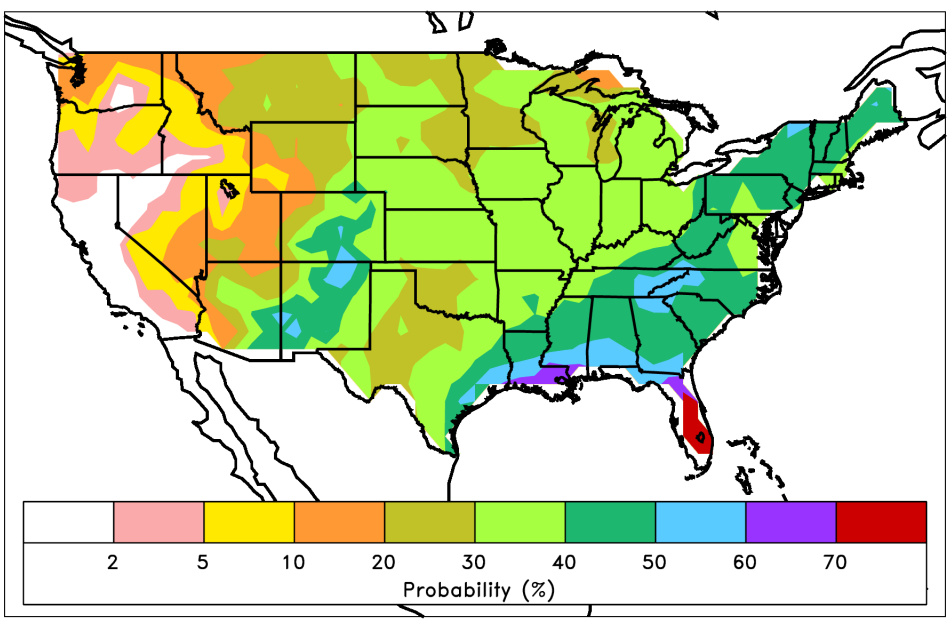
Calculation of climatological probabilities for BSS reference

- Based on climatology of Stage-4 data, 2005-2009. Trouble generating off EMC CCPA product.
- Climatological probabilities determined separately for each month.
- In very dry regions, probabilities directly estimated from event relative frequency will be distorted by "small" sample size.
- Remedy this by fitting Gamma distribution to nonzero precipitation amounts using maximum likelihood estimation technique to power transformed (^0.75) precipitation.
- Resulting fields look reasonable.

(a) 1 mm climatological probability, relative frequency Jul



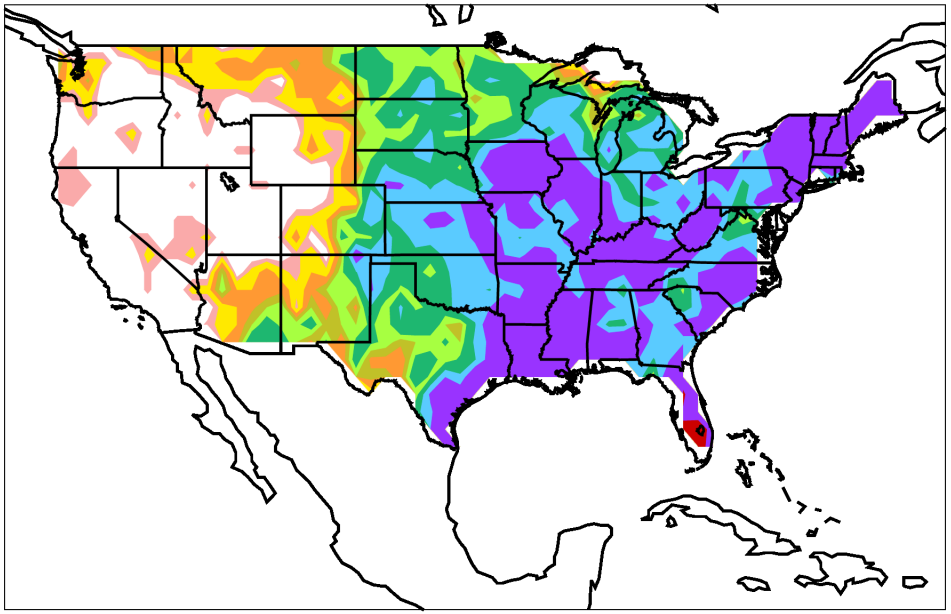
(b) 1 mm climatological probability, Gamma–distribution fitted, Jul



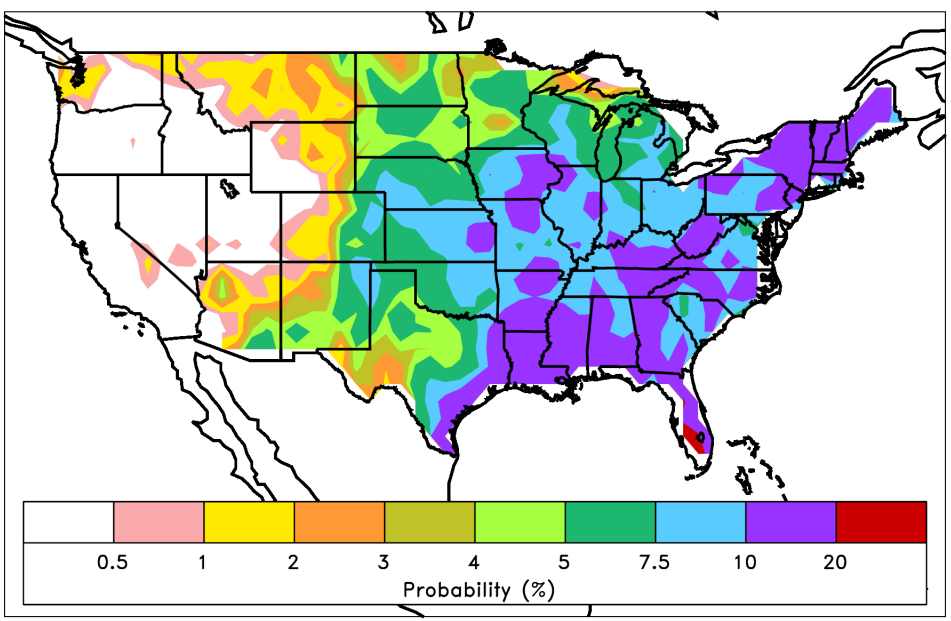
Climatological probability of > 1 mm / 24h, estimated from event relative frequency (top) and modeled (bottom)

you can find regions of slight underestimate (e.g., northern NM).

(a) 10 mm climatological probability, relative frequency Jul

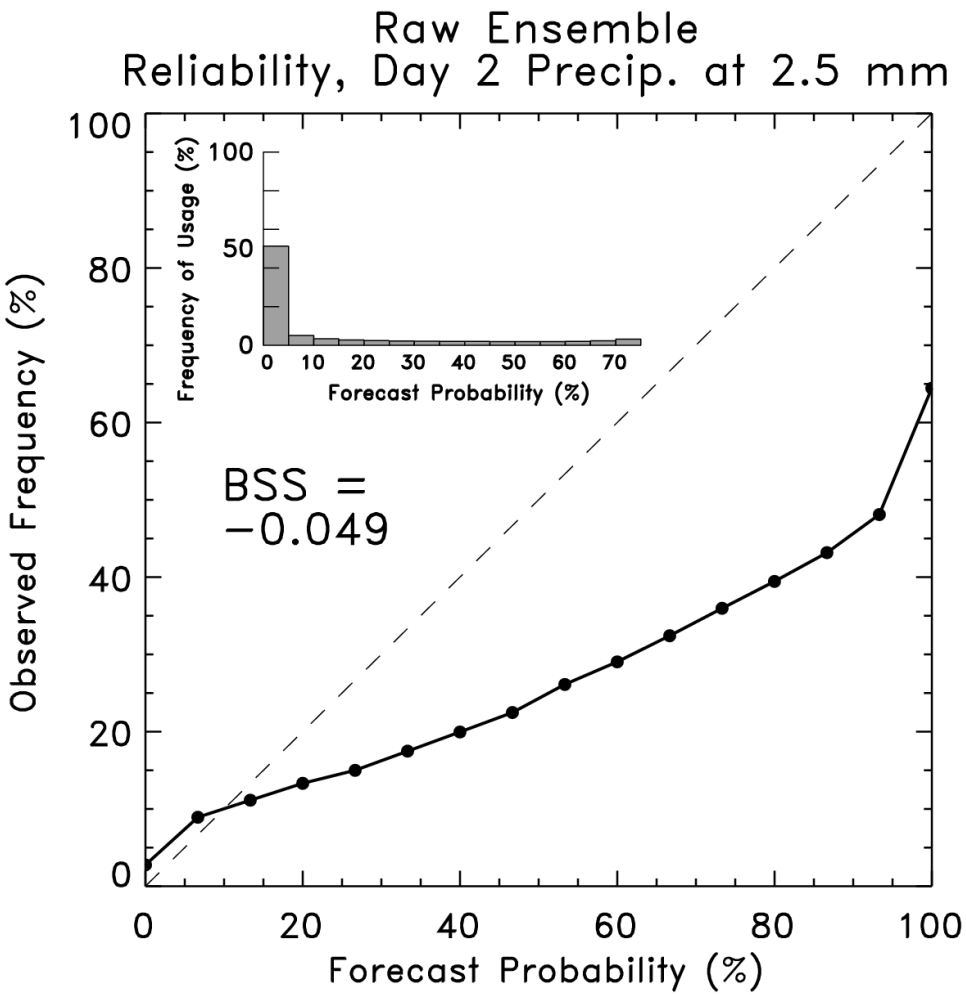


(b) 10 mm climatological probability, Gamma–distribution fitted, Jul

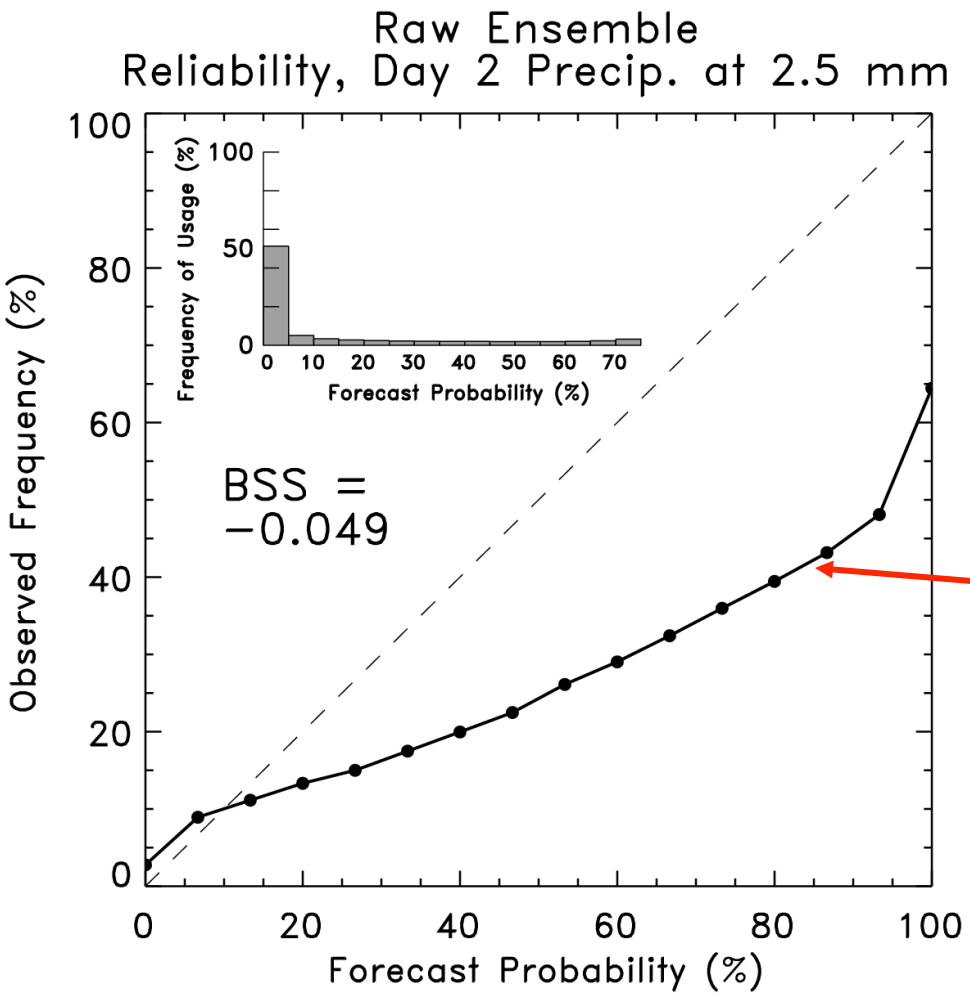


Climatological probability of > 10 mm / 24h, estimated from event relative frequency (top) and modeled (bottom)

Reliability diagrams



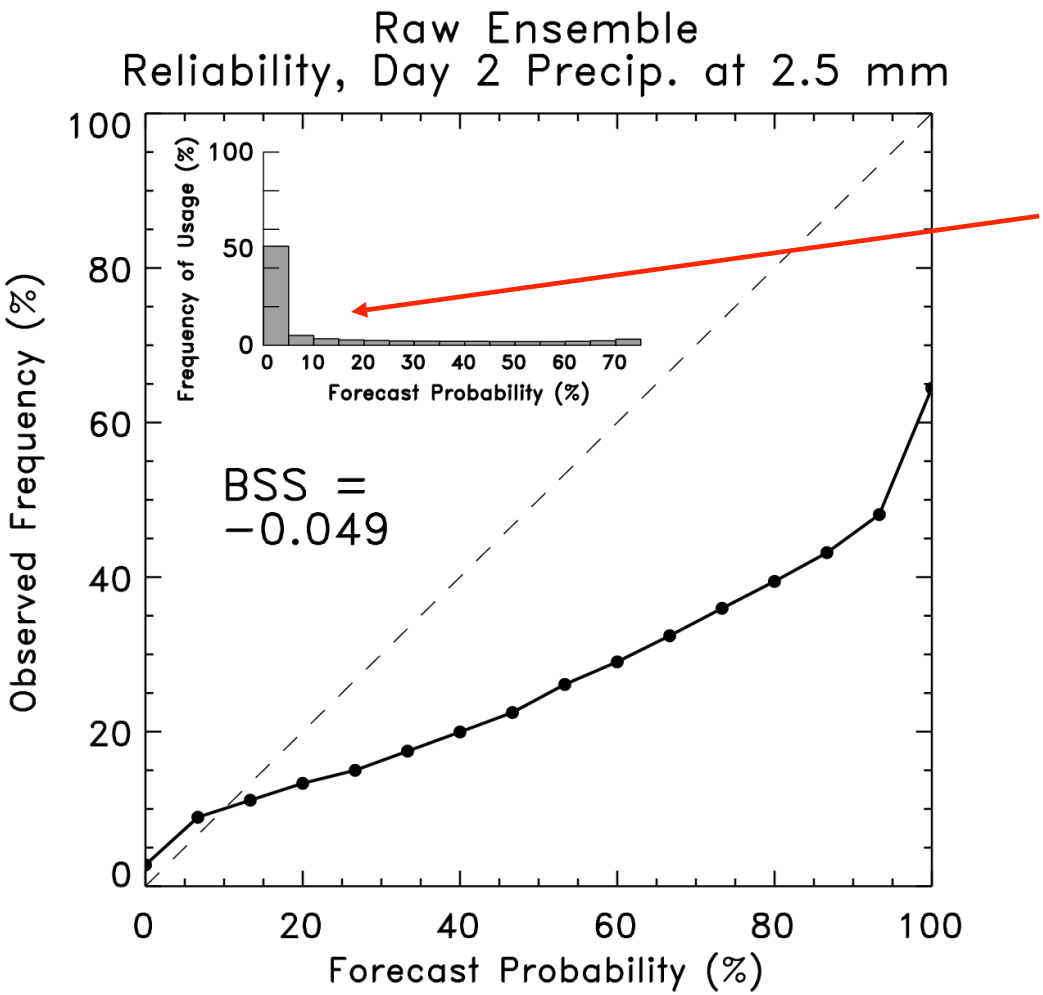
Reliability diagrams



Curve tells you what the observed frequency was each time you forecast a given probability. This curve ought to lie along $y = x$ line. Here this shows the ensemble-forecast system over-forecasts the probability of light rain.

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

Reliability diagrams

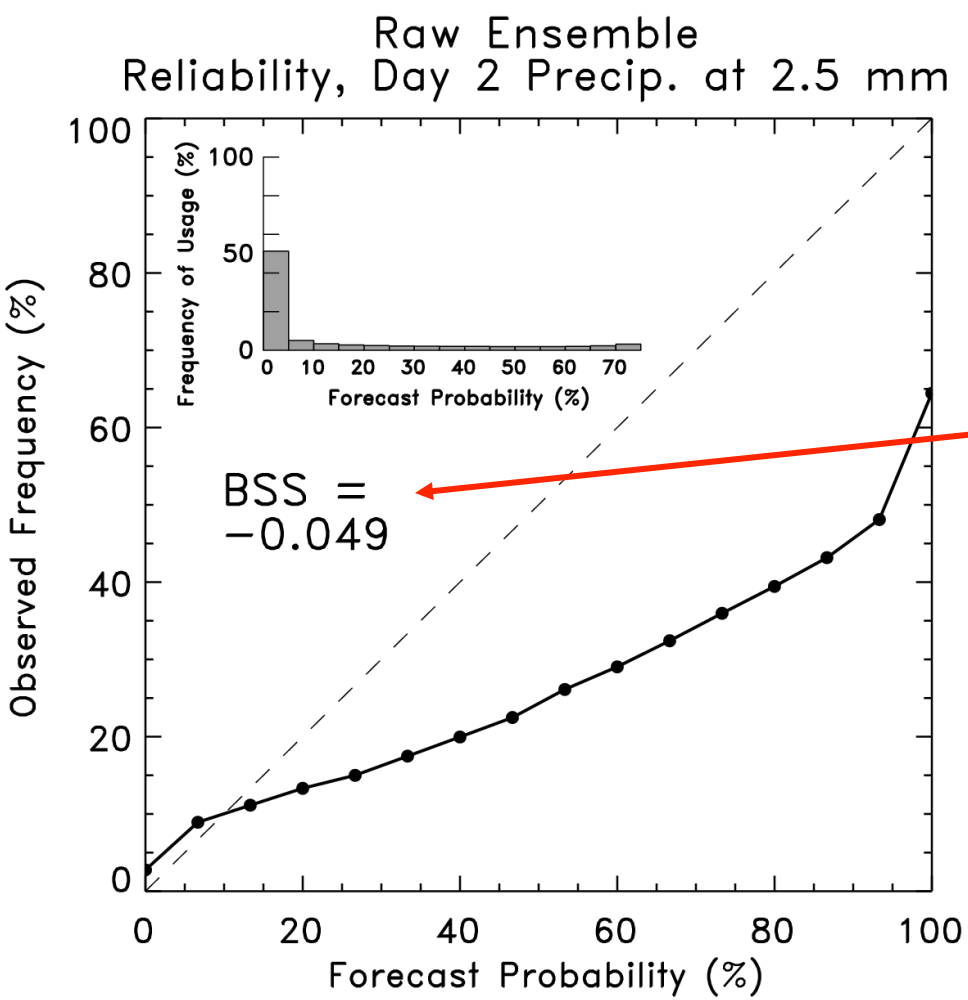


Inset histogram tells you how frequently each probability was issued.

Perfectly sharp: frequency of usage populates only 0% and 100%.

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

Reliability diagrams



BSS = Brier Skill Score

$$BSS = \frac{BS(Climo) - BS(Forecast)}{BS(Climo) - BS(Perfect)}$$

BS(•) measures the Brier Score, which you can think of as the squared error of a probabilistic forecast.

Perfect: *BSS* = 1.0

Climatology: *BSS* = 0.0

Ref: Wilks text, *Statistical Methods in the Atmospheric Sciences*

Equitable Threat Score

Forecast > Threshold ?	Observed > Threshold? YES	Observed > Threshold? NO
YES	a	b
NO	c	d

$$ETS = \frac{a - a_{ref}}{a - a_{ref} + b + c}$$

$$a_{ref} = (a+b)(a+c)/n$$

$$BIA = \frac{a+b}{a+c}$$

```
SUBROUTINE continuous_RPS_calc_ens (nmembers, fcstens,  obs, CRPS)

INTEGER, INTENT(IN) :: nmembers
REAL, INTENT(IN) :: fcstens(nmembers), obs
REAL, INTENT(OUT) :: CRPS
REAL x(400), fcdf(400), vcdf(400)

! ---- set up array of precip to calc cdf at (every 0.5 mm to 200 mm)
DO itemp= 1, 400
    x(itemp) = FLOAT(itemp-1)*0.5
END DO
! ---- cdf of observations, step function
vcdf = 0.
DO itemp = 1,400
    IF (x(itemp) .gt. obs) vcdf(itemp) = 1.
END DO
! ---- forecast cdf
fcdf(:) = 0.
DO itemp = 1, 400
  DO imem = 1, nmembers
   IF (x(itemp) .gt. fcstens(imem)) fcdf(itemp) = fcdf(itemp) + &
      1./FLOAT(nmembers)
  END DO
END DO
! ---- calculate CRPS
CRPS = 0.0
DO itemp = 1,400
  CRPS = CRPS + 0.5*(fcdf(itemp)-vcdf(itemp))**2
END DO
RETURN
END SUBROUTINE continuous_RPS_calc_ens
```