NOAA Earth System
Research Laboratory

Common verification methods for ensemble forecasts, and how to apply them properly

Tom Hamill

NOAA Earth System Research Lab,

Physical Sciences Division, Boulder, CO

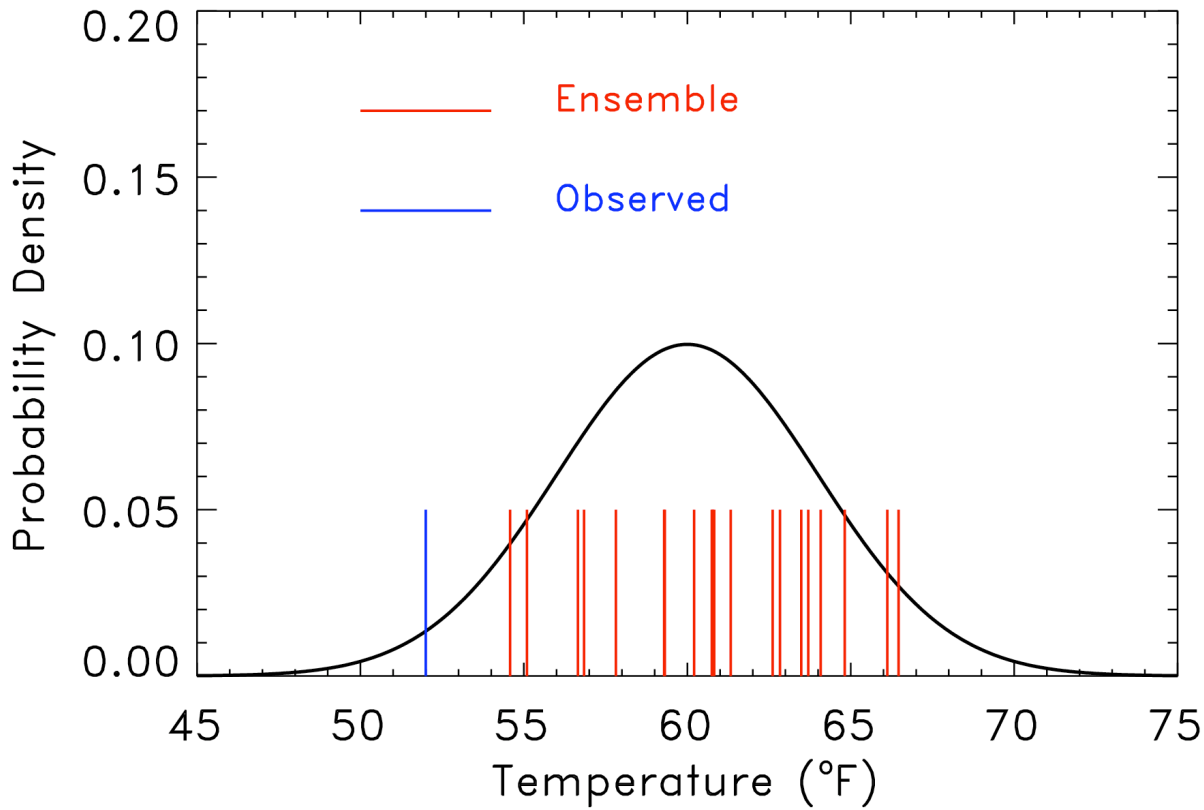
tom.hamill@noaa.gov

Part 1: two desirable properties of ensembles, and the challenges of evaluating these properties

- Property 1: **Reliability**, no matter how you slice and dice your ensemble data.

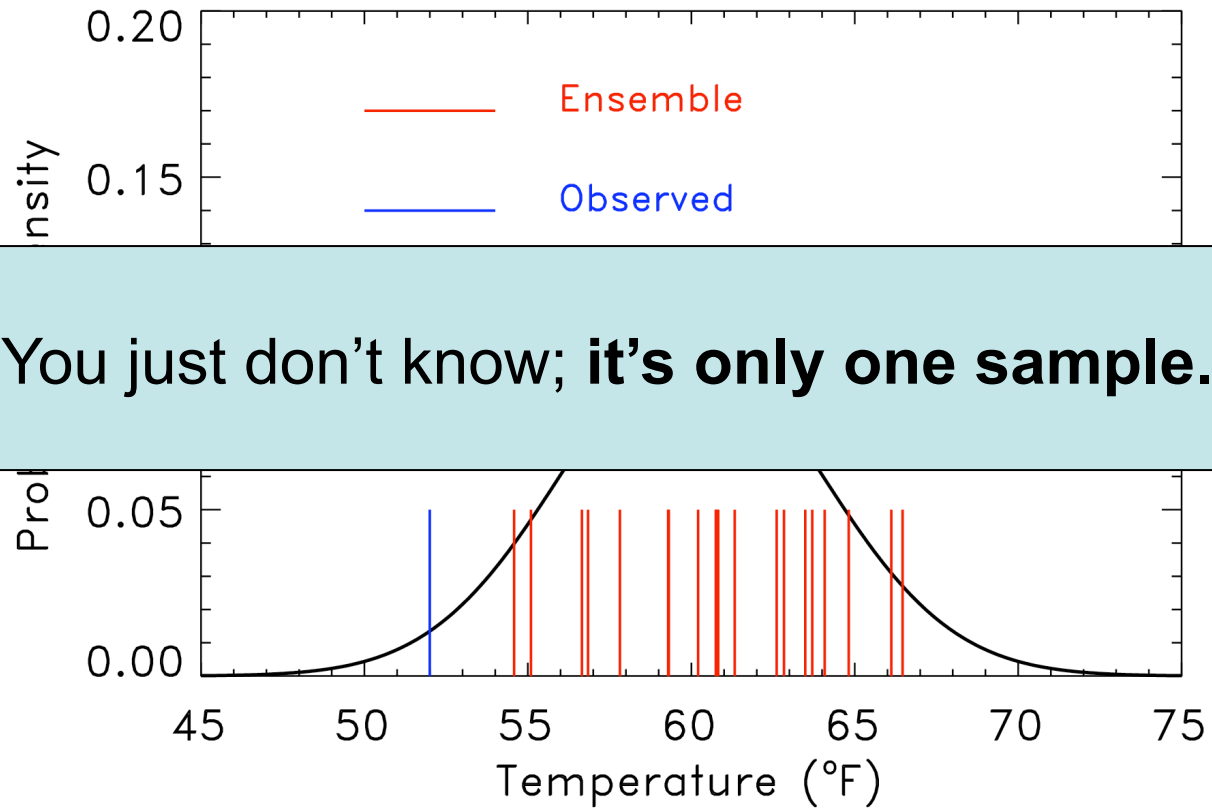
- Property 2: Specificity, i.e., **sharpness**.

Unreliable ensemble?



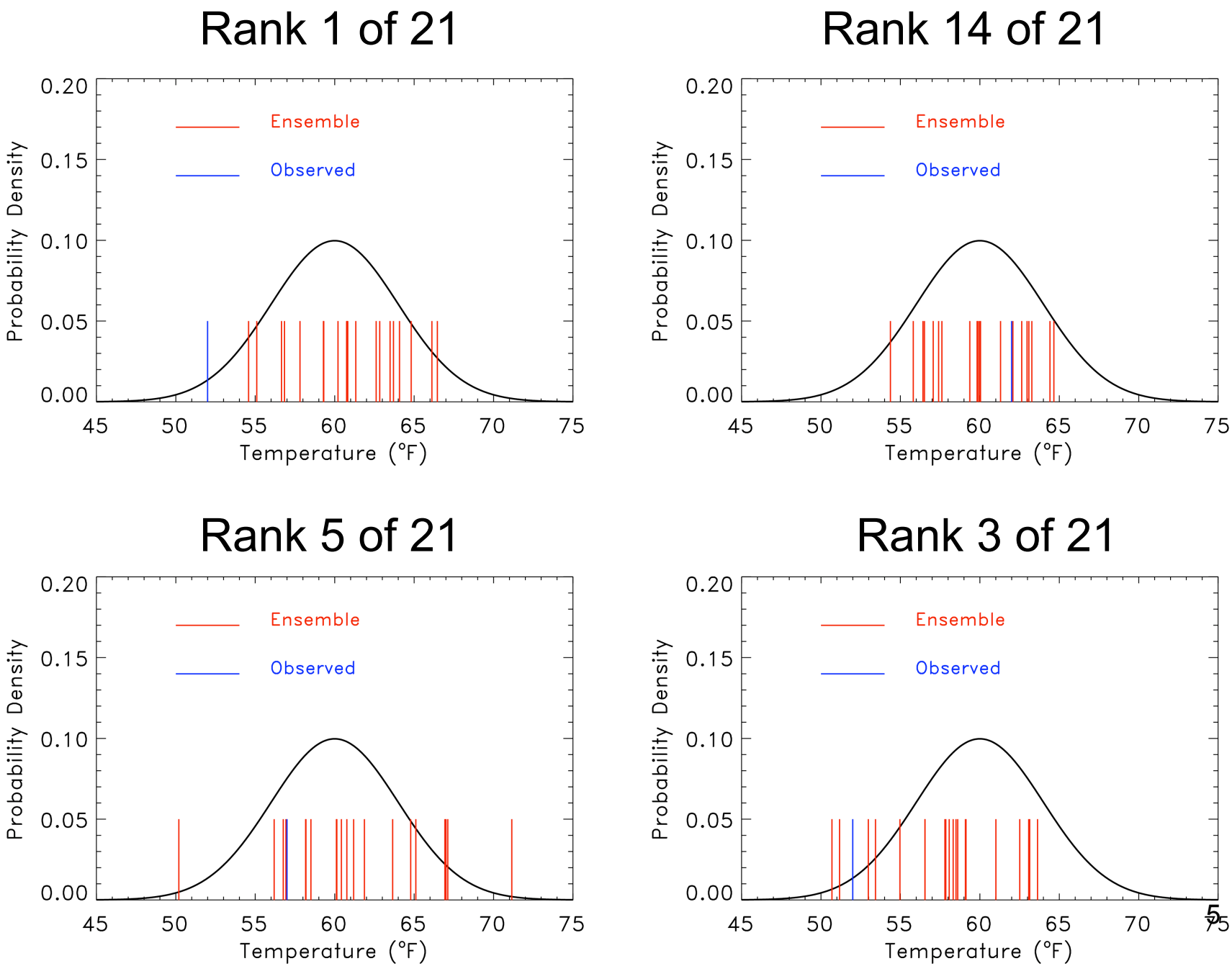
Here, the observed is outside of the range of the ensemble, which was sampled from the pdf shown. Is this a sign of a poor ensemble forecast?

Unreliable ensemble?



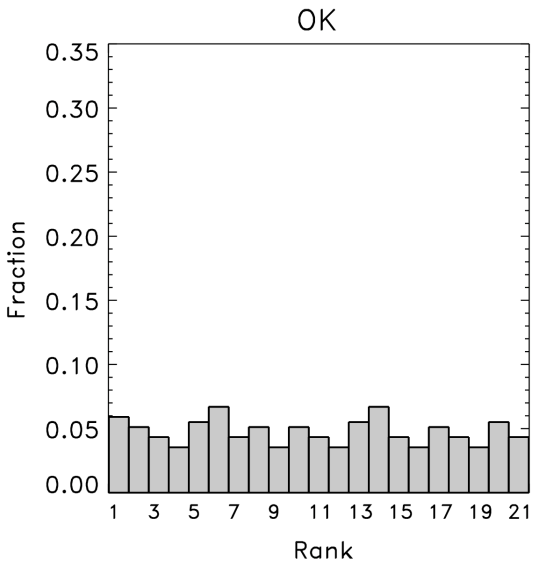
You just don't know; **it's only one sample.**

Here, the observed is outside of the range of the ensemble, which was sampled from the pdf shown. Is this a sign of a poor ensemble forecast?

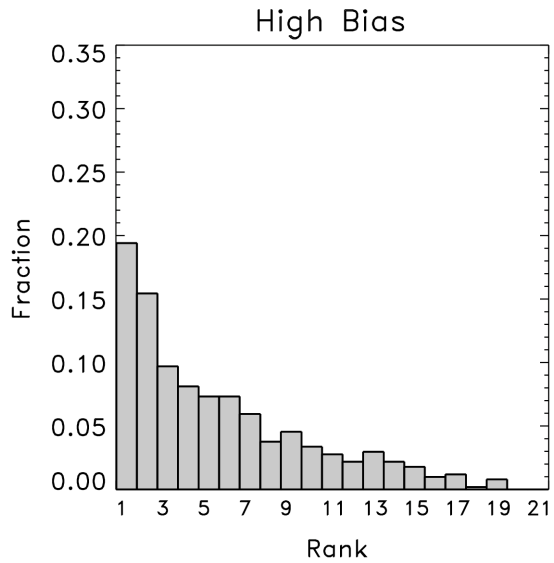

Rank 1 of 21
Rank 14 of 21
Rank 5 of 21
Rank 3 of 21
Ensemble
Observed
Probability Density
0.00
0.05
0.10
0.15
0.20
45
50
55
60
65
70
75
Temperature (°F)

“Rank histograms,” aka “Talagrand diagrams”

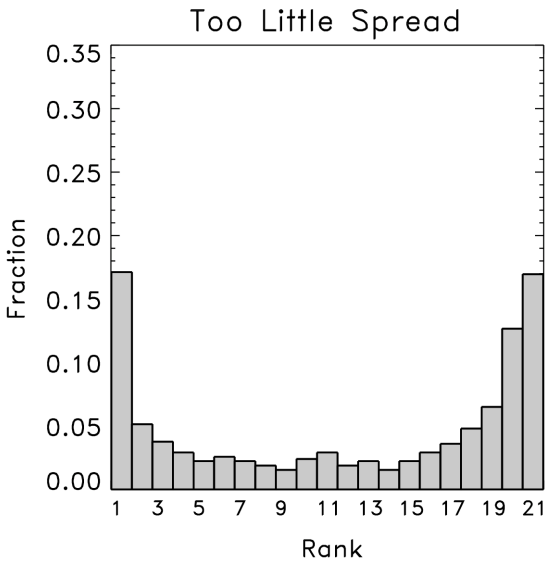
With **lots of samples** from many situations, can evaluate the characteristics of the ensemble.



Happens when observed is indistinguishable from any other member of the ensemble. Ensemble hopefully is reliable.



Happens when observed too commonly is lower than the ensemble members.



Happens when there are either some low and some high biases, or when the ensemble doesn’t spread out enough.

ref: Hamill, *MWR*, March 2001

Underlying mathematics

$$\mathbf{X} = (x_1, \ldots, x_n)$$ ← n-member *sorted* ensemble at some point

$$E\left[P\left(V < x_i\right)\right] = \frac{i}{n+1}$$ ← probability the truth V is less than the ith sorted member if V and **X**'s members sample the same distribution

$$E\left[P\left(x_{i-1} \le V < x_i\right)\right] = \frac{1}{n+1}$$ ← … equivalently

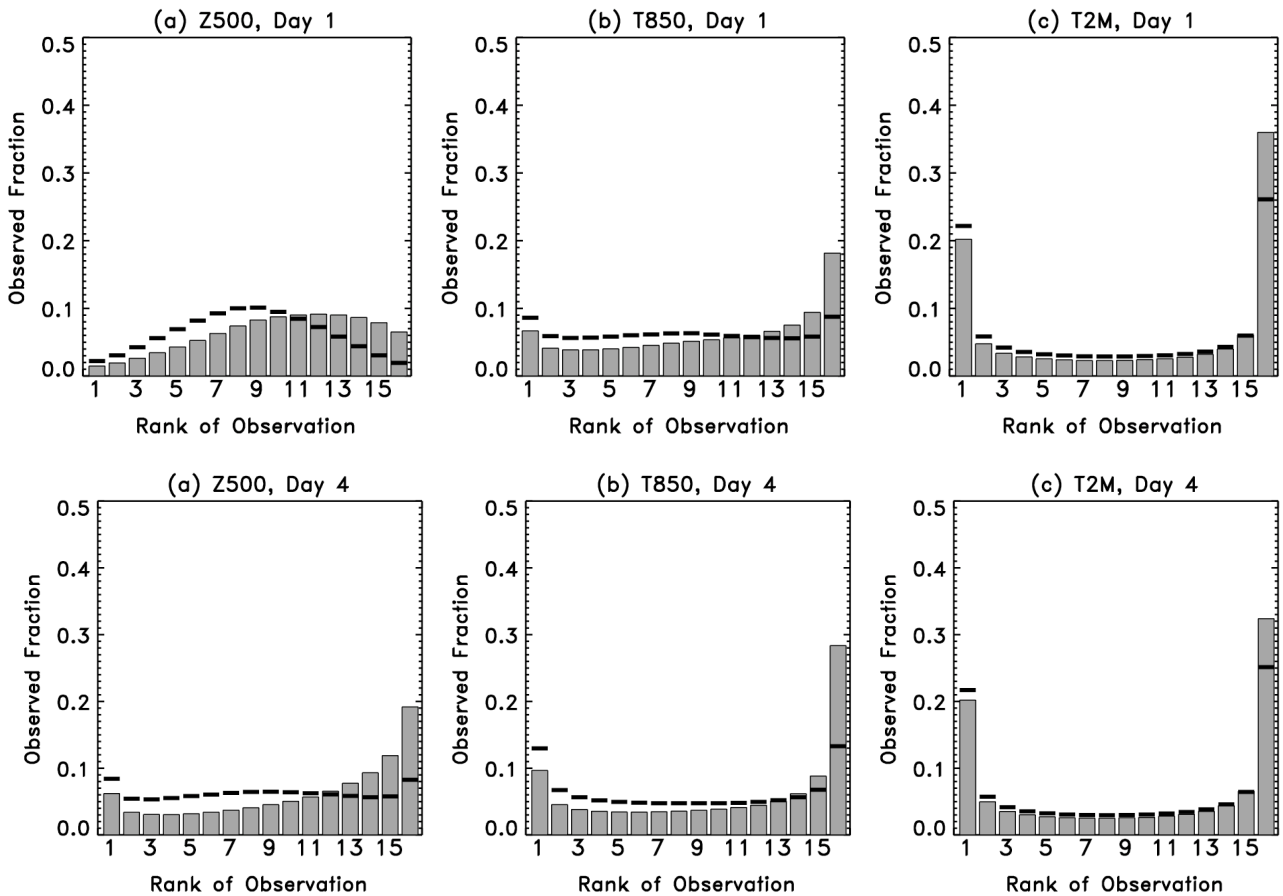
$$\mathbf{R} = \left(r_1, \ldots, r_{n+1}\right)$$ ← our rank histogram vector

$$r_j = \overline{P\left(x_{j-1} \le V < x_j\right)}$$ ← overbar denotes the sample average over many (hopefully independent) samples

ref: ibid.

Rank histograms of Z_{500}, T_{850}, T_{2m}

(from 1998 reforecast version of NCEP GFS)



Solid lines indicate ranks after bias correction.

Rank histograms are particularly U-shaped for T2M.

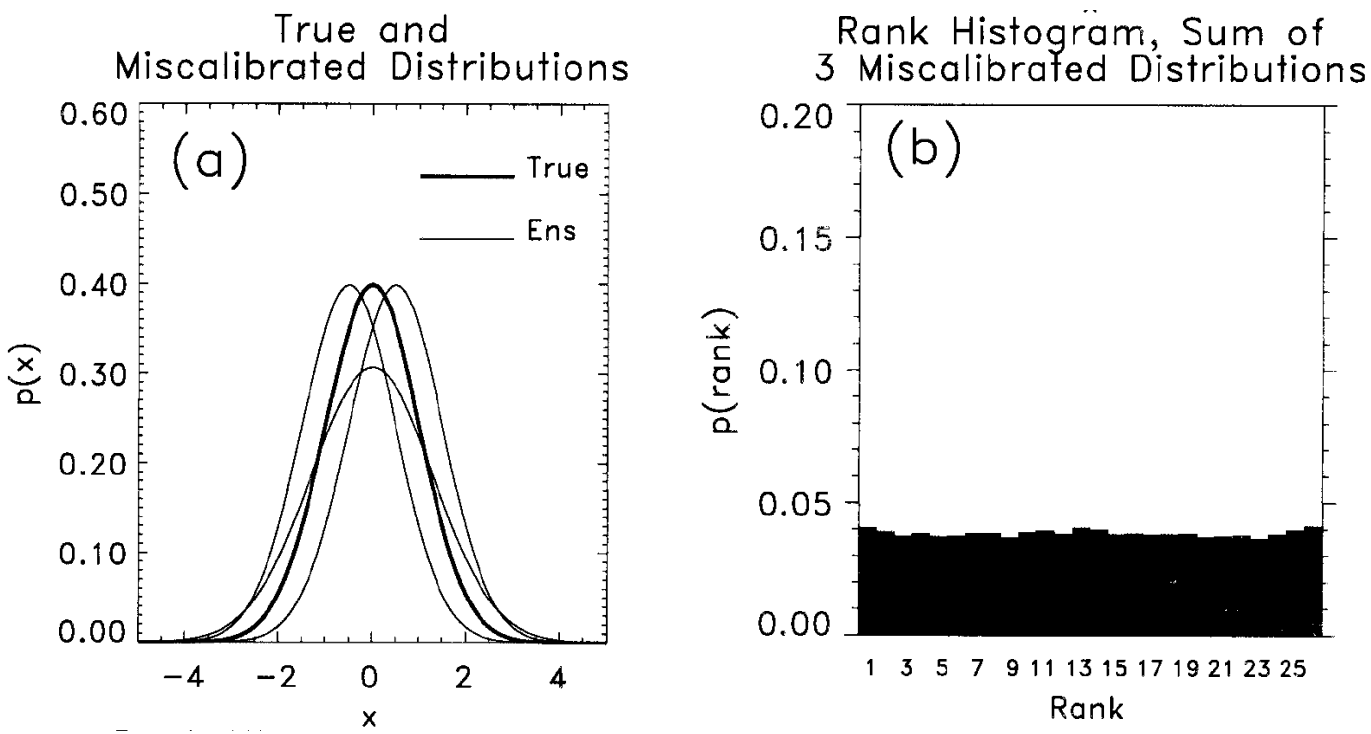
Ref: Hamill and Whitaker, MWR, Sep 2007.

Rank histograms… pretty simple, right?

Let’s consider some of the issues involved with this one seemingly simple verification metric.

Issues in using rank histograms

(1) Flat rank histograms can be created from combinations of uncalibrated ensembles.



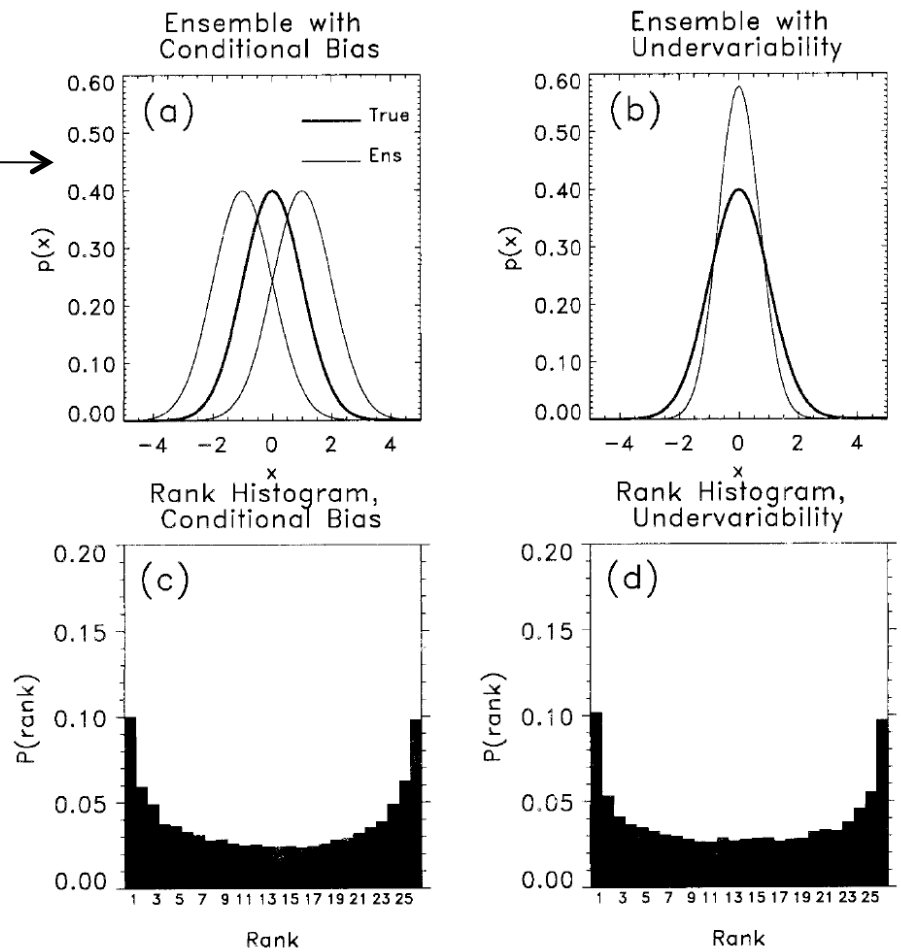
Lesson: you only be confident you have reliability if you see flatness of rank histogram having sliced the data many different ways.

Ref: Hamill, MWR, Mar 2001

Issues in using rank histograms

(2) Same rank histogram shape may result from ensembles with different deficiencies

Here, half of ensemble members from low-biased distribution, half from high-biased distribution

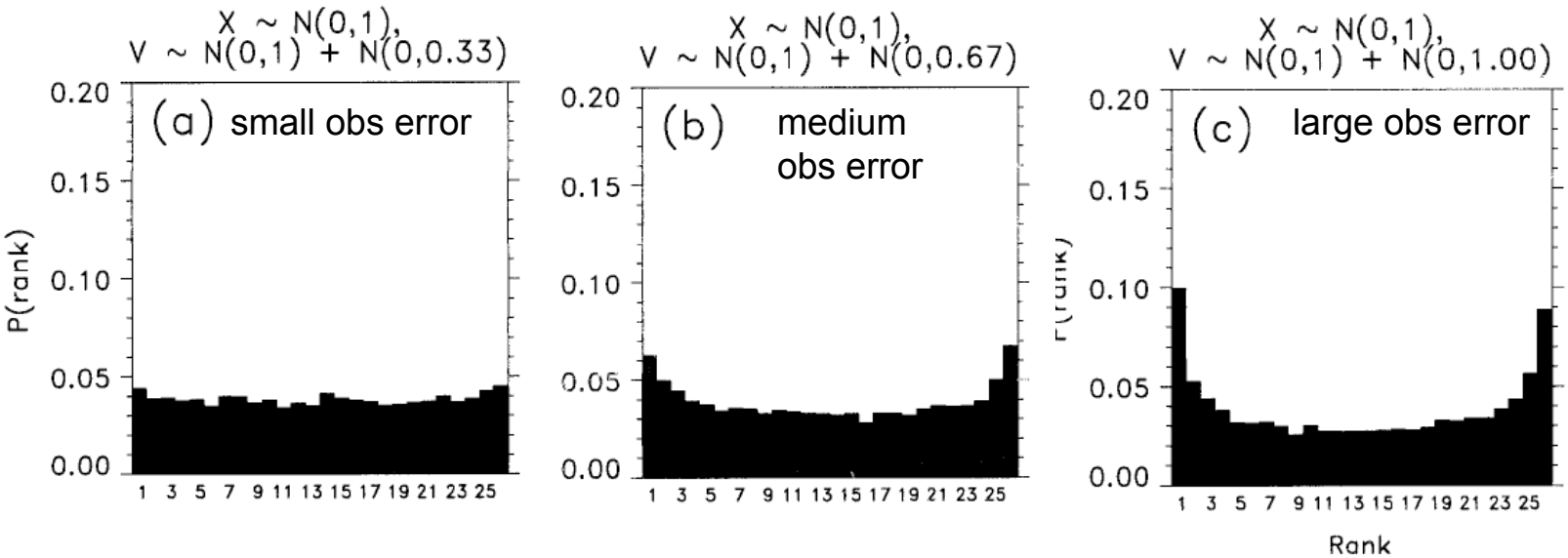


Here, all of ensemble members from under-spread distribution.

Ref: ibid

Issues in using rank histograms

(3) Evaluating ensemble relative to observations with errors distorts shape of rank histograms.



A solution of sorts is to dress every ensemble member with a random sample of noise, consistent with the observation errors.

Ref: ibid.

Rank histogram shapes under "location" and "scale" errors

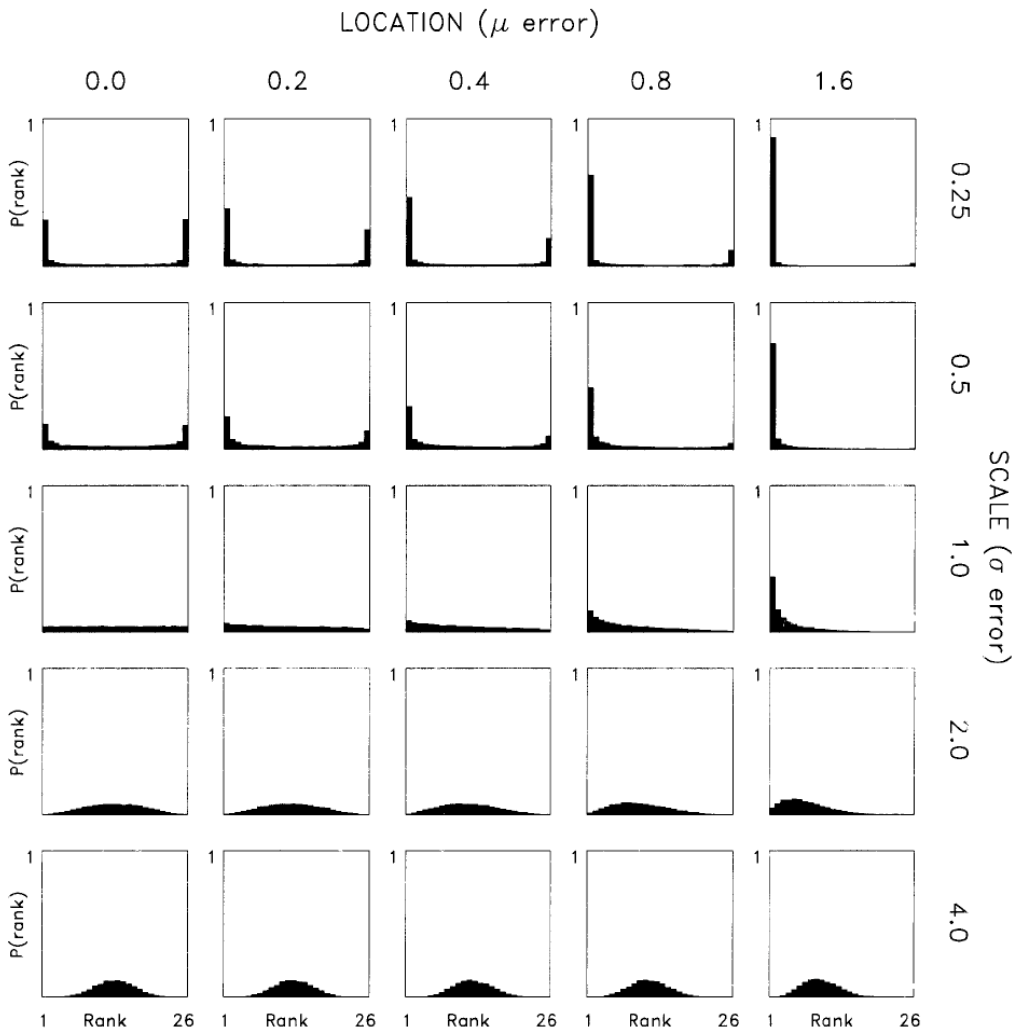


FIG. 1. Rank histograms where verification is sampled from a $N(0, 1)$ distribution and the ensemble ($n = 25$ members) is sampled from a $N(\mu, \sigma)$ distribution. The rank of the verification is tallied 10 000 times in each panel.

truth from N(0,1)

forecast from N(μ,σ); no implicit correlations between members or between member and truth

Ref: ibid.

Caren Marzban's question: what if we did have correlations between members, and w. truth?

$$Truth\ y; \quad ensemble\left(x_1, \ldots, x_n\right)$$

$$\left(y, x_1, \ldots, x_n\right) \sim MVN\left(\left(0, \mu_x, \ldots, \mu_x\right), \Sigma\right)$$

$$\Sigma = \begin{bmatrix} 1 & R\sigma_x & R\sigma_x & \ldots & R\sigma_x \\ & \sigma_x^2 & r\sigma_x^2 & \ldots & r\sigma_x^2 \\ & & \sigma_x^2 & \ldots & r\sigma_x^2 \\ & & & \ddots & \vdots \\ & & & & \sigma_x^2 \end{bmatrix}$$

r is correlation between ensemble members

R is correlation between ensemble member and the observation

Ref: Marzban et al., MWR 2010, in press.

Rank histograms, $r = R = 0.9$

Marzban's rank histograms include box & whiskers to quantify sampling variability (nice!).

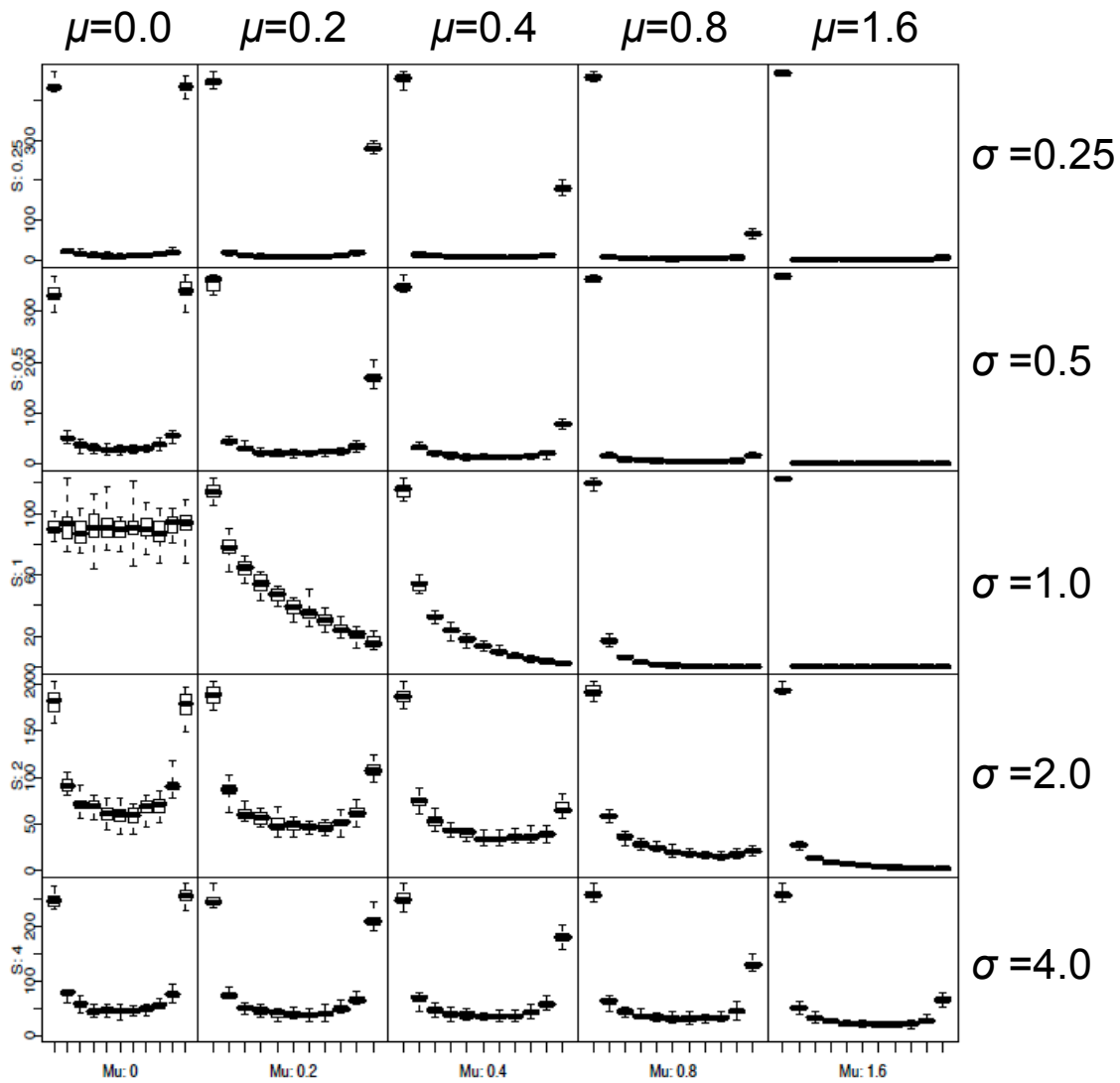


FIG. 5. Rank Histograms for different values of the common forecast mean μ (along x-axis) and variance within ensemble member σ (along y-axis), when there is a strong correlation between ensemble members and an equally strong correlation between the ensemble members and the observation (i.e., $R = r = 0.9$).

Rank histograms, $r = R = 0.9$

Marzban's rank histograms include box & whiskers to quantify sampling variability (nice!).

What's going on here? Why do we now appear to have diagnosed under-spread with an ensemble with apparent excess spread?

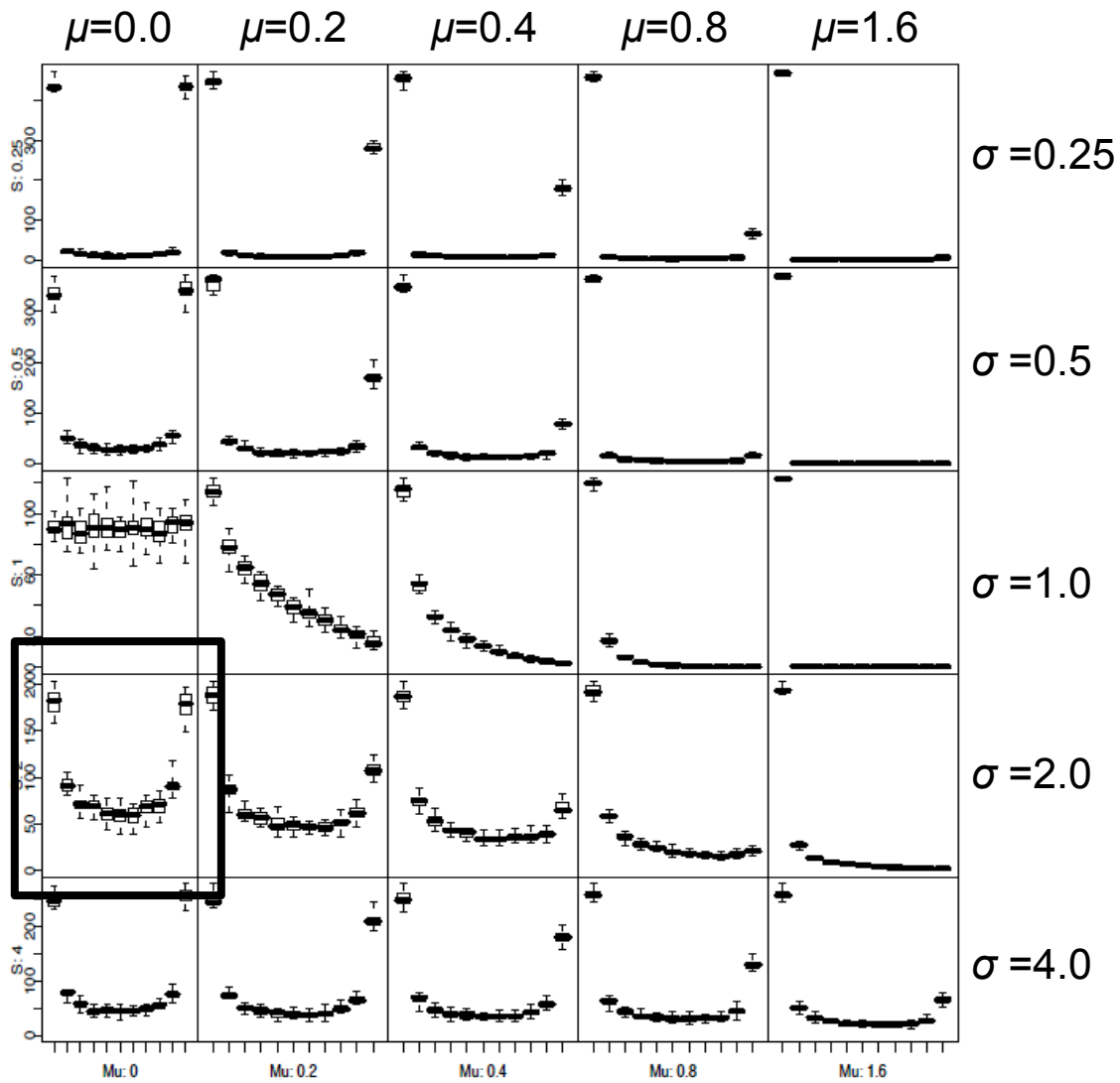


FIG. 5. Rank Histograms for different values of the common forecast mean μ (along x-axis) and variance within ensemble member σ (along y-axis), when there is a strong correlation between ensemble members and an equally strong correlation between the ensemble members and the observation (i.e., $R = r = 0.9$).

Dan Wilks' insight into this curious phenomenon…

Property of multivariate Gaussian distribution:

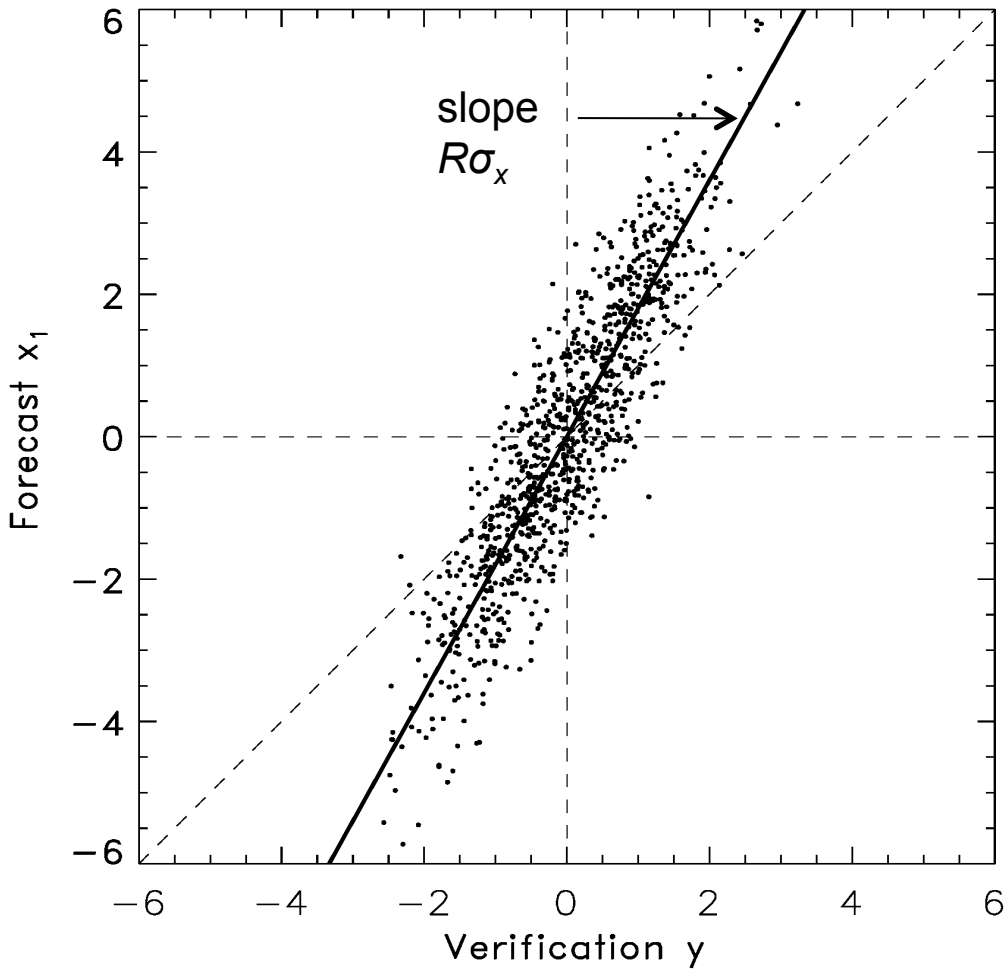
$$\mu_{x|y} = \mu_x + R\frac{\sigma_x}{\sigma_y}\left(y - \underset{\uparrow}{\mu_y}\right) = \mu_x + R\frac{\sigma_x}{\sigma_y}y$$

(= 0, since in this case assumed y ~ N(0,1))

When μ_x= 0 (unbiased forecast), and given σ_y =1, expected value for ensemble member is $R\sigma_x y$, which will be more extreme (further from origin) than y for relatively large σ_x.

Ref: Wilks, 2010 MWR, submitted.

Illustration, $\mu_x=0$



Many realizations of truth and synthetic 1-member ensemble were created, $\mu_x = 0, \sigma_x = 2$, $r = R = 0.9$

Here, for a 1-member ensemble, that forecast member x_1 tends to be further from the origin than the verification y. If we generated another ensemble member, it would tend to be further from the origin in the same manner. Hence, you have an ensemble clustered together, away from the origin, and the verification near the origin, i.e., at extreme ranks relative to the sorted ensemble.

Rank histograms in higher dimensions: the "*minimum spanning tree*" histogram

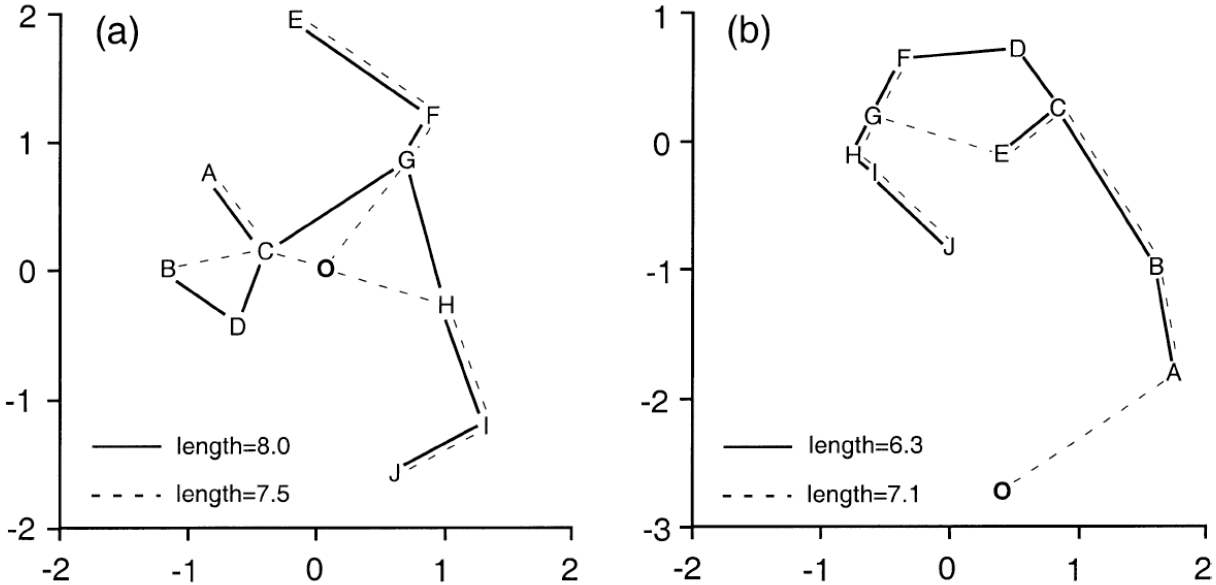


FIG. 1. Hypothetical example MSTs in $K = 2$ dimensions. The $n_{ens} = 10$ ensemble members are labeled A–J, and the corresponding observation is O. Solid lines indicate MSTs for the ensemble as forecast, and dashed lines indicate MSTs that result from the observation being substituted for ensemble member D. (a) A configuration that could result from an overdispersed ensemble, where the observation is interior to the point cloud of the ensemble. (b) A configuration that could result from an underdispersed ensemble and/or a substantial ensemble mean error.

- Solid lines: minimum spanning tree (MST) between 10-member forecasts
- Dashed line: MST when observed O is substituted for member D
- Calculate MST's sum of line segments for all forecasts, and observed replacing each forecast member. Tally rank of pure forecast sum relative to sum where observed replaced a member.
- Repeat for independent samples, build up a histogram

Ref: Wilks, MWR, June 2004. See also Smith and Hansen, MWR, June 2004

Rank histograms in higher dimensions: the "*minimum spanning tree*" histogram

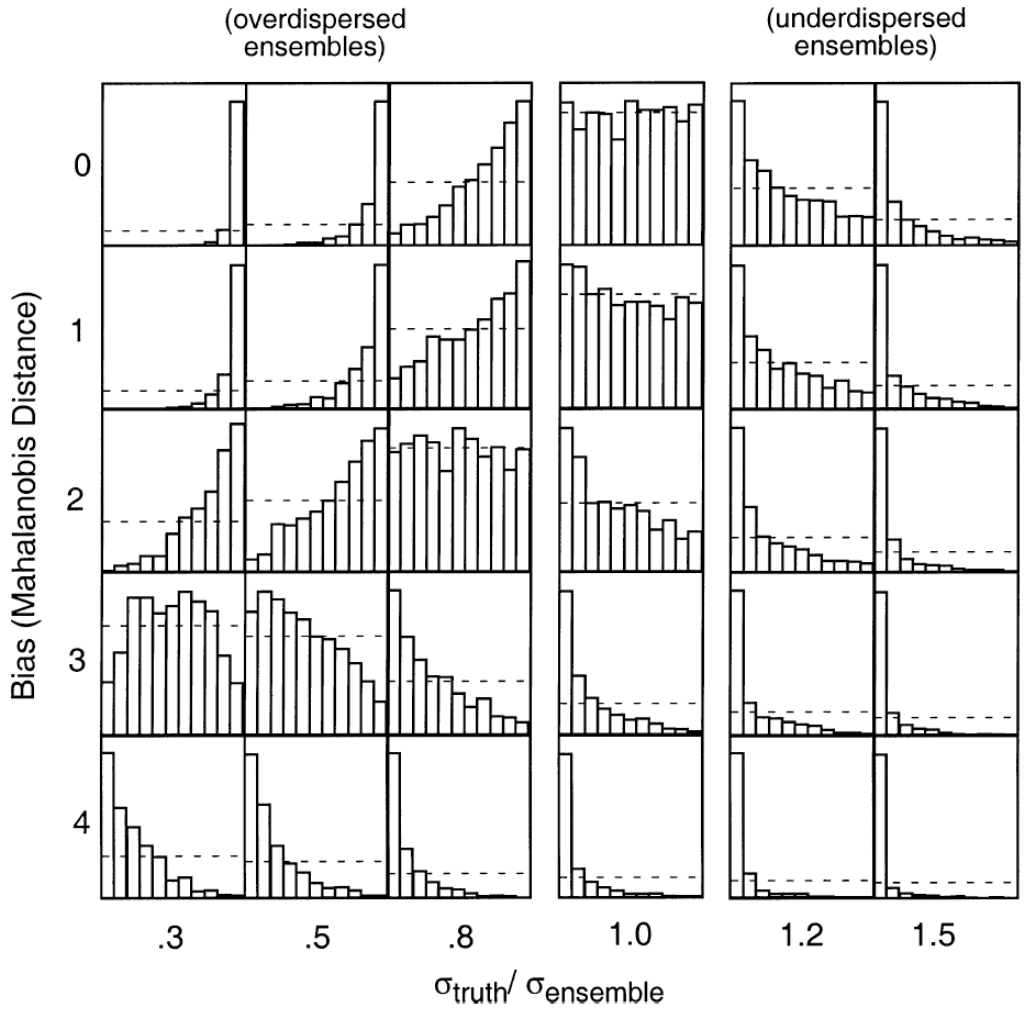
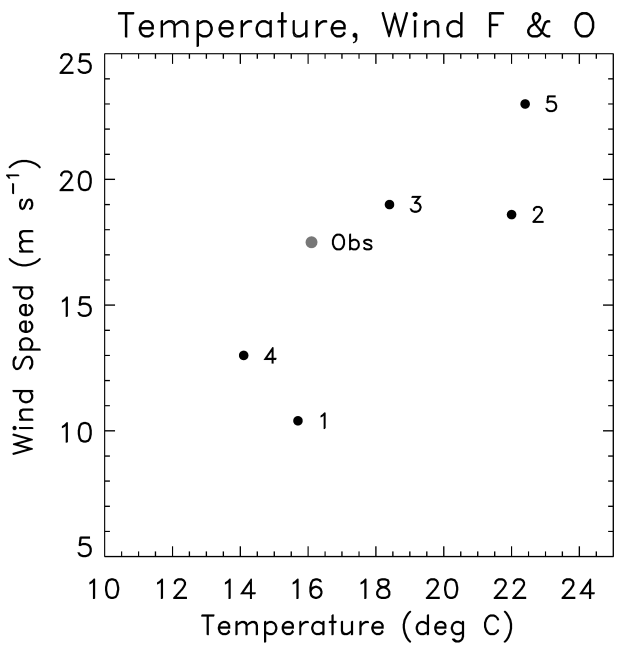


FIG. 2. Behaviors of MST histograms for $n_{ens} = 10$ in $K = 10$ dimensions, as functions of ensemble bias (vertical) and ensemble underdispersion (horizontal), from independent samples of size $n = 1000$. Vertical scales on each histogram have been varied for clarity of presentation, with the level of the expected number per bin under uniformity (1000/11 = 91) indicated in each case by the dashed line.

- Graphical interpretation of MST is different and perhaps less intuitive than it is for uni-dimensional rank histogram.

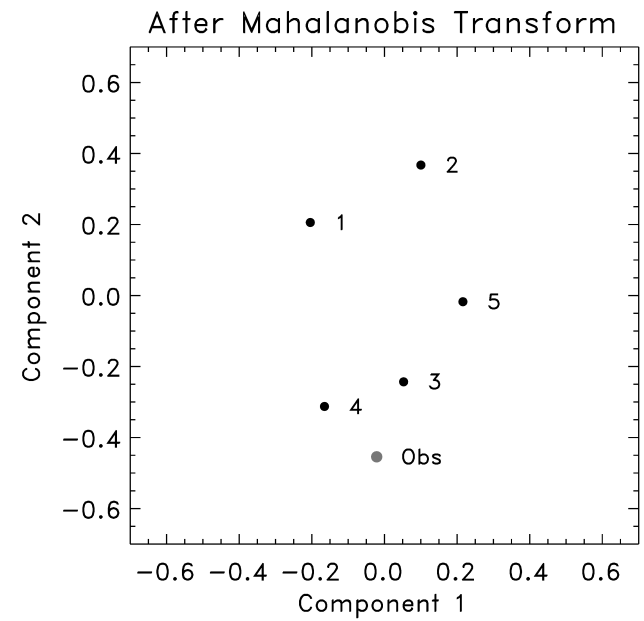
Ref: Wilks, MWR, June 2004. See also Smith and Hansen, MWR, June 2004

Multi-variate rank histogram



"Mahalanobis" transform (S is forecasts' sample covariance)

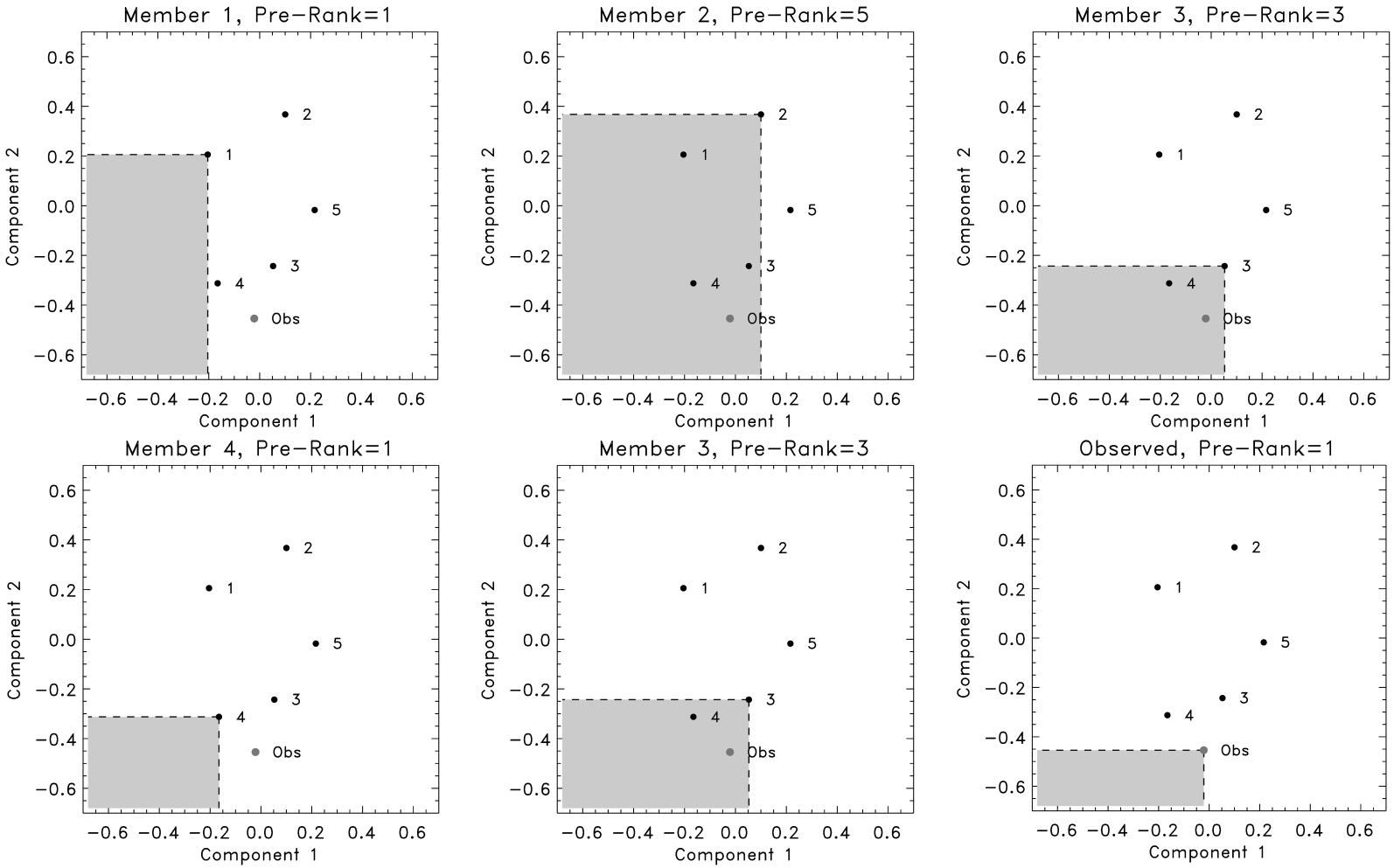
$$z_i = [S]^{-1/2}(x_i - \bar{x})$$



- Standardize and rotate using Mahalanobis transformation (see Wilks 2006 text).
- For each of *n* members of forecast and observed, define "pre-rank" as the number of vectors to its lower left (a number between 1 and *n*+1)
- The multi-variate rank is the rank of the observation pre-rank, with ties resolved at random
- Composite multi-variate ranks over many independent samples and plot rank histogram.
- Same interpretation as scalar rank histogram (e.g., U-shape = under-dispersive).

based on Tilmann Gneiting's presentation at Probability and Statistics, 2008 AMS Annual Conf., New Orleans.
and http://www.springerlink.com/content/q58j4167355611g1/

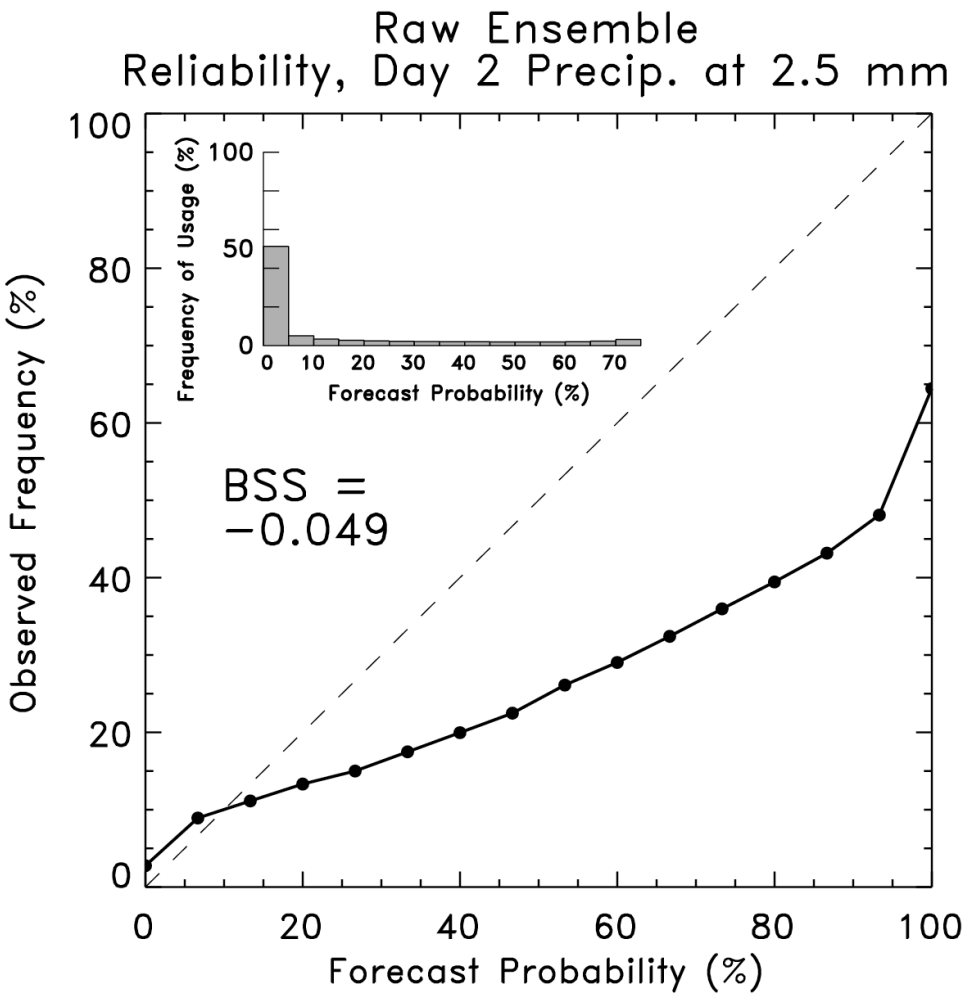
Multi-variate rank histogram calculation



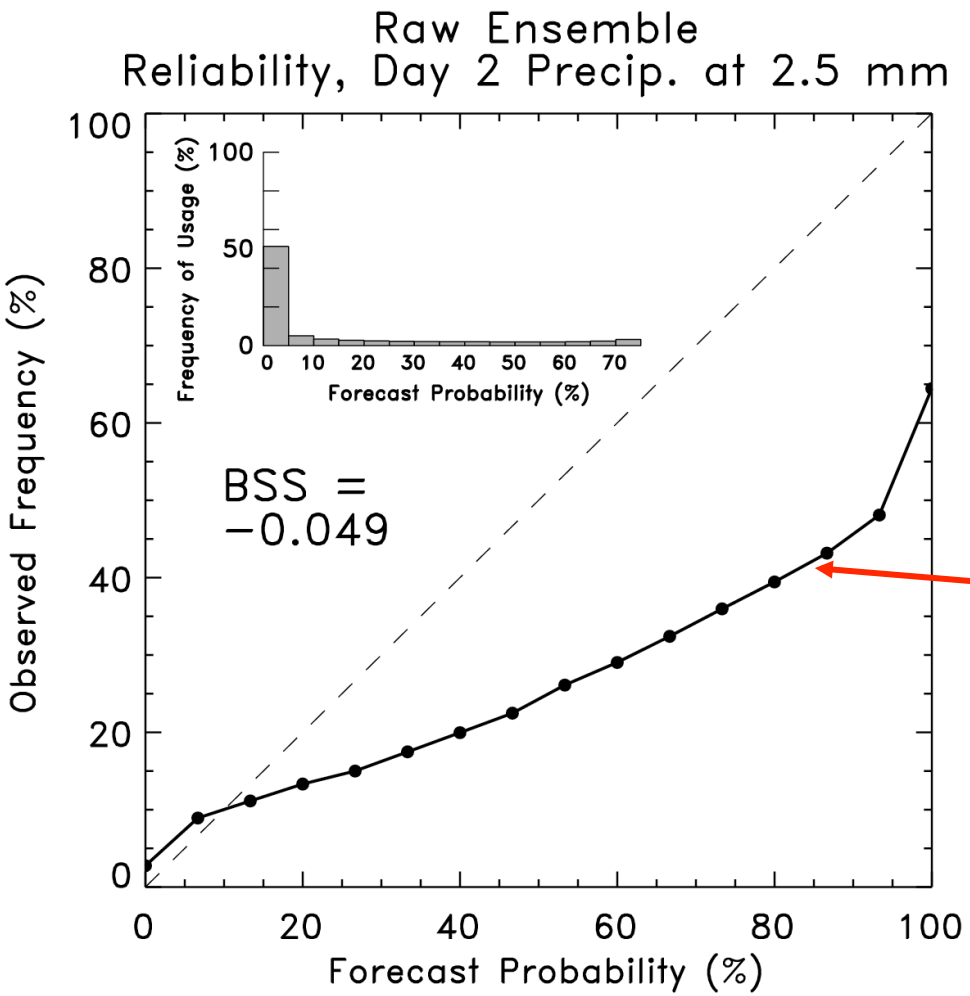
F_1, F_2, F_3, F_4, F_5, O pre-ranks: [1, 5, 3, 1, 4, 1] ; sorted: obs = either rank 1, 2, or 3 with p=1/3.

based on Tilmann Gneiting's presentation at Probability and Statistics, 2008 AMS Annual Conf., New Orleans

Reliability diagrams



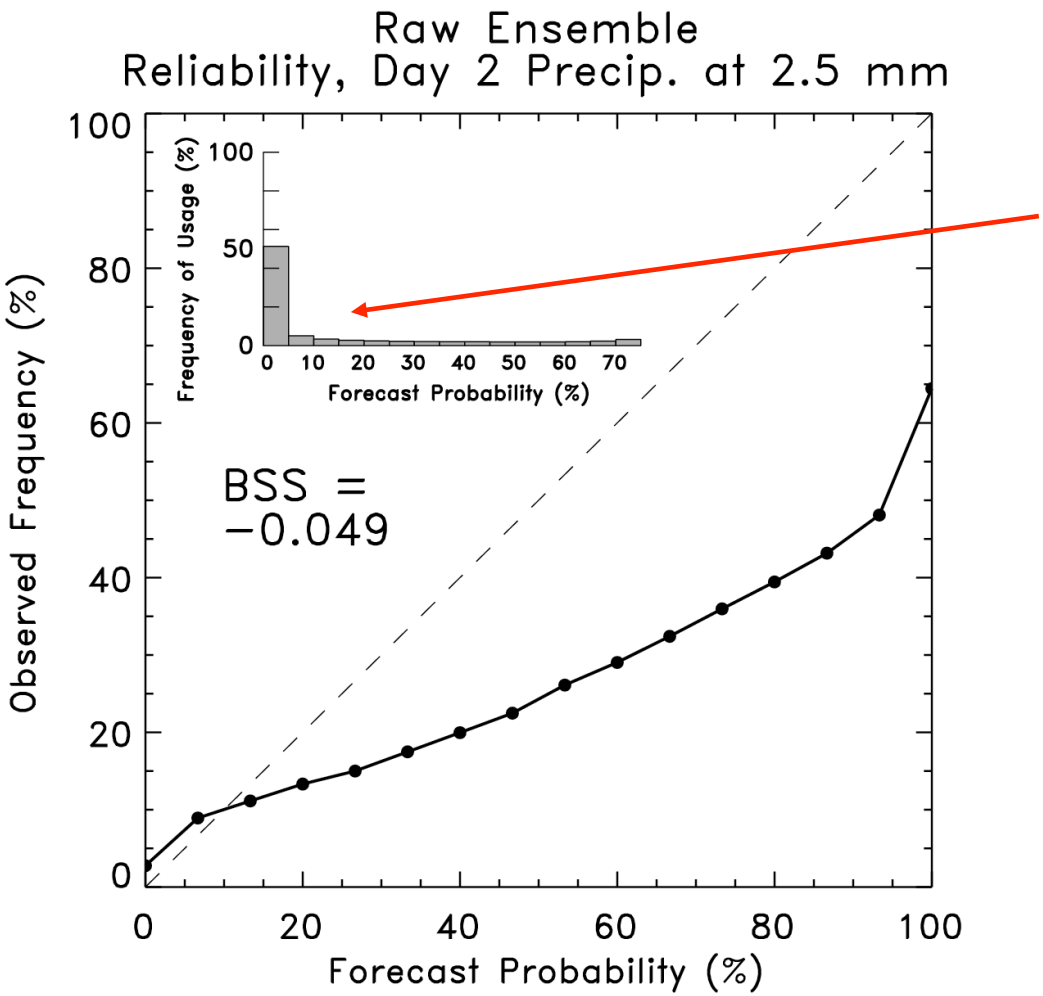
Reliability diagrams



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

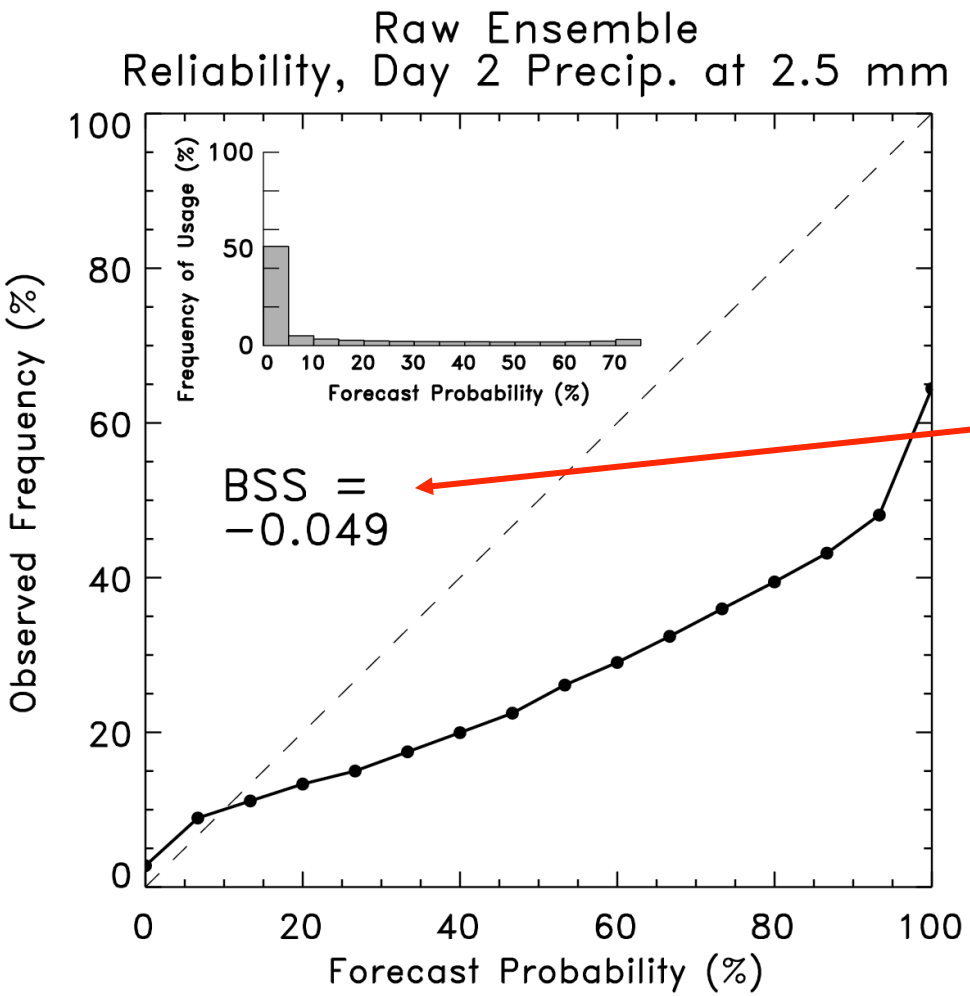


Inset histogram tells you how frequently each probability was issued.

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*

Brier score

- Define an event, e.g., precip > 2.5 mm.
- Let y_i be the forecast probability for the ith forecast case.
- Let o_i be the observed probability (1 or 0). Then

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

(So the Brier score is the averaged squared error of the probabilistic forecast)

Ref: Wilks 2006 text

Brier score decomposition

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

*S*uppose I allowable forecast values, e.g., I=10 for (5%, 15%, ... , 95%). Frequency of usage N_i.

Suppose overall climatological probability $\overline{o}$, and conditional average observed probability $\overline{o}_i$ for *I*th value.

$$\mathrm{BS} = \frac{1}{n}\sum_{i=1}^{I} N_i\left(y_i - \overline{o}_i\right)^2 - \frac{1}{n}\sum_{i=1}^{I} N_i\left(\overline{o}_i - \overline{o}\right)^2 + \overline{o}\left(1 - \overline{o}\right)$$

("reliability") ("resolution") ("uncertainty")

Ref: 2006 Wilks text

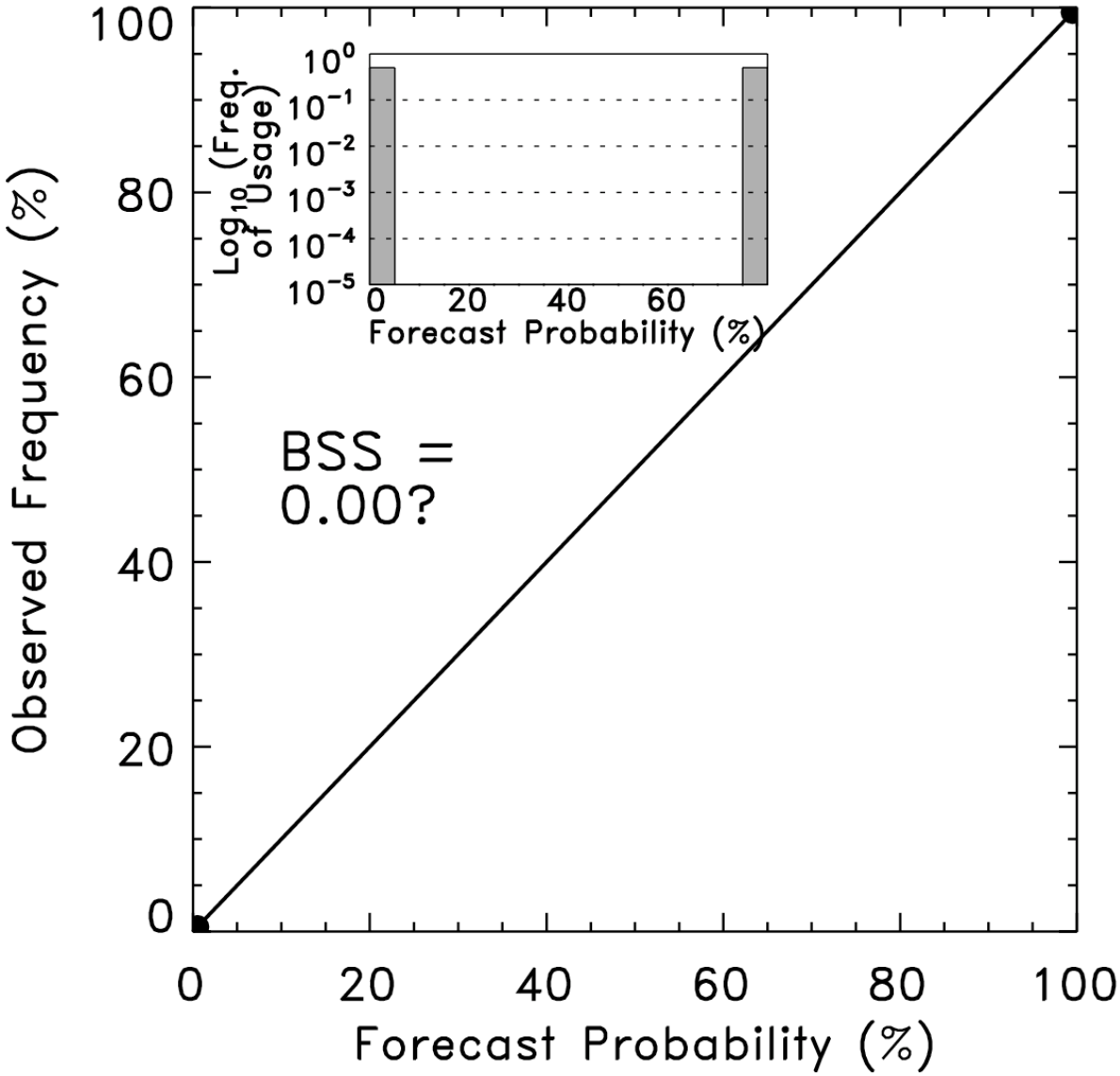
Brier score decomposition

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

Decomposition only makes sense when every sample is drawn from a distribution with an overall climatology of $\overline{o}$. For example, don't use if your forecast sample mixes together data from both desert and rainforest locations. For more, see Hamill and Juras, Oct 2006 QJ

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i\left(y_i - \overline{o}_i\right)^2}_{\text{("reliability")}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i\left(\overline{o}_i - \overline{o}\right)^2}_{\text{("resolution")}} + \underbrace{\overline{o}\left(1 - \overline{o}\right)}_{\text{("uncertainty")}}$$

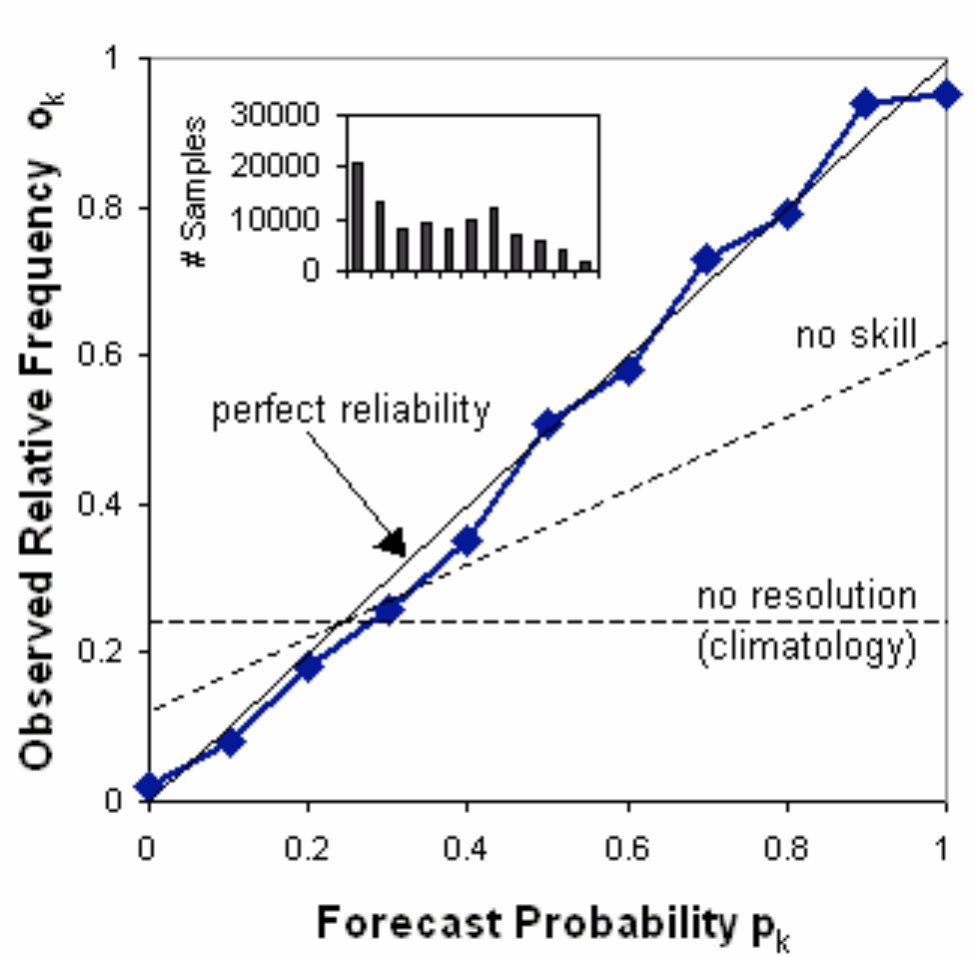
Perfectly Sharp, Perfect Reliability: Is BSS 1.0 or 0.0?



Degenerate case:

Skill might appropriately be 0.0 if all samples with 0.0 probability are drawn from climatology with 0.0 probability, and all samples with 1.0 are drawn from climatology with 1.0 probability.

“Attributes diagram”



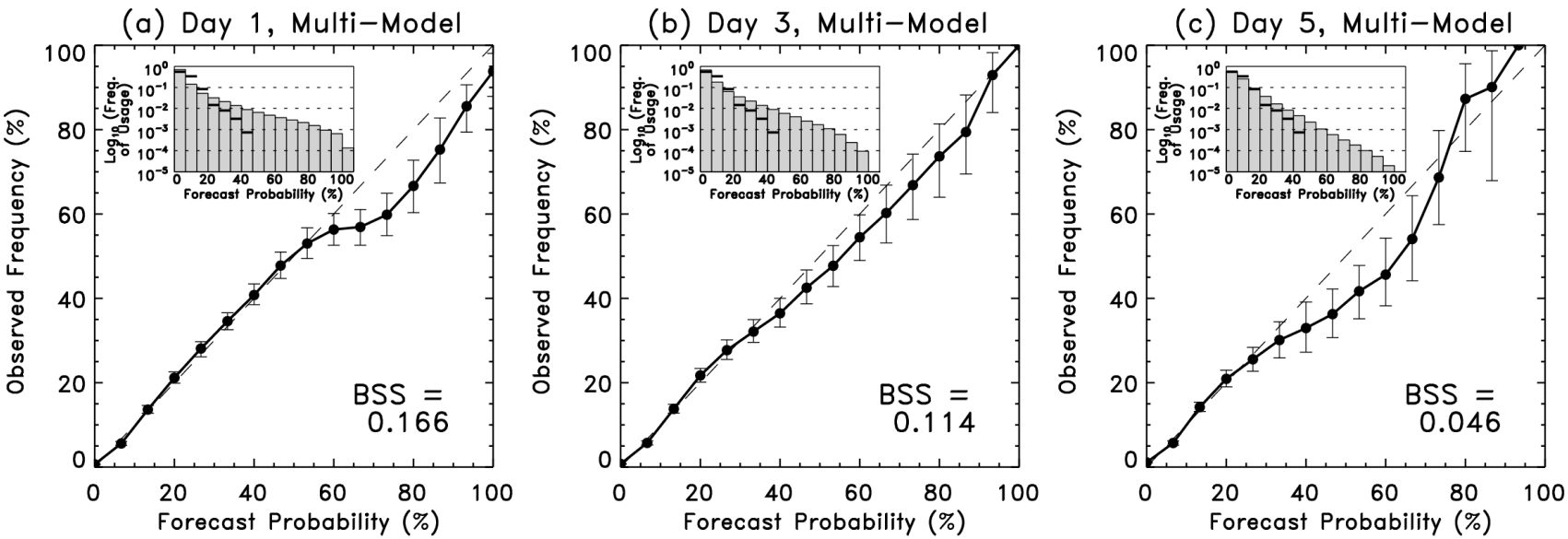
$$\text{BSS} = \frac{\text{“Resolution” - “Reliability”}}{\text{“Uncertainty”}}$$

Uncertainty term always positive, so probability forecasts will exhibit positive skill if resolution term is larger in absolute value than reliability term. Geometrically, this corresponds to points on the attributes diagram being closer to 1:1 perfect reliability line than horizontal no-resolution line (from Wilks text, 2006, chapter 7)

Again, this geometric interpretation of the attributes diagram makes sense only if all samples used to populate the diagram are drawn from the same climatological distribution.If you are mixing samples from locations with different climatologies, this interpretation is no longer correct! (Hamill and Juras, Oct 2006 *QJRMS*)

www.bom.gov.au/bmrc/wefor/staff/eee/verif/ReliabilityDiagram.gif, from Beth Ebert’s verification web page, http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html based on Hsu and Murphy, 1986, *Int’l Journal of Forecasting*

Proposed modifications to standard reliability diagrams



(1) Block-bootstrap techniques (perhaps each forecast day is a block) to provide confidence intervals. See Hamill, *WAF*, April 1999, and Bröcker and Smith, *WAF*, June 2007.

(2) BSS calculated in a way that does not attribute false skill to varying climatology (talk later this morning)

(3) Distribution of climatological forecasts plotted as horizontal bars on the inset histogram. Helps explain why there is small skill for a forecast that appears so reliable (figure from Hamill et al., *MWR*, 2008).

When does reliability ≠ reliability?

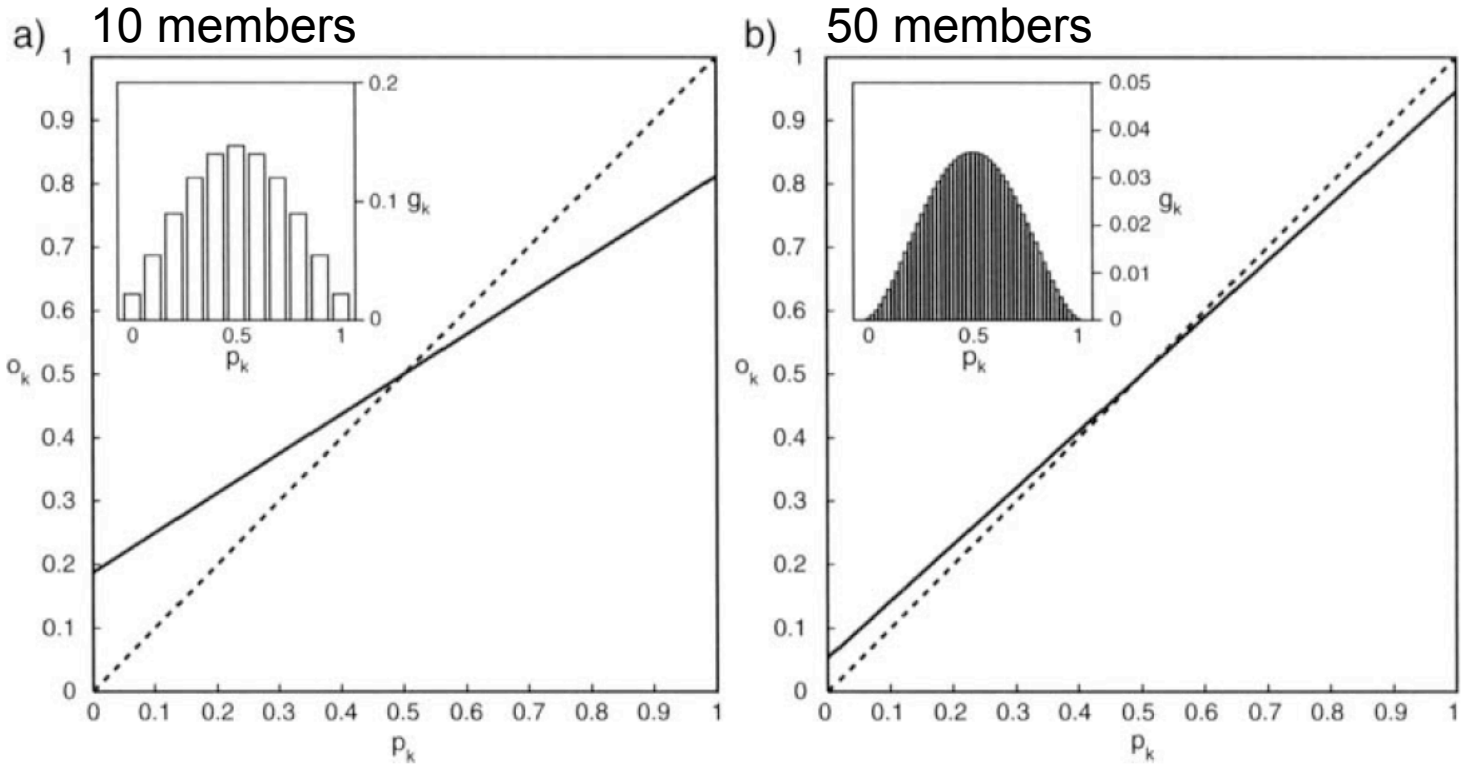
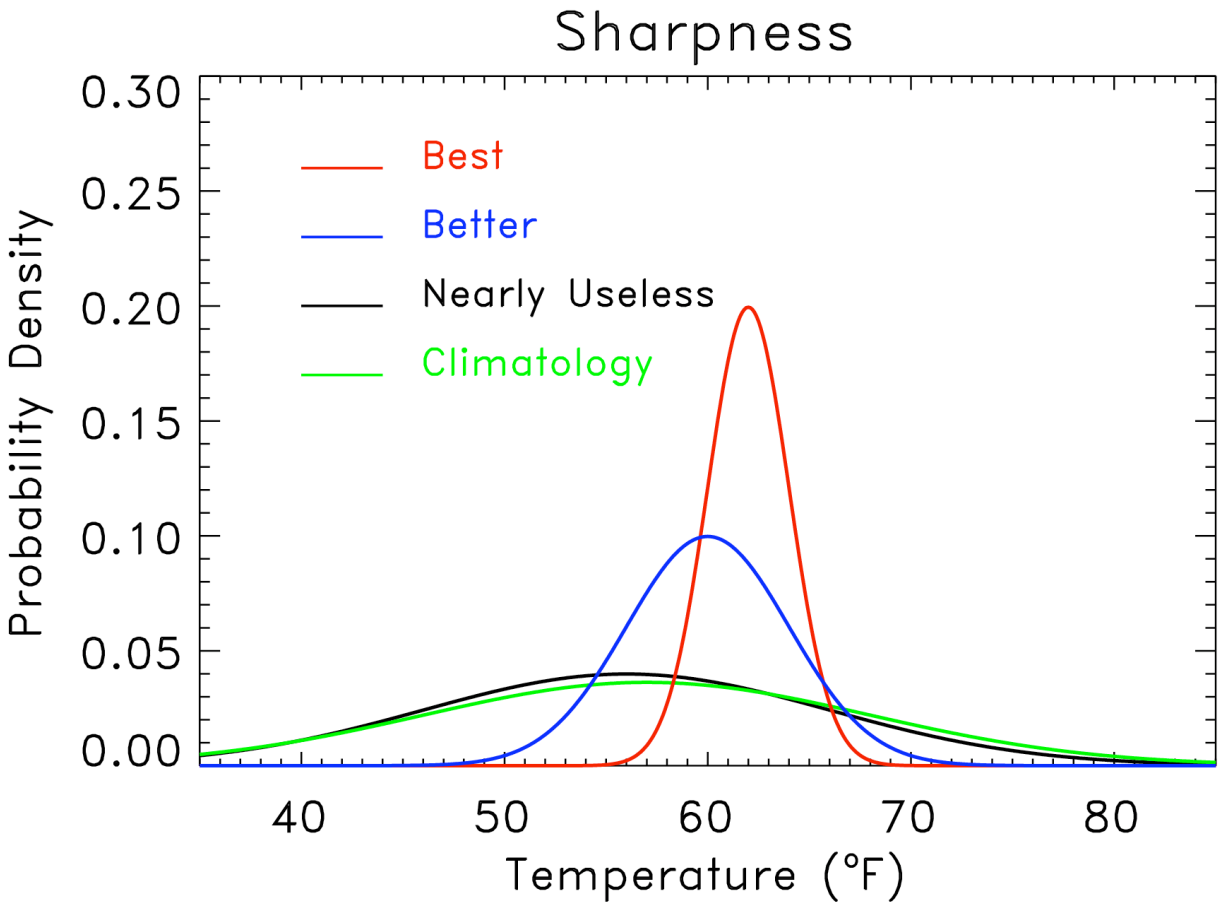


Figure 3. Reliability diagrams for theoretical ensemble forecasts for (a) a 10-member ensemble prediction system (EPS) and (b) a 50-member EPS. Distribution of underlying forecast probabilities is completely reliable and specified by a beta distribution with $r = s = 3$. See text for details and explanation of symbols.

Probabilities directly estimated from a “reliable” ensemble system with only a few members may not produce diagonal reliability curves due to sampling variability.

ref: Richardson 2001 QJRMS

Sharpness also important



"**Sharpness**" measures the specificity of the probabilistic forecast. Given two reliable forecast systems, the one producing the sharper forecasts is preferable.

But: don't want sharp if not reliable. Implies unrealistic confidence.

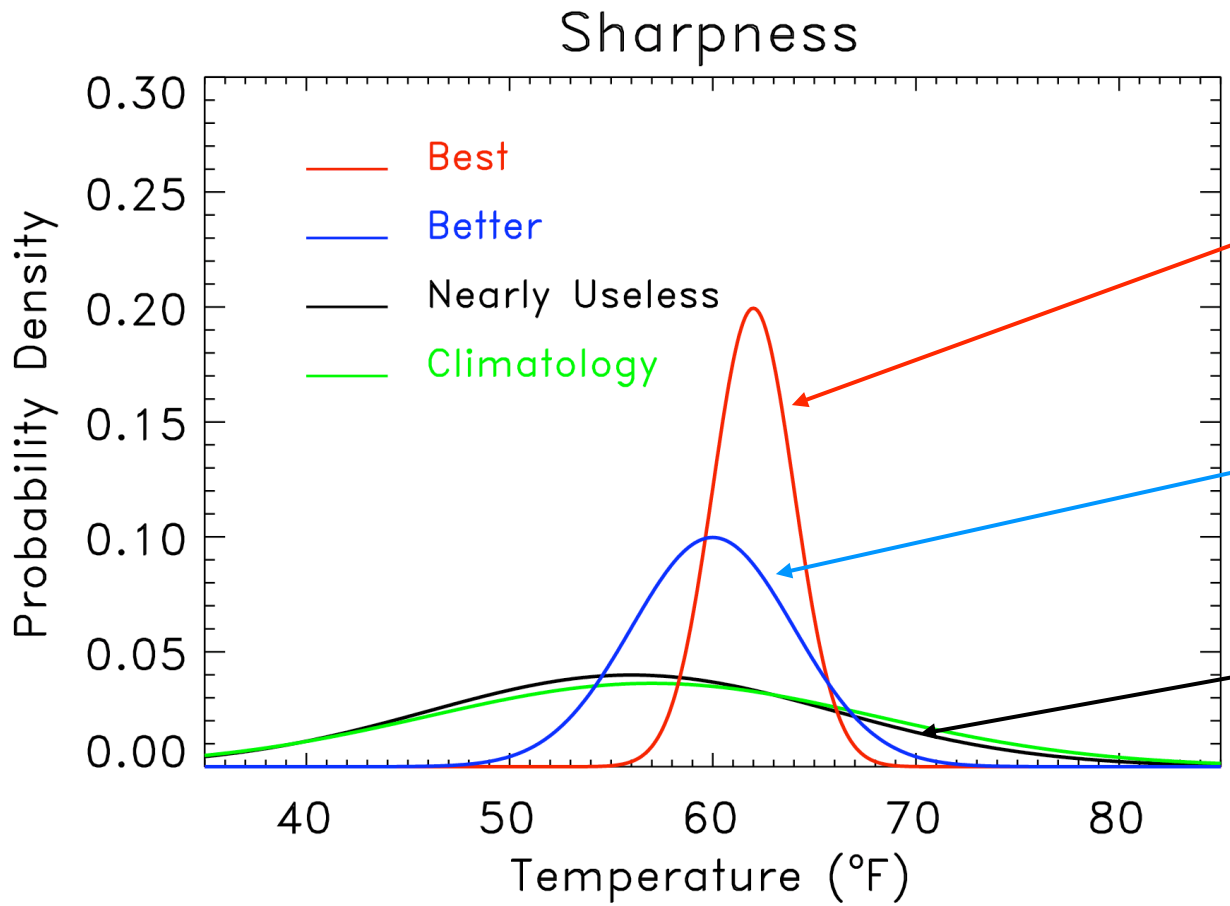
Sharpness ≠ resolution

- Sharpness is *a property of the forecasts alone*; a measure of sharpness in Brier score decomposition would be how populated the extreme N_i*'s are* is.

$$\mathrm{BS} = \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i \left(y_i - \bar{o}_i\right)^2}_{\text{("reliability")}} - \underbrace{\frac{1}{n}\sum_{i=1}^{I} N_i \left(\bar{o}_i - \bar{o}\right)^2}_{\text{("resolution")}} + \underbrace{\bar{o}\left(1 - \bar{o}\right)}_{\text{("uncertainty")}}$$

“Spread-error” relationships are important, too.

Small-spread ensemble forecasts should have less ensemble-mean error than large-spread forecasts, in some sense a conditional reliability dependent upon amount of sharpness.



ensemble-mean error from a sample of this pdf on avg. should be low.

ensemble-mean error should be moderate on avg.

ensemble-mean error should be large on avg.

Why would you expect spread-error relationship?

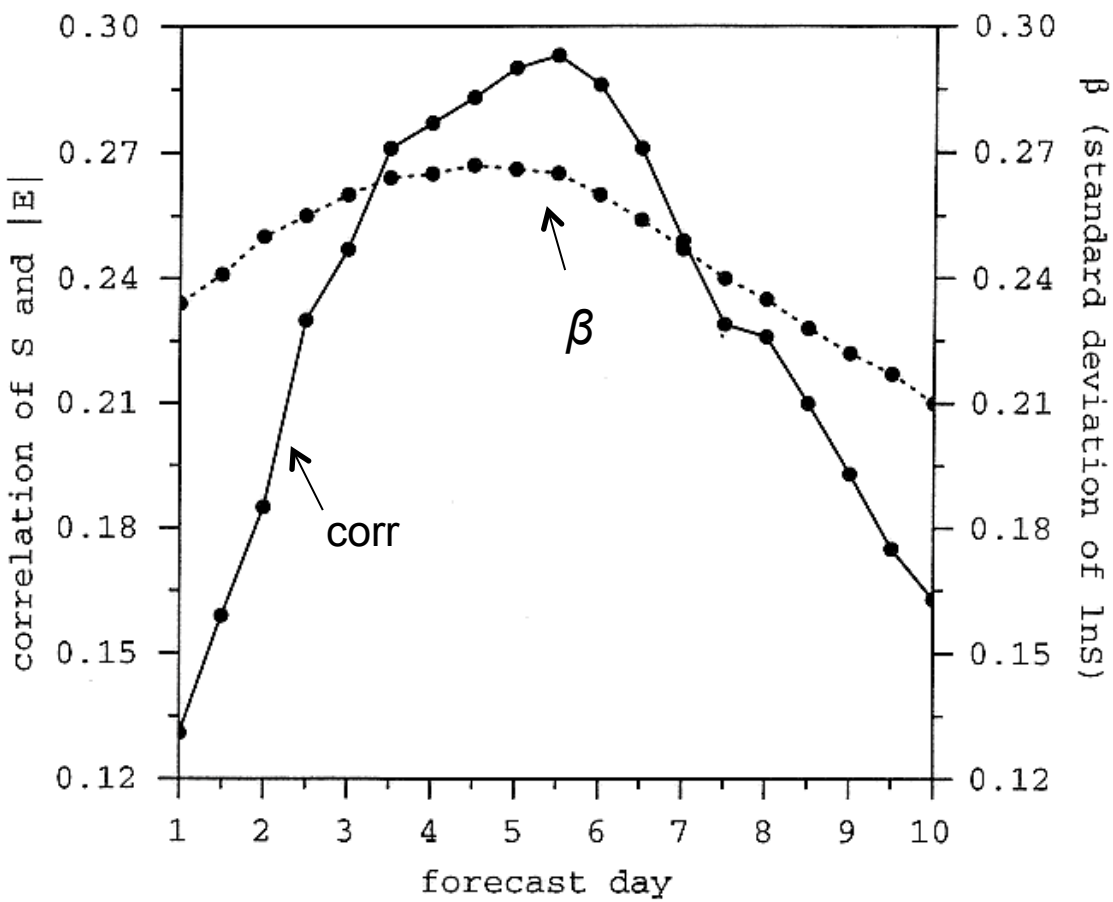
- Ensemble-mean error ought to be the same expected value as ensemble spread.
- If V, x_i are sampled from the same distribution, then

$$E\left(x_i - \overline{x}\right)^2 = E\left(V - \overline{x}\right)^2$$

(spread)^2 (ens. mean error)^2

- Sometimes quantified with a correlation between spread and error.

Spread-error correlations with 1990's NCEP GFS



At a given grid point, spread S is assumed to be a random variable with a lognormal distribution

$$\ln S \sim N\left(\ln S_m, \beta\right)$$

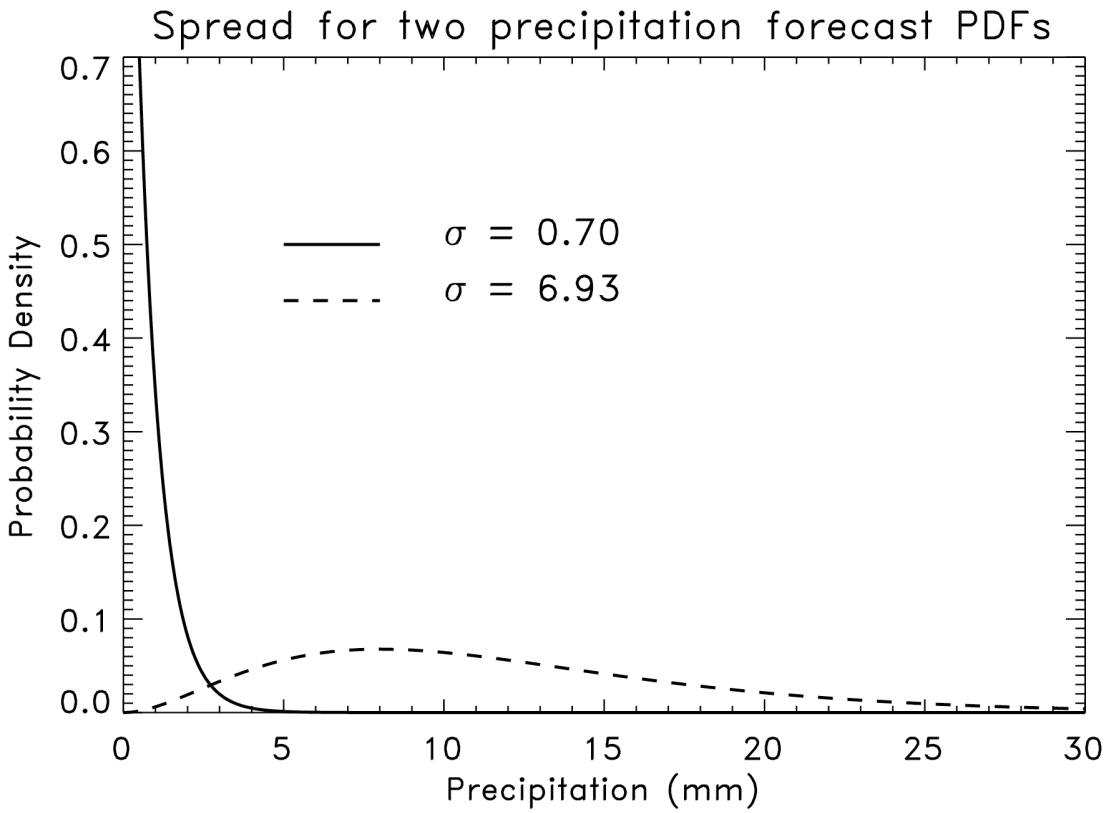
where S_m is the mean spread and β is its standard deviation.

As β increases, there is a wider range of spreads in the sample. One would expect then the possibility for a larger spread-skill correlation.

Lesson: spread-error correlations inherently limited by amount of variation in spread

Ref: Whitaker and Loughe, Dec. 1998 MWR

Spread-error relationships and precipitation forecasts



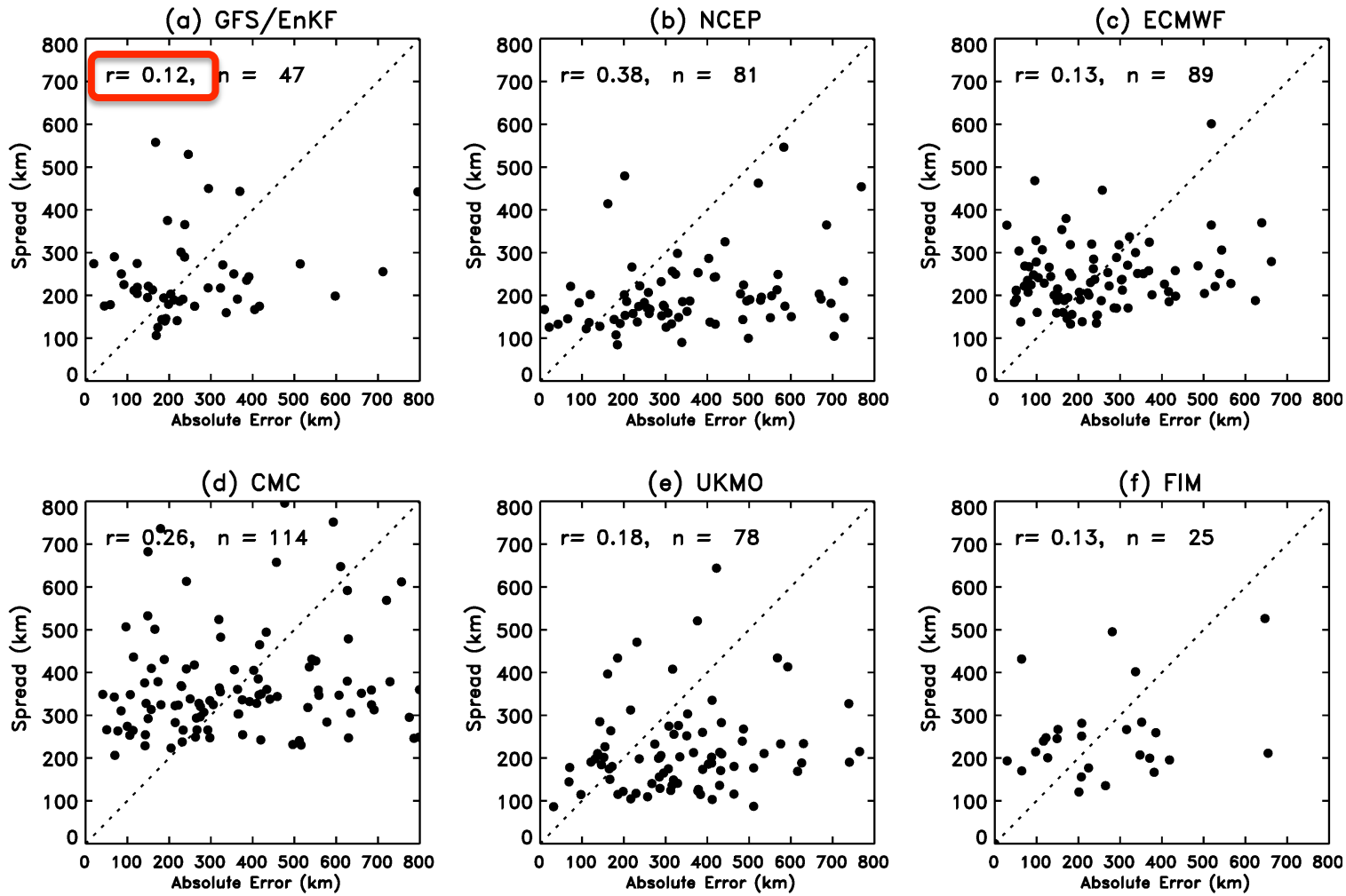
True spread-skill relationships harder to diagnose if forecast PDF is non-normally distributed, as they are typically for precipitation forecasts.

Commonly, **spread is no longer independent of the mean value**; it's larger when the amount is larger.

Hence, you may get an apparent spread-error relationship, but this may reflect variations in the mean forecast rather than real spread-skill.

Ref & possible solution: Hamill and Colucci, MWR, Mar. 1998

Is the spread-error correlation as high as it could be?

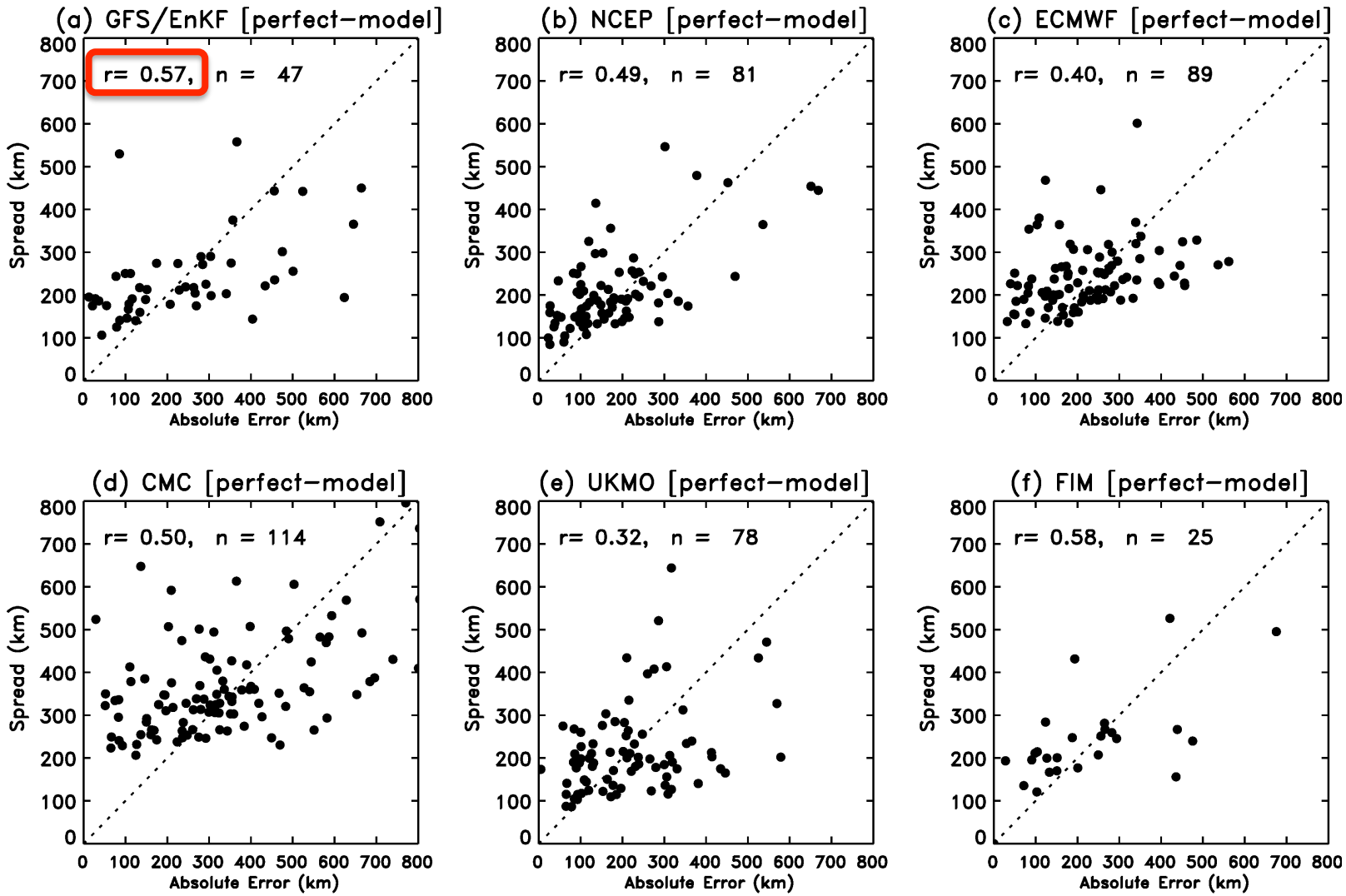


Hurricane track error and spread

40

Ref: Hamill et al., MWR 2010 conditionally accepted.

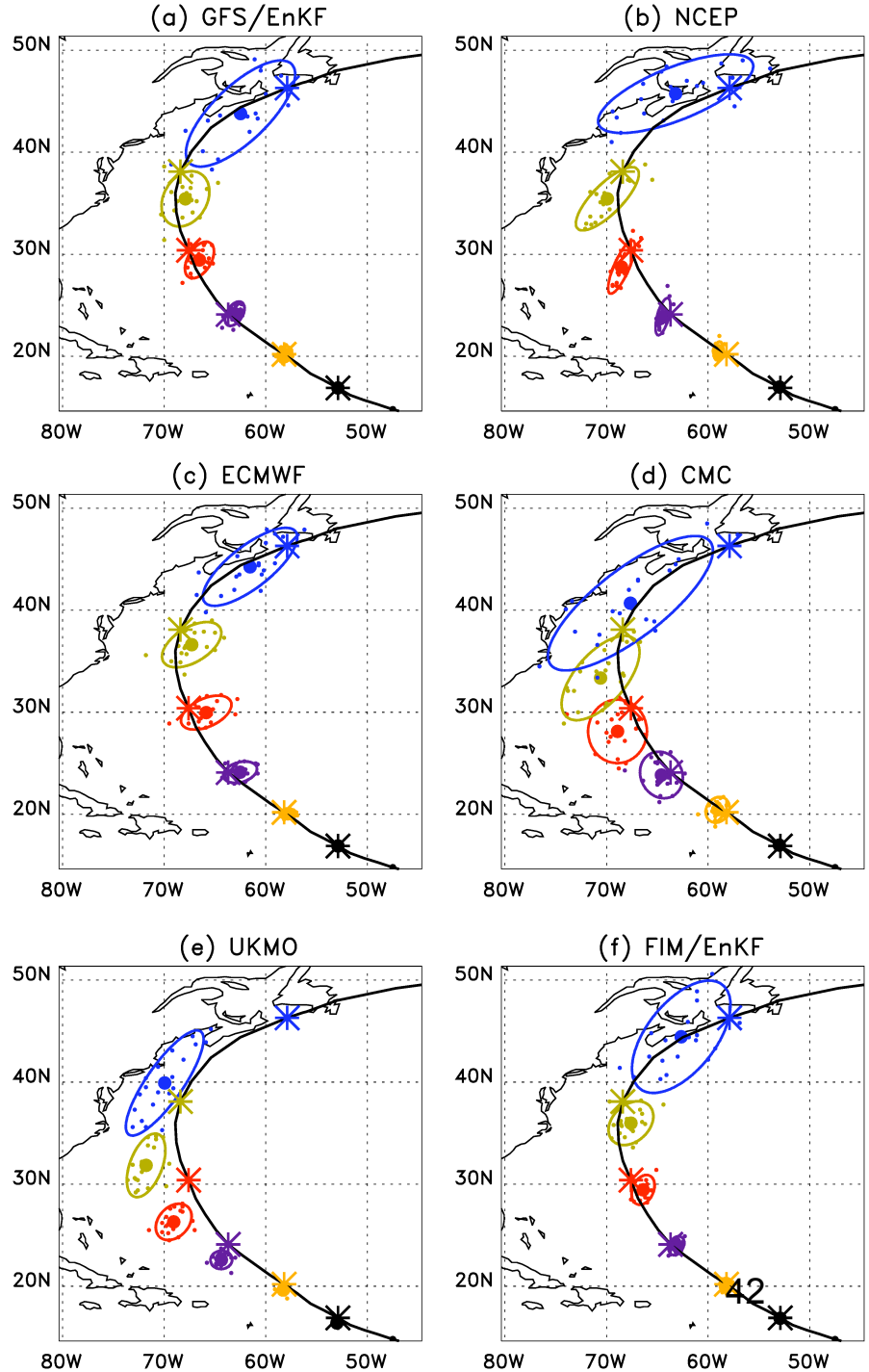
…use one member as synthetic verification to gauge potential spread-error correlation



Spread vs. error in more than one dimension: verification of ellipse eccentricity

Hurricane Bill, initialized 00 UTC 19 August 2009.

Fitted bi-variate normal to each ensemble's forecast track positions. Ellipse encloses 90% of the fitted probability.



Ref: Hamill et al., MWR 2010 conditionally accepted.

Ellipse eccentricity analysis

Question: are errors projections larger along the direction where the ellipse is stretched out?

$$\mathbf{x}'_\lambda = \left(x_{\lambda(1)} - \bar{x}_\lambda,\ \ldots\ ,x_{\lambda(nt)} - \bar{x}_\lambda\right)\Big/\left(nt-1\right)^{1/2}$$

$$\mathbf{x}'_\phi = \left(x_{\phi(1)} - \bar{x}_\phi,\ \ldots\ ,x_{\phi(nt)} - \bar{x}_\phi\right)\Big/\left(nt-1\right)^{1/2}$$

$$\lambda = longitude, \quad \phi = latitude, \quad nt = \#\,tracked$$

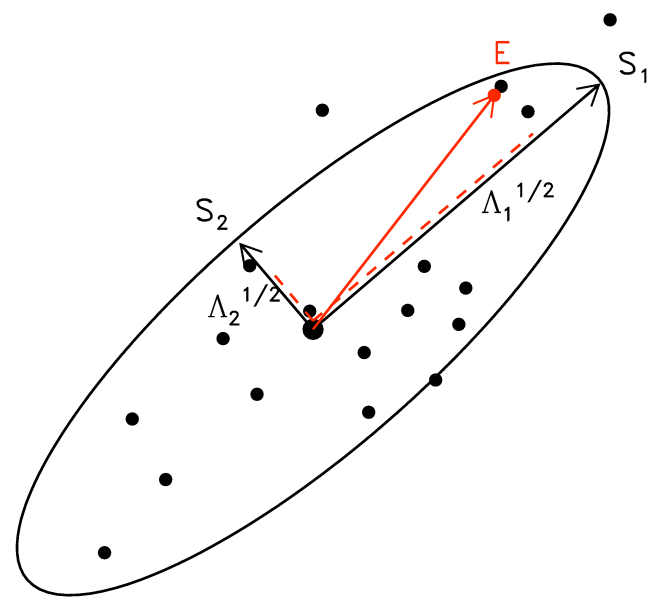
$$\mathbf{X} = \begin{bmatrix} \mathbf{x}'_\lambda \\ \mathbf{x}'_\phi \end{bmatrix}$$

$$\mathbf{F} = \mathbf{X}\mathbf{X}^{\mathrm{T}} = \mathbf{S}\,\Lambda\,\mathbf{S}^{-1} = \mathbf{S}\,\Lambda\,\mathbf{S}^{\mathrm{T}} = \left(\mathbf{S}\,\Lambda^{1/2}\right)\left(\mathbf{S}\,\Lambda^{1/2}\right)^{\mathrm{T}}$$

$\left\langle \left|\mathbf{E}\bullet\mathbf{S}_1\right| \right\rangle$ should be consistent with $\left\langle\left\langle \left|\mathbf{X}_{i\bullet}\bullet\mathbf{S}_1\right| \right\rangle\right\rangle$

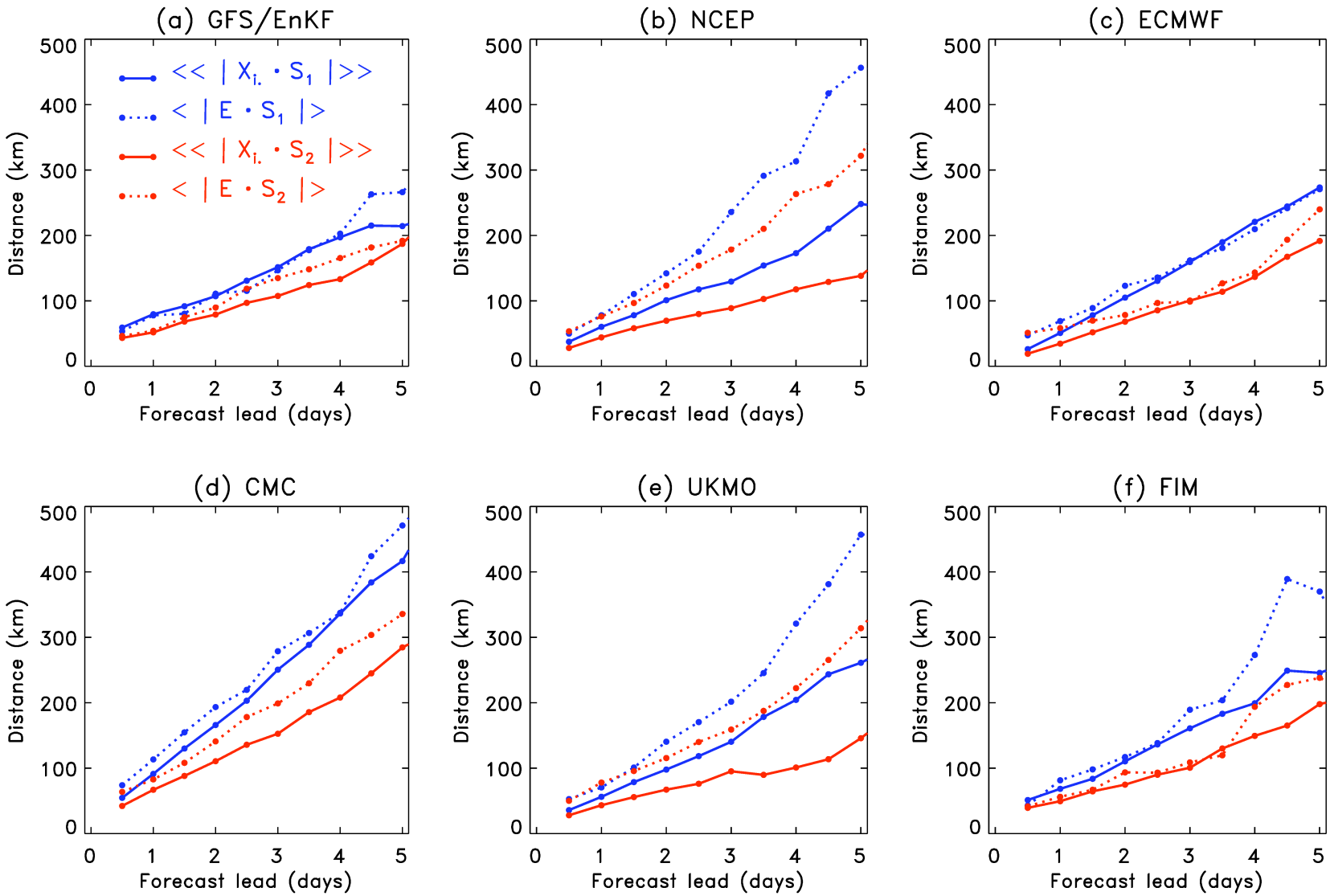
$\left\langle \left|\mathbf{E}\bullet\mathbf{S}_2\right| \right\rangle$ should be consistent with $\left\langle\left\langle \left|\mathbf{X}_{i\bullet}\bullet\mathbf{S}_2\right| \right\rangle\right\rangle$

$\langle\bullet\rangle$ = average over cases; $\langle\langle\bullet\rangle\rangle$ = average over cases, members



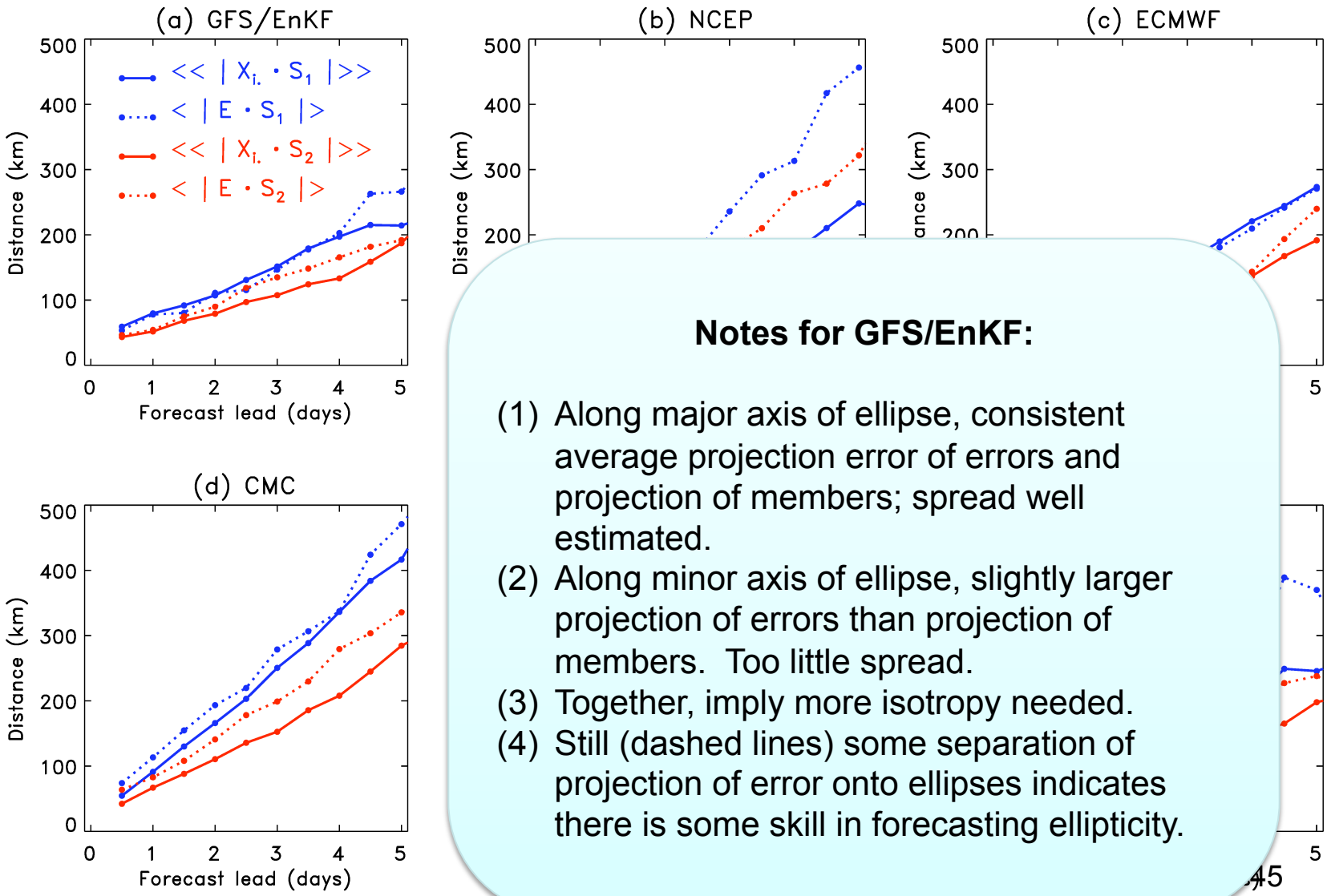
Ellipse eccentricity analysis

(non-homogeneous)



Ellipse eccentricity analysis

(non-homogeneous)



Notes for GFS/EnKF:

(1) Along major axis of ellipse, consistent average projection error of errors and projection of members; spread well estimated.
(2) Along minor axis of ellipse, slightly larger projection of errors than projection of members. Too little spread.
(3) Together, imply more isotropy needed.
(4) Still (dashed lines) some separation of projection of error onto ellipses indicates there is some skill in forecasting ellipticity.

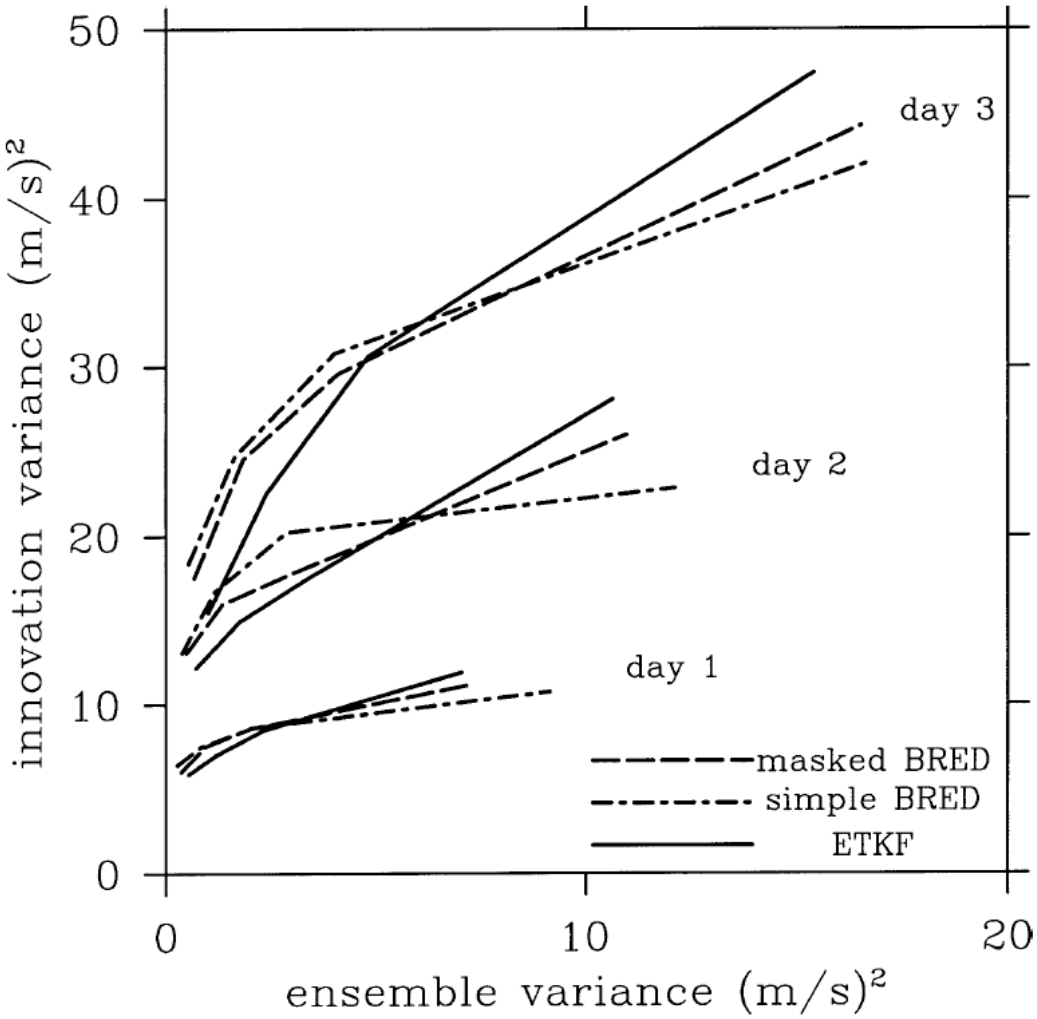


FIG. 8. This figure is plotted by first drawing a scatterplot (not shown) of squared 500-hPa *U* wind innovation vs 500-hPa *U* wind ensemble variance at a particular forecast lead time for each mid-latitude observation location for all forecasts during boreal summer of 2000, dividing the points into four equally populated bins, arranged in order of increasing ensemble variance, and then averaging the squared innovation and ensemble variance in each bin. What is shown is averaged squared innovation vs the averaged ensemble variance from 1- to 3-day forecasts.

Another way of conveying spread-error relationships.

Averaging the individual samples into bins, then plotting a curve.

Ref: Wang and Bishop, JAS, May 2003.

Part 2: other common (and uncommon) ensemble-related evaluation tools

- What do they tell us about the ensemble?
- How are they computed?
- Is there some relation to reliability, sharpness, or other previously discussed attributes?
- What are some issues in their usage?

Isla Gilmour's non-linearity index

how long does a linear regime persist?

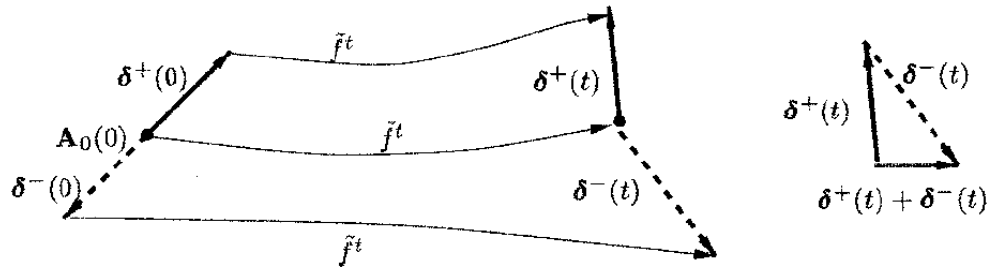


FIG. 2. Defining Θ: equal and opposite perturbations at $t = 0$, $\boldsymbol{\delta}^{\pm}(0)$, evolve so as to be no longer symmetric at time t. The error in assuming linear dynamics, $\|\boldsymbol{\delta}^+(t) + \boldsymbol{\delta}^-(t)\|$, is scaled by the average magnitude of the evolved perturbations to give the relative nonlinearity Θ.

the *relative nonlinearity* of evolution Θ, given by

$$\Theta(\hat{\boldsymbol{\delta}}, \|\boldsymbol{\delta}\|, t) = \frac{\|\boldsymbol{\delta}^+(t) + \boldsymbol{\delta}^-(t)\|}{0.5\{\|\boldsymbol{\delta}^+(t)\| + \|\boldsymbol{\delta}^-(t)\|\}}, \qquad (3)$$

where $\hat{\boldsymbol{\delta}}$ is the unit vector and $\|\cdot\|$ is one of several possible metrics

Ref: Gilmour et al., JAS, Nov. 2001

Power spectra and saturation

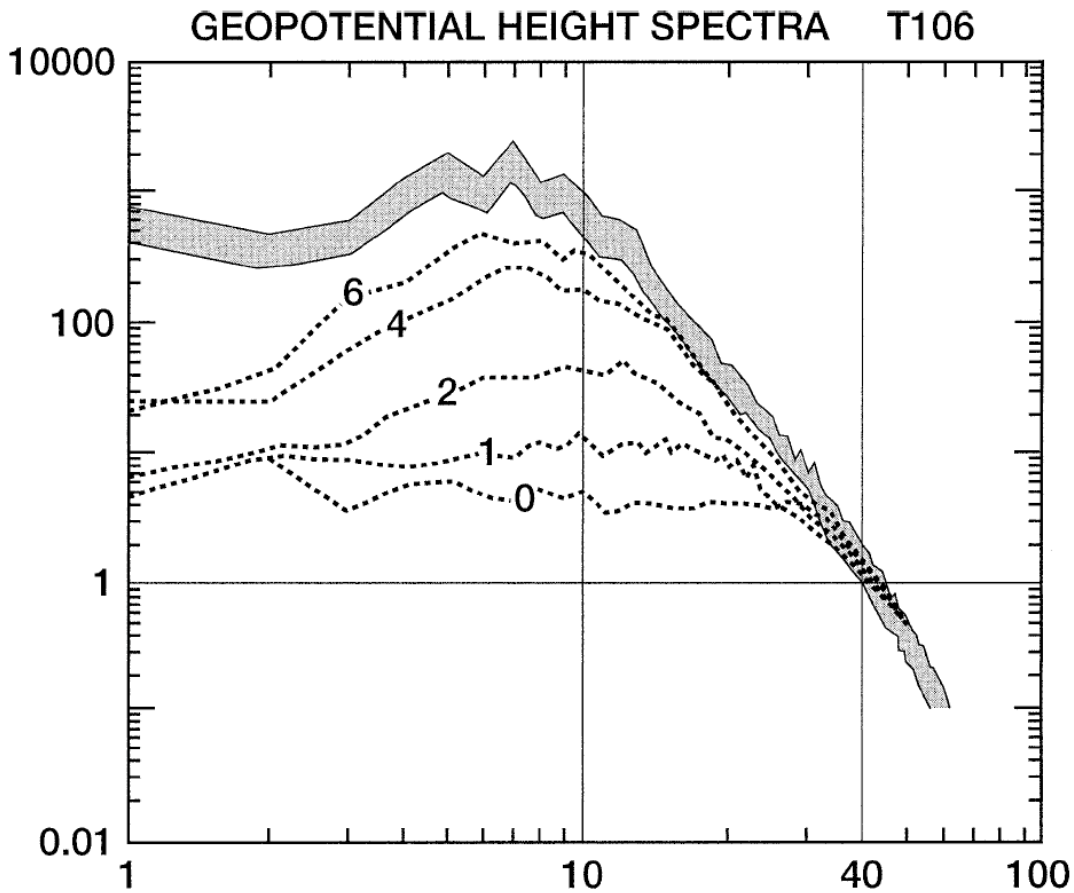


FIG. 8. Same as Fig. 7, but for ensemble dispersion error growth from errors specified in the initial condition. Gray shading indicates σ difference from total 2σ saturation.

The 0–curve represents the power difference between analyses as a function of global wavenumber.

Other numbered curves indicate the power difference at different forecast leads, in days.

Diagnostics like this can be useful in evaluating upscale growth characteristics of ensemble uncertainty.

Ref: Tribbia and Baumhefner, MWR, March 2004

Satterfield and Szunyogh's "Explained Variance"

$$EV = \frac{\left\| \delta \xi^{(\|)} \right\|}{\left\| \xi \right\|} = \frac{\left\| \delta \xi^{(\|)} \right\|}{\left\| \delta \xi^{(\|)} + \delta \xi^{(\perp)} \right\|}$$

How much of the error ξ (difference of truth from ens. mean) lies in the space spanned by the ensemble (||) vs. orthogonal to it ($\perp$)? (calculation is done in a limited-area region).

Would expect that as EV decreases, more U-shaped multi-dimensional rank histograms would be encountered.

Ref: Satterfield and Szunyogh, MWR, March 2010

Example with forecasts from “local ETKF”

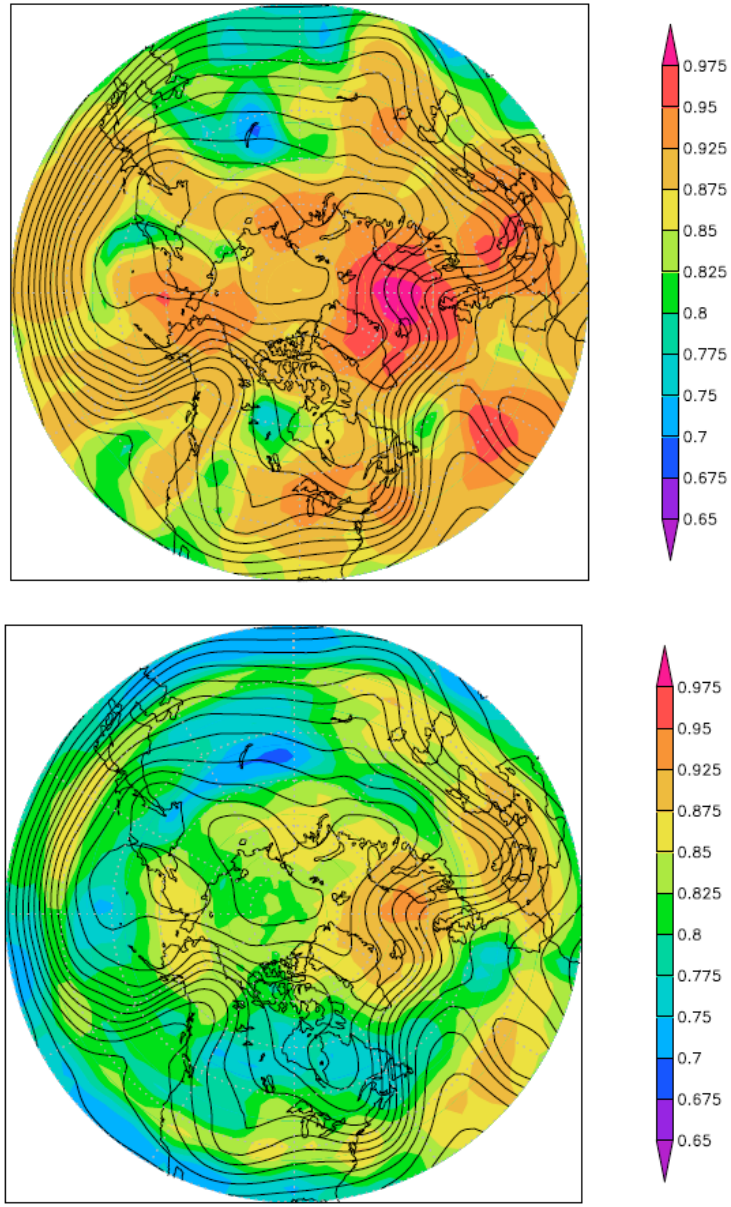


FIG. 4. Explained variance (shades) and geopotential height control (contours) at the 250-hPa level shown for the experiment that assimilates conventional observations for a local region size of 5x5 (top panels) and 10x10 (bottom panels). Results are shown for the 5-day forecast started on 4 Feb 2004.

top: 5x5 degree boxes

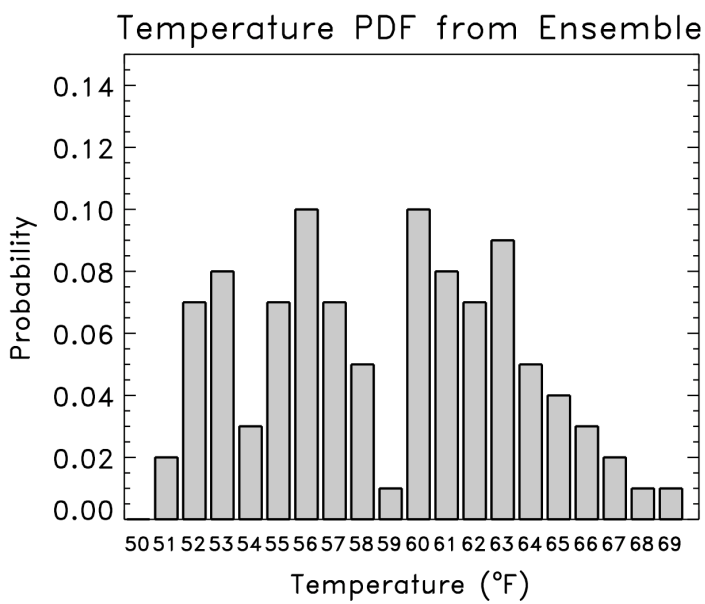
bottom: 10x10 boxes.

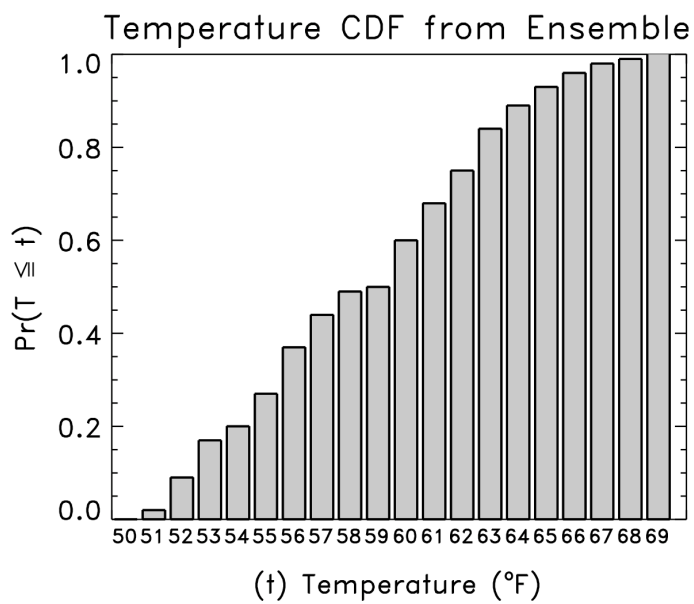
the larger the box, the smaller the EV, but patterns similar.

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.

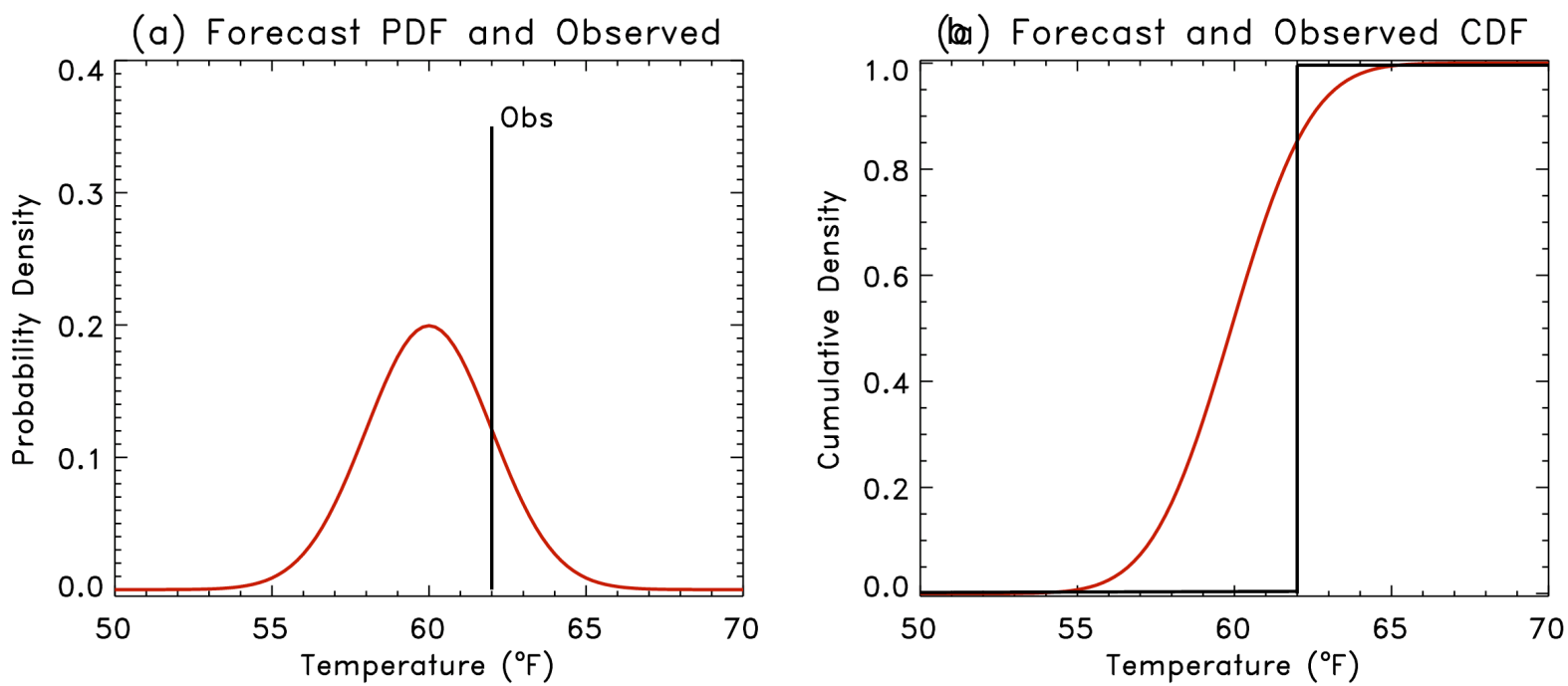




Continuous Ranked Probability Score

- 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).

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

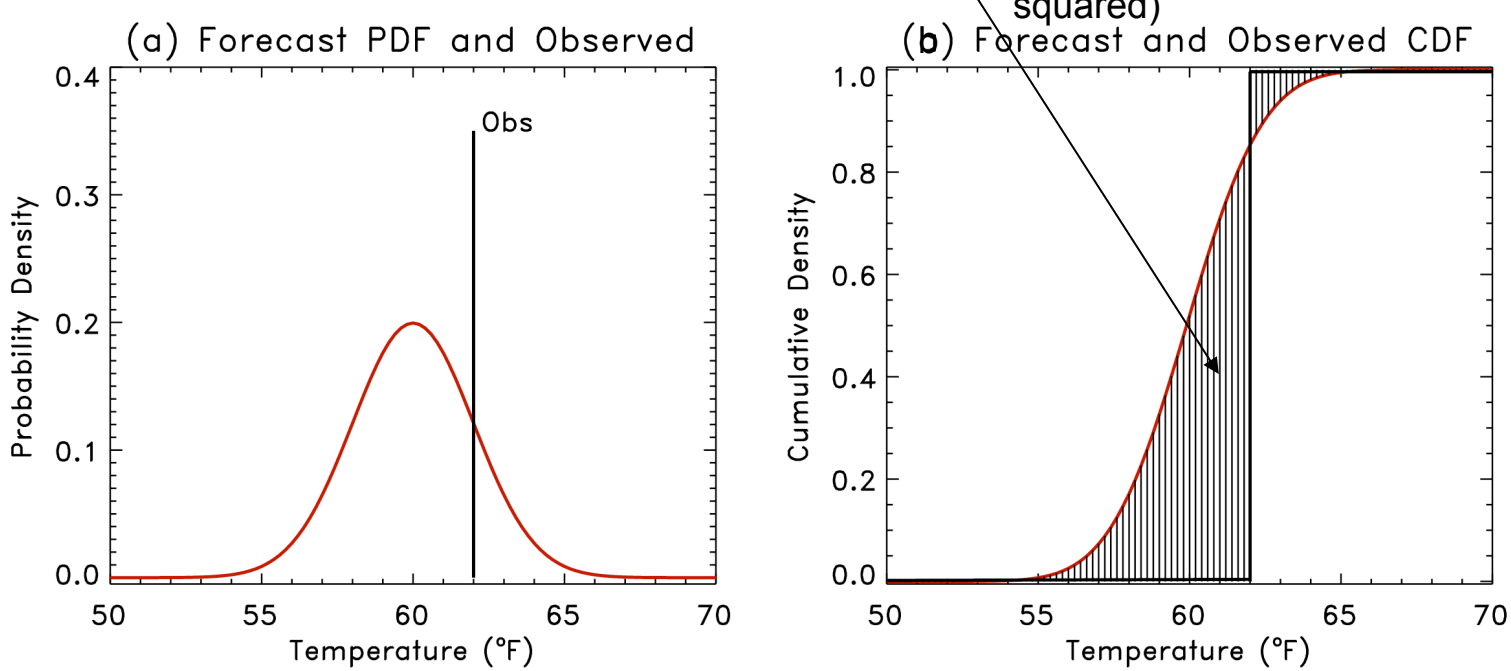


Continuous Ranked Probability Score

- 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

Continuous Ranked Probability **Skill** Score (CRPSS)

Like the Brier score, it's common to convert this to a skill score by normalizing by the skill of a reference forecast, perhaps climatology.

$$CRPSS = \frac{\overline{CRPS}(forecast) - \overline{CRPS}(climo)}{\overline{CRPS}(perfect) - \overline{CRPS}(climo)}$$

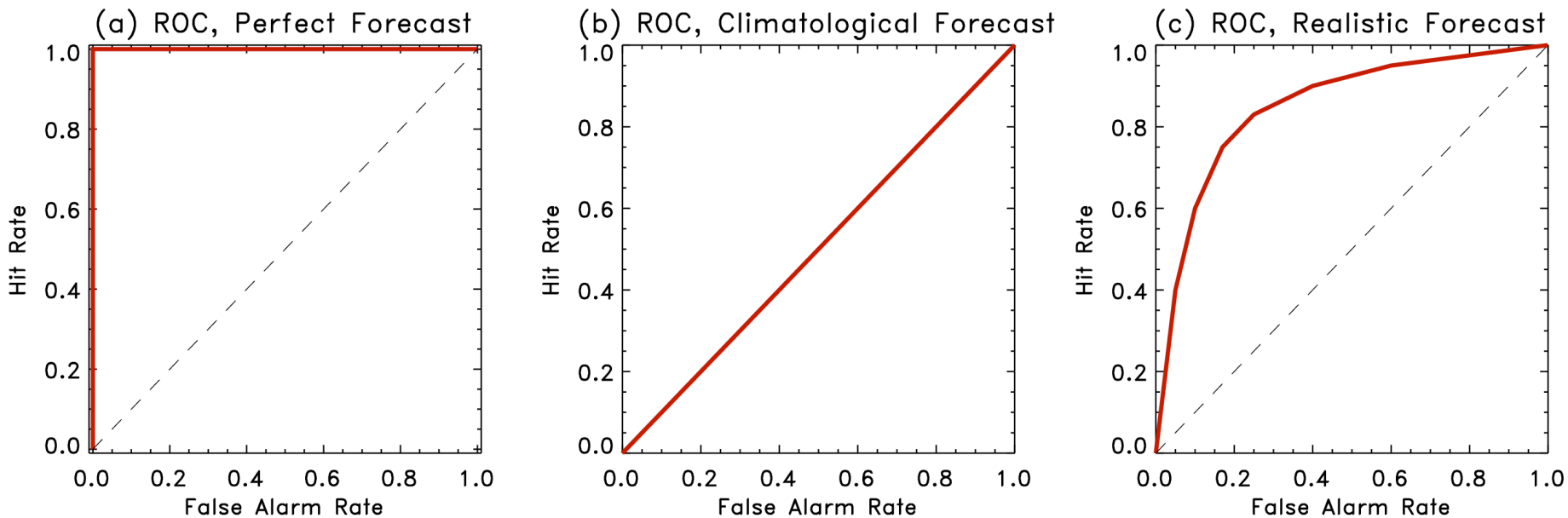
Danger: can over-estimate forecast skill. See later presentation on this.

Decomposition of CRPS

- Like Brier score, there is a decomposition of CRPS into reliability, resolution, uncertainty.
- Like Brier score, interpretation of this decomposition only makes sense if all samples are draws from a distribution with the same climatology.

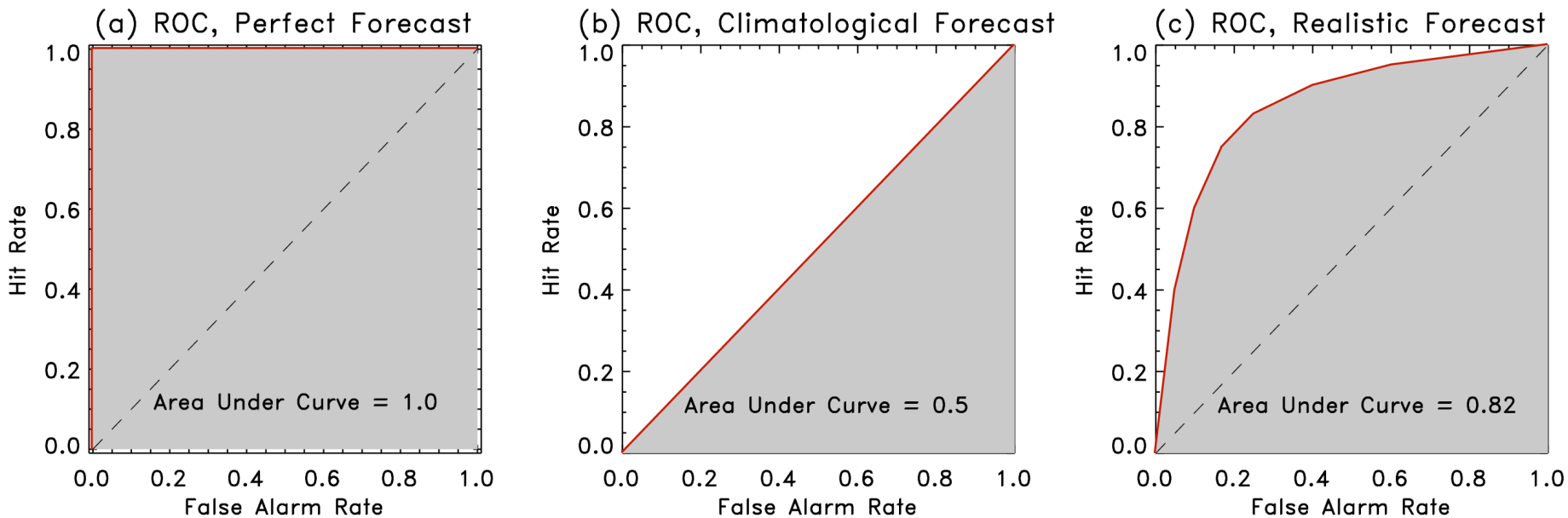
Ref: Hersbach, WAF, Oct 2000, + Hamill and Juras, QJ, Oct 2006

Relative Operating Characteristic (ROC)



Hit Rate = H / (H+M) FAR = F / (F+C)

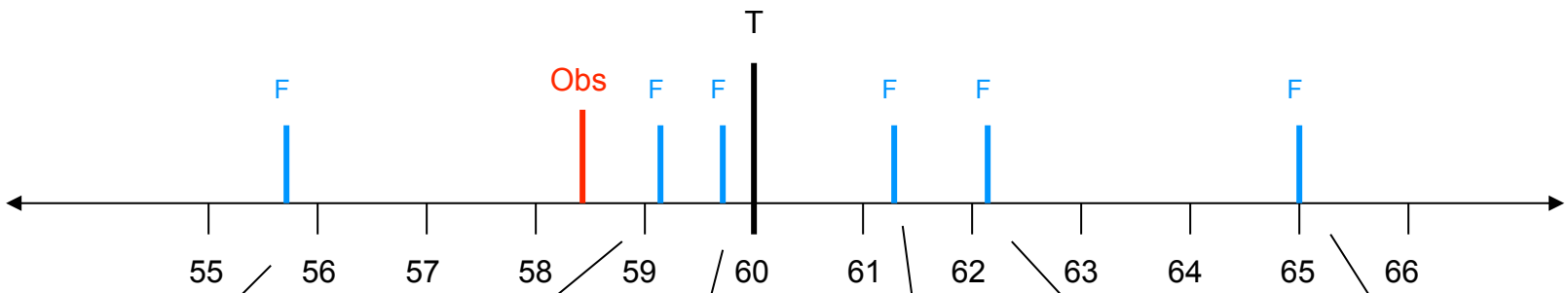
Relative Operating Characteristic (ROC)



$$ROCSS = \frac{AUC_f - AUC_{clim}}{AUC_{perf} - AUC_{clim}} = \frac{AUC_f - 0.5}{1.0 - 0.5} = 2AUC_f - 1$$

Method of calculation of ROC: parts 1 and 2

(1) Build contingency tables for each sorted ensemble member



Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	0
N	0	1

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	0
N	0	1

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	0
N	0	1

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	1
N	0	0

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	1
N	0	0

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	0	1
N	0	0

(2) Repeat the process for other locations, dates, building up contingency tables for sorted members.

Method of calculation of ROC: part 3

(3) Get hit rate and false alarm rate for each from contingency table for each sorted ensemble member.

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	H	F
N	M	C

$$HR = H / (H+M) \qquad FAR = F / (F+C)$$

Sorted Member 1

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	1106	3
N	5651	73270

HR = 0.163
FAR = 0.000

Sorted Member 2

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	3097	176
N	3630	73097

HR = 0.504
FAR = 0.002

Sorted Member 3

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	4020	561
N	2707	72712

HR = 0.597
FAR = 0.007

Sorted Member 4

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	4692	1270
N	2035	72003

HR = 0.697
FAR = 0.017

Sorted Member 5

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	5297	2655
N	1430	70618

HR = 0.787
FAR = 0.036

Sorted Member 6

Obs ≥ T?

Fcst ≥ T?	Y	N
Y	6603	44895
N	124	28378

HR = 0.981
FAR = 0.612

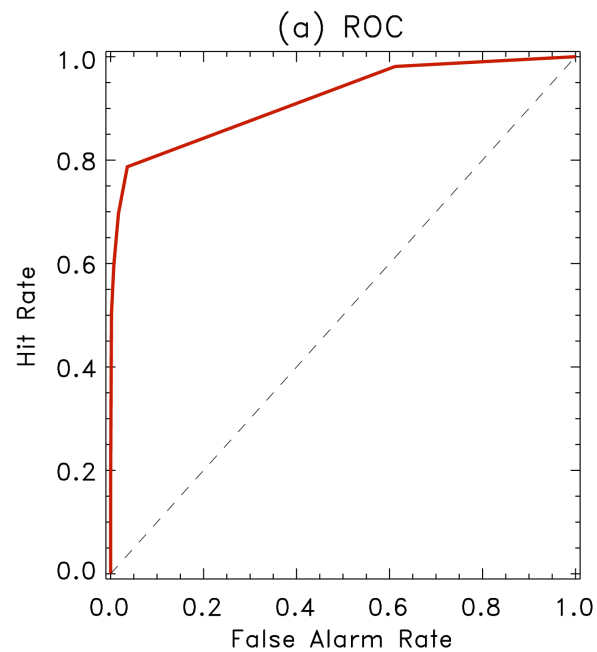
Method of calculation of ROC: part 3

HR = 0.163	HR = 0.504	HR = 0.597	HR = 0.697	HR = 0.787	HR = 0.981
FAR = 0.000	FAR = 0.002	FAR = 0.007	FAR = 0.017	FAR = 0.036	FAR = 0.612

HR = [0.000, 0.163, 0.504, 0.597, 0.697, 0.787, 0.981, 1.000]

FAR = [0.000, 0.000, 0.002, 0.007, 0.017, 0.036, 0.612, 1.000]

(4) Plot hit rate vs. false alarm rate



Again, can overestimate forecast skill.

ROC connection with hypothesis testing

Observed ≥ T?

Fcst ≥ T? (our test statistic)	YES	NO
YES	H (HIT)	(F) FALSE ALARM
NO	(M) MISS	(C) CORRECT NO

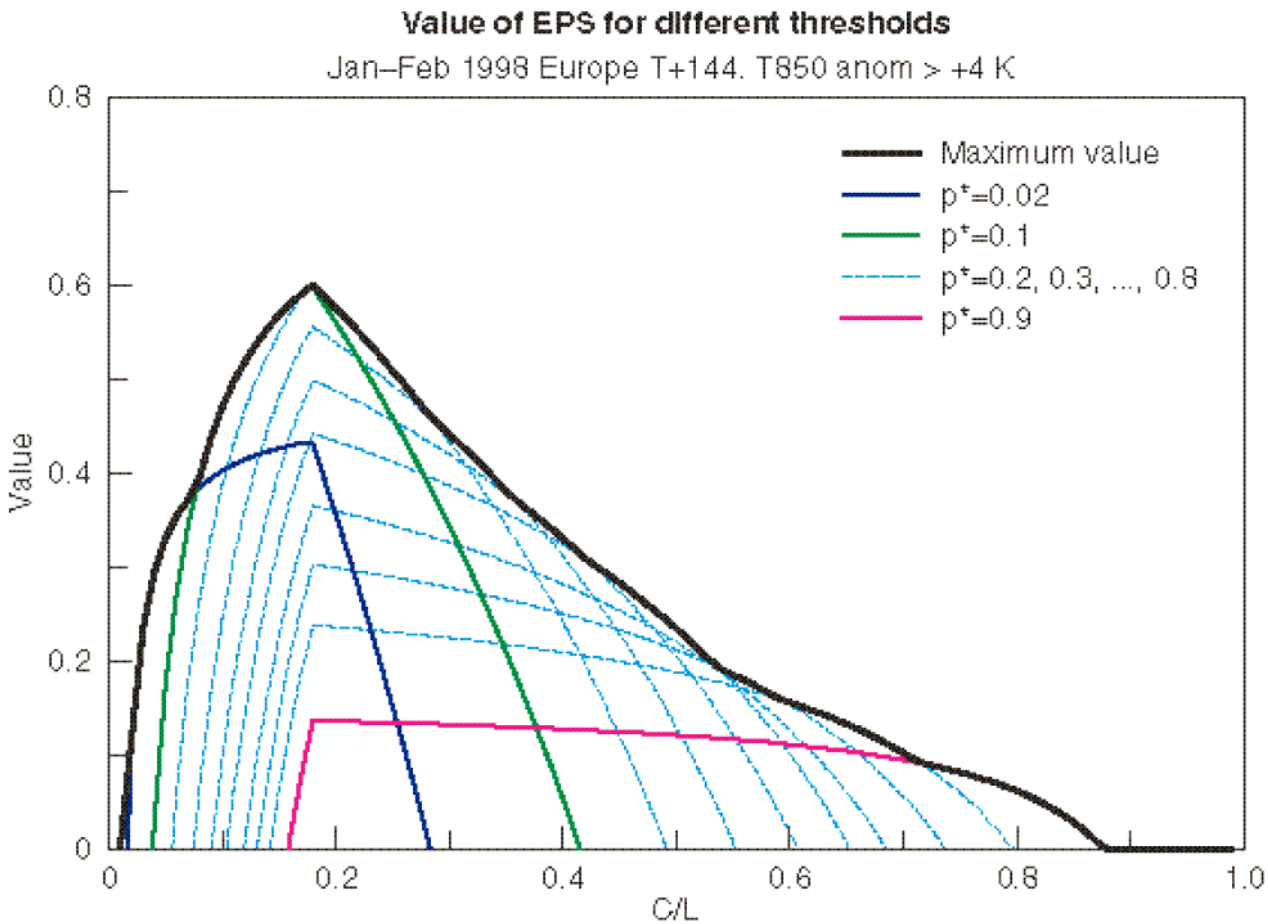
Point on ROC curve provides evaluation of Type I (inappropriate rejection of null hypothesis, which here is Obs < T) vs. 1. - Type II statistical errors (inappropriate acceptance of alternate hypothesis, Obs ≥ T). ROC curve provides tradeoff as different ensemble data is used as test statistic.

FAR = F / F+C = P(Type I error)

HR = H / H+M = 1 - P(Type II error)

Economic value diagrams

Motivated by search for a metric that relates ensemble forecast performance to things that customers will actually care about.



These diagrams tell you the potential economic value of your ensemble forecast system applied to a particular forecast aspect. Perfect forecast has value of 1.0, climatology has value of 1.0. Value differs with user's cost/loss ratio.

Economic value: calculation method

Contingency table indicating the costs and losses accrued by the use of weather forecasts, depending on forecast and observed events.

		Forecast/action		
		Yes	**No**	
Observation	**Yes**	Hit (h) Mitigated loss ($C + L_u$)	Miss (m) Loss ($L = L_p + L_u$)	$h + m = \bar{o}$
	No	False Alarm (f) Cost (C)	Correct rejection (c) No cost (N)	$f + c = 1 - \bar{o}$

Assumes decision maker alters actions based on weather forecast info.

C = Cost of protection
$L = L_p + L_u$ = total cost of a loss, where …
L_p = Loss that can be protected against
L_u = Loss that can't be protected against.
N = No cost

Economic value, continued

. Contingency table indicating the costs and losses accrued by the use of weather forecasts, depending on forecast and observed events.

		Forecast/action: Yes	Forecast/action: No	
Observation	Yes	Hit (h) Mitigated loss ($C + L_u$)	Miss (m) Loss ($L = L_p + L_u$)	$h + m = \bar{o}$
	No	False Alarm (f) Cost (C)	Correct rejection (c) No cost (N)	$f + c = 1 - \bar{o}$

Suppose we have the contingency table of forecast outcomes, [*h, m, f, c*].

Then we can calculate the expected value of the expenses from a forecast, from climatology, from a perfect forecast.

$$E_{forecast} = fC + h\left(C + L_u\right) + m\left(L_p + L_u\right)$$

$$E_{climate} = Min\left[\bar{o}\left(L_p + L_u\right),\ C + \bar{o}L_u\right] = \bar{o}L_u + Min\left[\bar{o}L_p,\ C\right]$$

$$E_{perfect} = \bar{o}\left(C + L_u\right)$$

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}} = \frac{Min\left[\bar{o}L_p,\ C\right] - (h+f)C - mL_p}{Min\left[\bar{o}L_p,\ C\right] - \bar{o}C}$$

Note that value will vary with C, L_p, L_u;

Different users with different protection costs may experience a different value from the forecast system.

Ref: Zhu et al. 2002 BAMS

From ROC to economic value

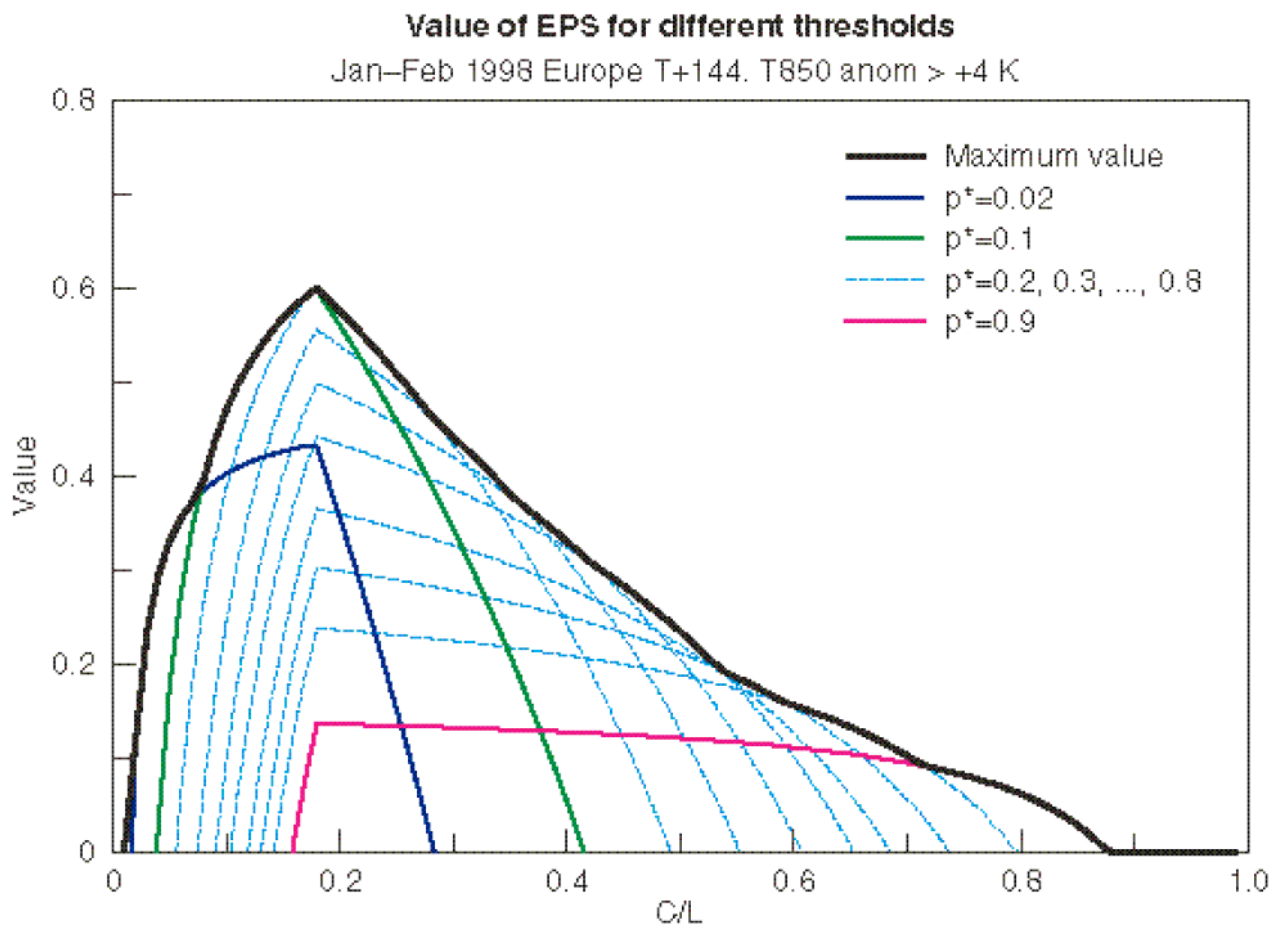
$$HR = \frac{h}{\bar{o}} \qquad FAR = \frac{f}{1-\bar{o}} \qquad m = \bar{o} - HR\,\bar{o}$$

$$V = \frac{Min\left[\bar{o}, C/L_p\right] - (h+f)C/L_p - m}{Min\left[\bar{o}, C/L_p\right] - or}$$

$$= \frac{Min\left[\bar{o}, C/L_p\right] - \left(C/L_p\right)FAR\left(1-\bar{o}\right) + HR\,\bar{o}\left(1 - C/L_p\right) - \bar{o}}{Min\left[\bar{o}, C/L_p\right] - \bar{o}r}$$

Value is now seen to be related to FAR and HR, the components of the ROC curve. A (HR, FAR) point on the ROC curve will thus map to a value curve (as a function of C/L)

Ref: ibid.



The red curve is from the ROC data for the member defining the 90th percentile of the ensemble distribution. Green curve is for the 10th percentile. Overall economic value is the maximum (use whatever member for decision threshold that provides the best economic value).

Conclusions

- Many ensemble verification techniques out there.
- A good principle is to thought-test every verification technique you seek to use; is there some way it could be misleading you about the characteristics of the forecast (commonly, the answer will be yes).

Useful references

- **Good overall references** for forecast verification:
  - (1): Wilks, D.S., 2006: *Statistical Methods in the Atmospheric Sciences (2nd Ed).* Academic Press, 627 pp.
  - (2) Beth Ebert's forecast verification web page, http://tinyurl.com/y97c74
- **Rank histograms**: Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev*., **129**, 550-560.
- **Spread-skill relationships**: Whitaker, J.S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev*., **126**, 3292-3302.
- **Brier score, continuous ranked probability score, reliability diagrams**: Wilks text again.
- **Relative operating characteristic**: Harvey, L. O., Jr, and others, 1992: The application of signal detection theory to weather forecasting behavior. *Mon. Wea. Rev*., **120**, 863-883.
- **Economic value diagrams**:
  - (1)Richardson, D. S., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Royal Meteor. Soc*., **126**, 649-667.
  - (2) Zhu, Y, and others, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc*., **83**, 73-83.
- **Overforecasting skill**: Hamill, T. M., and J. Juras, 2006: Measuring forecast skill: is it real skill or is it the varying climatology? *Quart. J. Royal Meteor. Soc.*, Jan 2007 issue. http://tinyurl.com/kxtct