Measuring forecast skill:
is it *real* skill or
is it the varying climatology?

Tom Hamill

NOAA Earth System Research Lab, Boulder, Colorado

tom.hamill@noaa.gov; www.cdc.noaa.gov/people/tom.hamill

Josip Juras

University of Zagreb, Croatia

Hypothesis

- If climatological event probability varies among samples, then many verification metrics will credit a forecast with extra skill it doesn’t deserve - the extra skill comes from the variations in the climatology.

Example: Brier Skill Score

Brier Score: Mean-squared error of probabilistic forecasts.

$$\overline{BS}^{f} = \frac{1}{n}\sum_{k=1}^{n}\left(p_k^f - o_k\right)^2, \quad o_k = \begin{cases} 1.0 & if\ kth\ observation \geq threshold \\ 0.0 & if\ kth\ observation < threshold \end{cases}$$

Brier Skill Score: Skill relative to some reference, like climatology.
1.0 = perfect forecast, 0.0 = skill of reference.

$$BSS = \frac{\overline{BS}^{f} - \overline{BS}^{ref}}{\overline{BS}^{perfect} - \overline{BS}^{ref}} = \frac{\overline{BS}^{f} - \overline{BS}^{ref}}{0.0 - \overline{BS}^{ref}} = 1.0 - \frac{\overline{BS}^{f}}{\overline{BS}^{ref}}$$

Overestimating skill: example

5-mm threshold

Location A: P^f = 0.05, P^{clim} = 0.05, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.05-0)^2} = 0.0$$

Location B: P^f = 0.05, P^{clim} = 0.25, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.25-0)^2} = 0.96$$

Locations A and B:

$$BSS = 1.0 - \frac{\overline{BS}^f}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2 + (.05-0)^2}{(.25-0)^2 + (.05-0)^2} = 0.923$$

Overestimating skill: example

5-mm threshold

Location A: $P^f = 0.05$, $P^{clim} = 0.05$, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^{f}}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.05-0)^2} = 0.0$$

Location B: $P^f = 0.05$, $P^{clim} = 0.25$, Obs = 0

$$BSS = 1.0 - \frac{\overline{BS}^{f}}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2}{(.25-0)^2} = 0.96$$

Locations A and B:

$$BSS = 1.0 - \frac{\overline{BS}^{f}}{\overline{BS}^{clim}} = 1.0 - \frac{(.05-0)^2 + (.05-0)^2}{(.25-0)^2 + (.05-0)^2} = 0.923$$

why not 0.48?

for more detail, see Hamill and Juras, QJRMS, Oct 2006 (c)

Another example of unexpected skill: two islands, zero meteorologists

Imagine a planet with a global ocean and two isolated islands. Weather forecasting other than climatology for each island is impossible.

Island 1: Forecast, observed uncorrelated, ~ N (+α, 1)

Island 2: Forecast, observed uncorrelated, ~ N (–α, 1)

$$0 \leq \alpha \leq 5$$

Event: Observed > 0

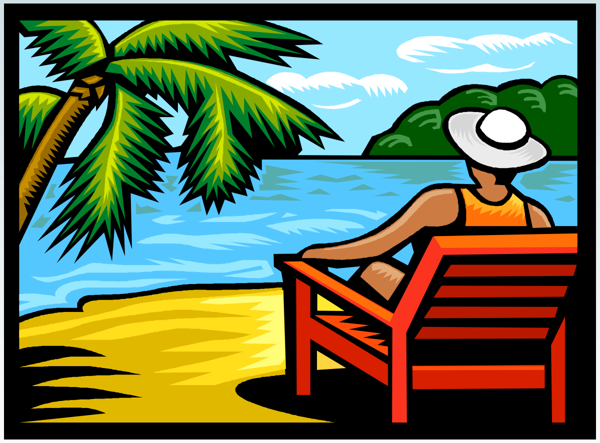
Forecasts: random ensemble draws from climatology

Two islands

As α increases…

Island 2

Island 1



But still, each island’s forecast is no better than
a random draw from its climatology. Expect no skill.

Consider three metrics…

(1) Brier Skill Score

(2) Relative Operating Characteristic

(3) Equitable Threat Score

(each will show this tendency to have scores vary depending on how they’re calculated)

Relative Operating Characteristic: standard method of calculation

Populate 2x2 contingency tables, separate one for each sorted ensemble member. The contingency table for the *i*th sorted ensemble member is

		Event forecast by *i*th member?	
		YES	NO
Event Observed?	YES	a_i	b_i
	NO	c_i	d_i

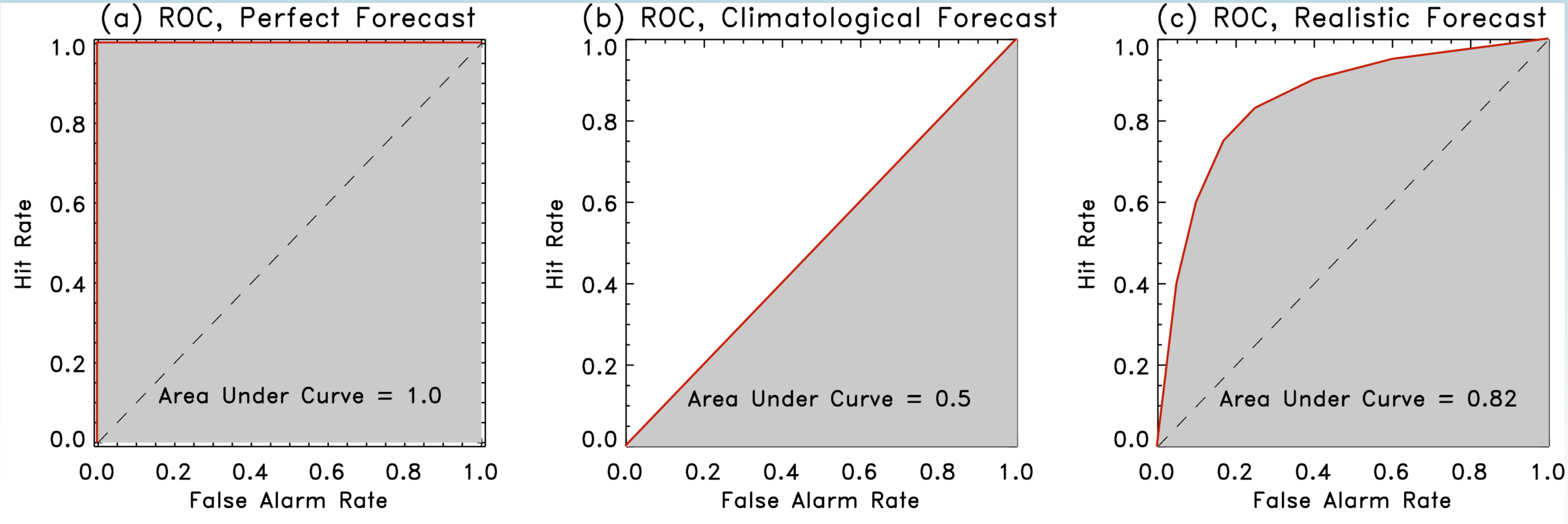
$$(a_i + b_i + c_i + d_i = 1)$$

$$HR_i = \frac{a_i}{a_i + b_i}$$ (hit rate)

$$FAR_i = \frac{c_i}{c_i + d_i}$$ (false alarm rate)

ROC is a plot of hit rate (y) vs. false alarm rate (x). Commonly summarized by “area under curve” (AUC), 1.0 for perfect forecast, 0.5 for climatology.

Relative Operating Characteristic (ROC) skill score



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

Equitable Threat Score: standard method of calculation

Assume we have a deterministic forecast

		Event forecast?	
		YES	NO
Event Observed?	YES	a	b
	NO	c	d

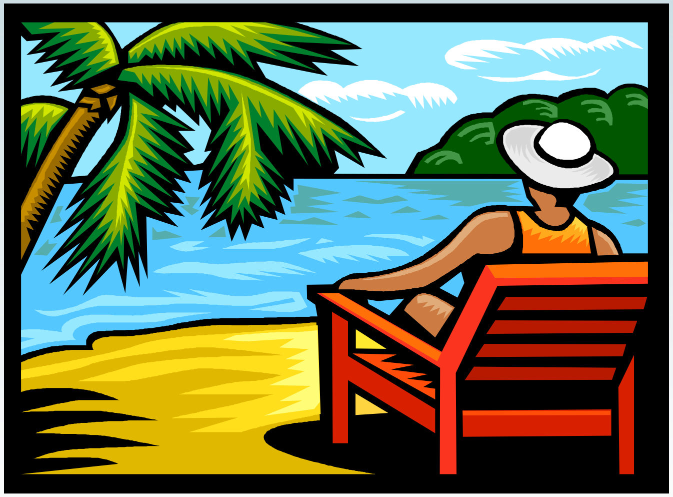
$$ETS = \frac{a - a_r}{a + b + c - a_r}$$ where $$a_r = \{a + c\}\,\{a + b\}$$

Two islands

As α increases…

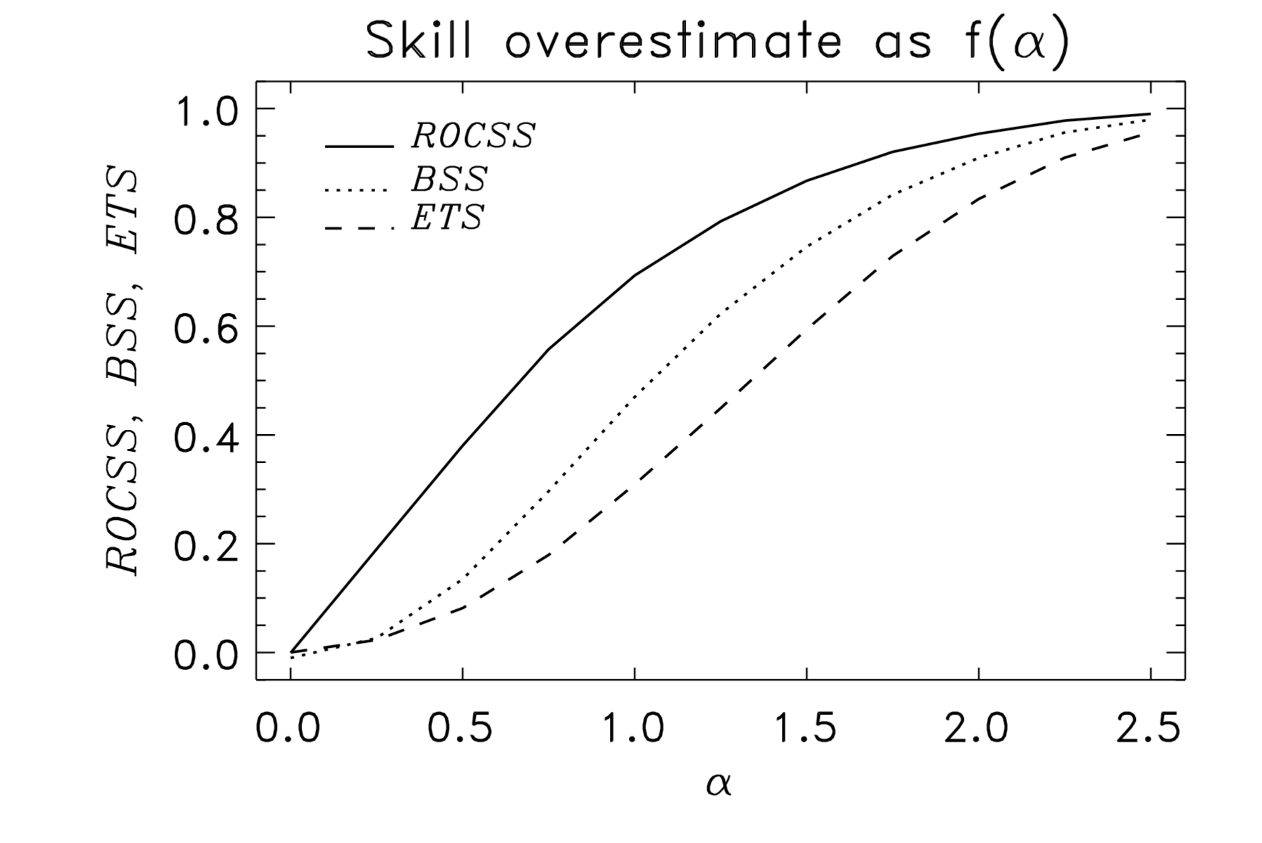
Island 2

Island 1



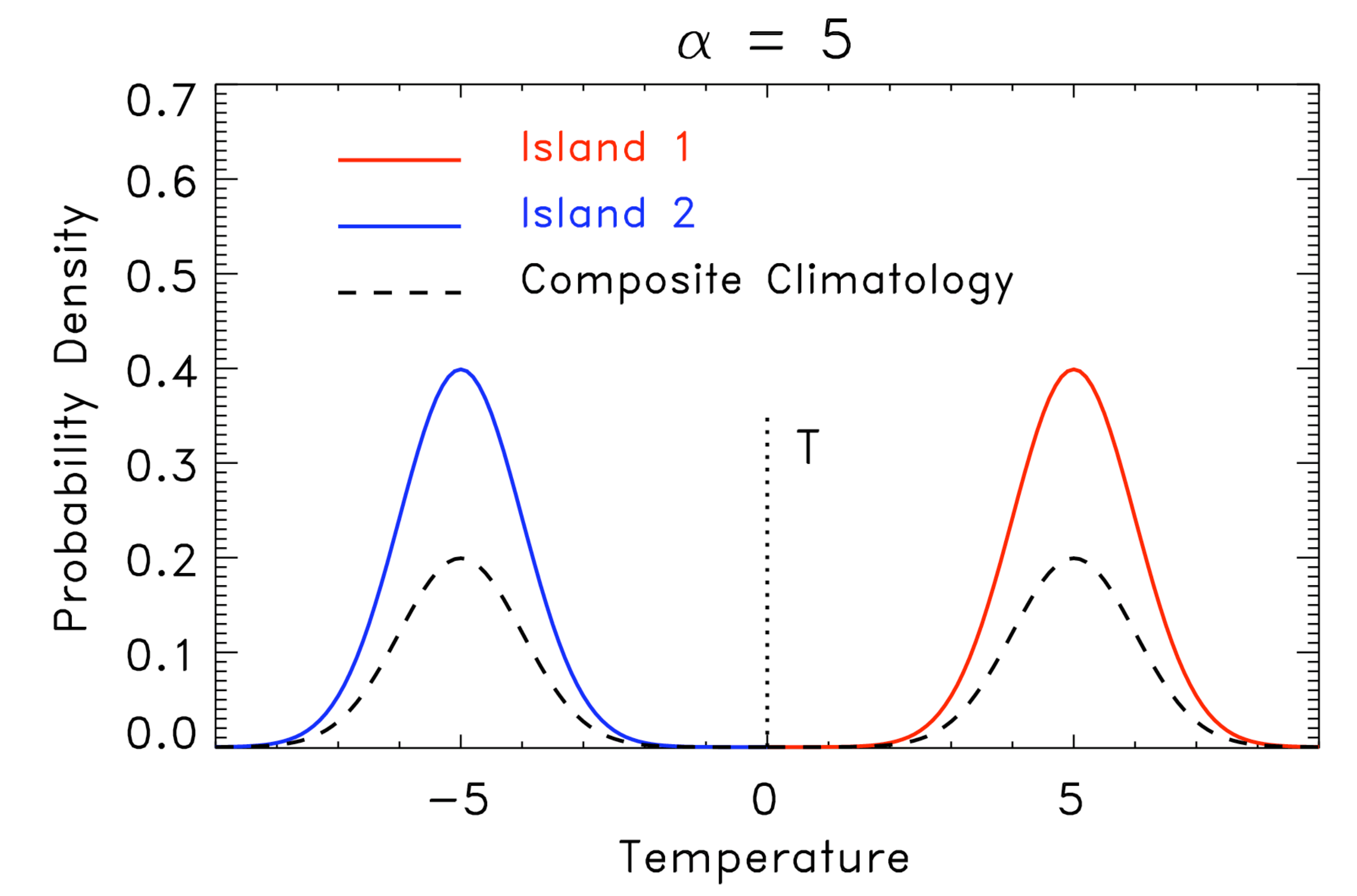
But still, each island’s forecast is no better than
a random draw from its climatology. Expect no skill.

Skill with conventional methods of calculation



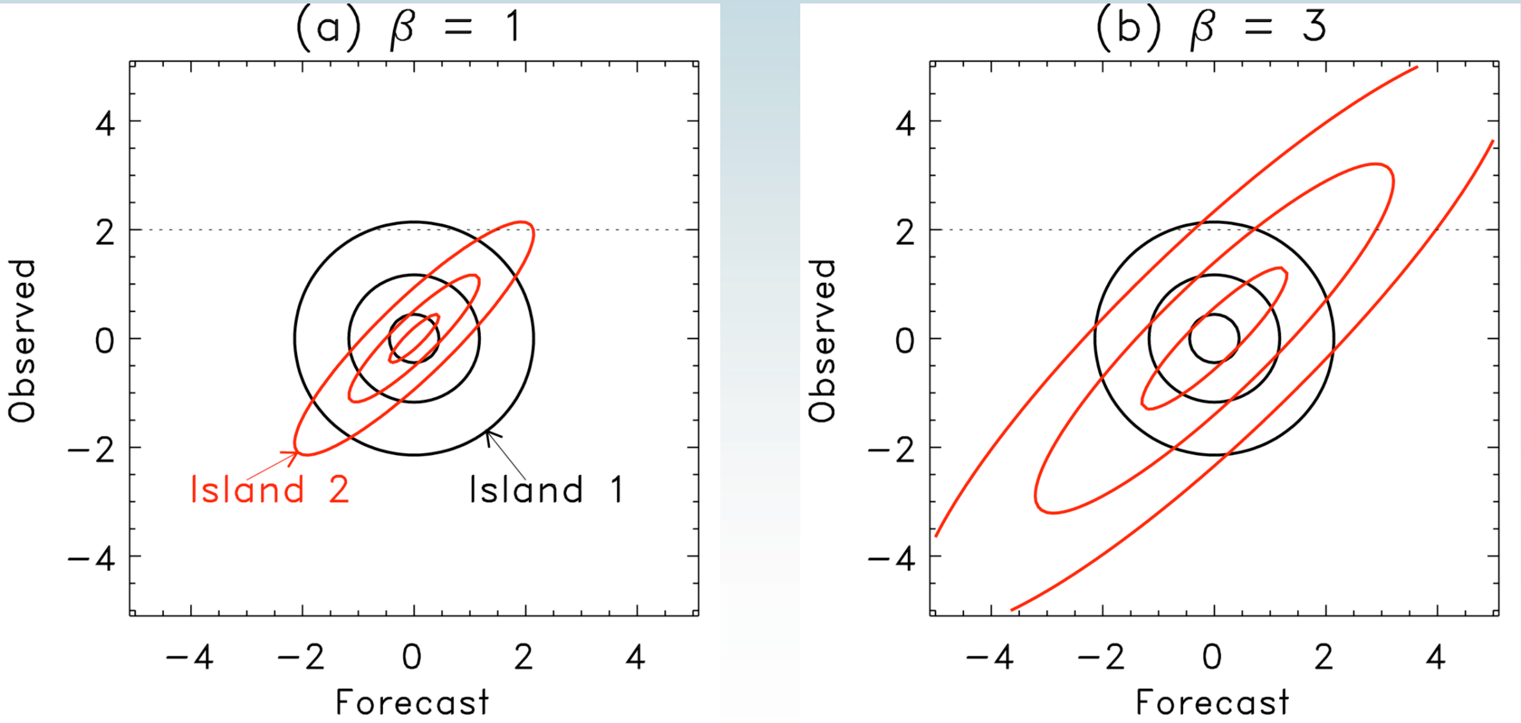
Reference climatology implicitly becomes
$N(+\alpha,1)$ + $N(-\alpha,1)$ not $N(+\alpha,1)$ OR $N(-\alpha,1)$

The new implicit reference climatology



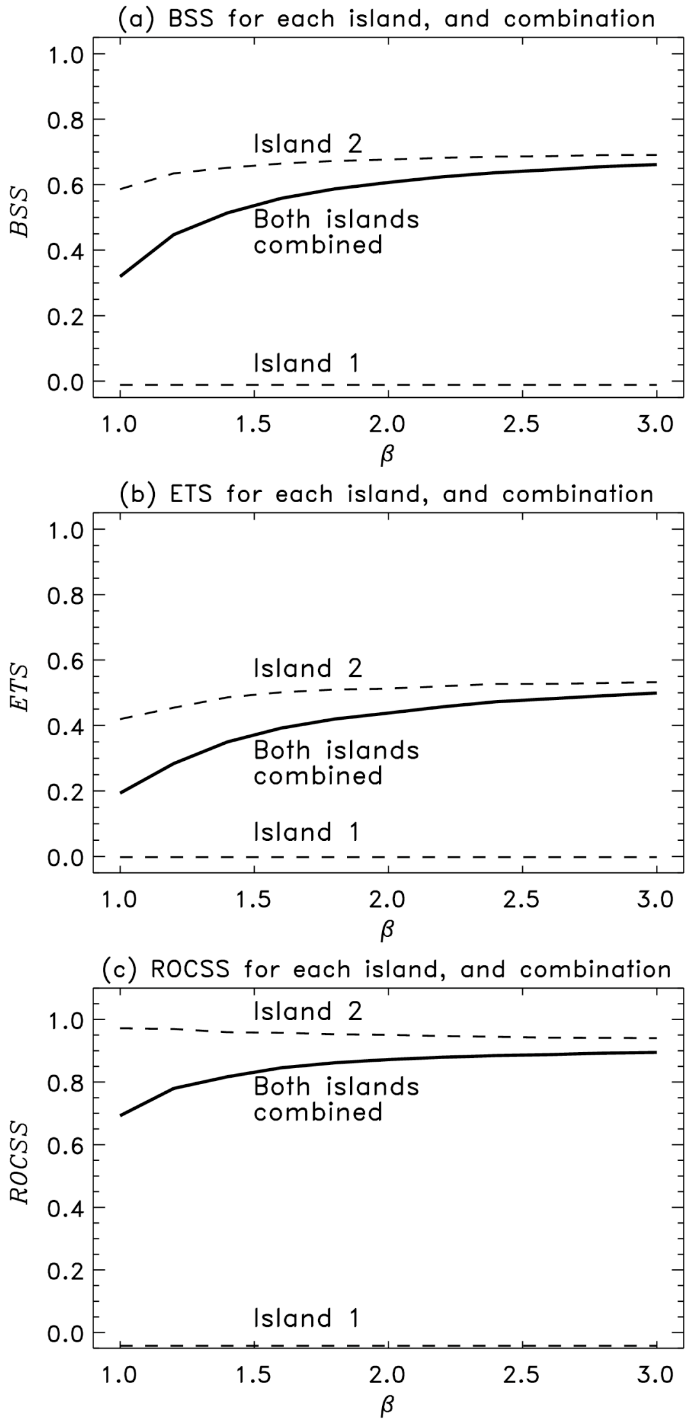
Related problem when means are the same but climatological variances differ

- **Event**: v > 2.0
- **Island 1**: f ~ N(0,1), v ~ N(0,1), Corr (f,v) = 0.0
- **Island 2**: f ~ N(0,α), v ~ N(0,α), 1 ≤ α≤ 3, Corr (f,v) = 0.9



- **Expectation**: positive skill over two islands, but not a function of α

the island with the greater climatological uncertainty of the observed event ends up dominating the calculations.



more

Are standard methods *wrong*?

- **Assertion**: we've just re-defined climatology, they're the correct scores with reference to that climatology.
- **Response**: You *can* calculate them this way, but you shouldn't.

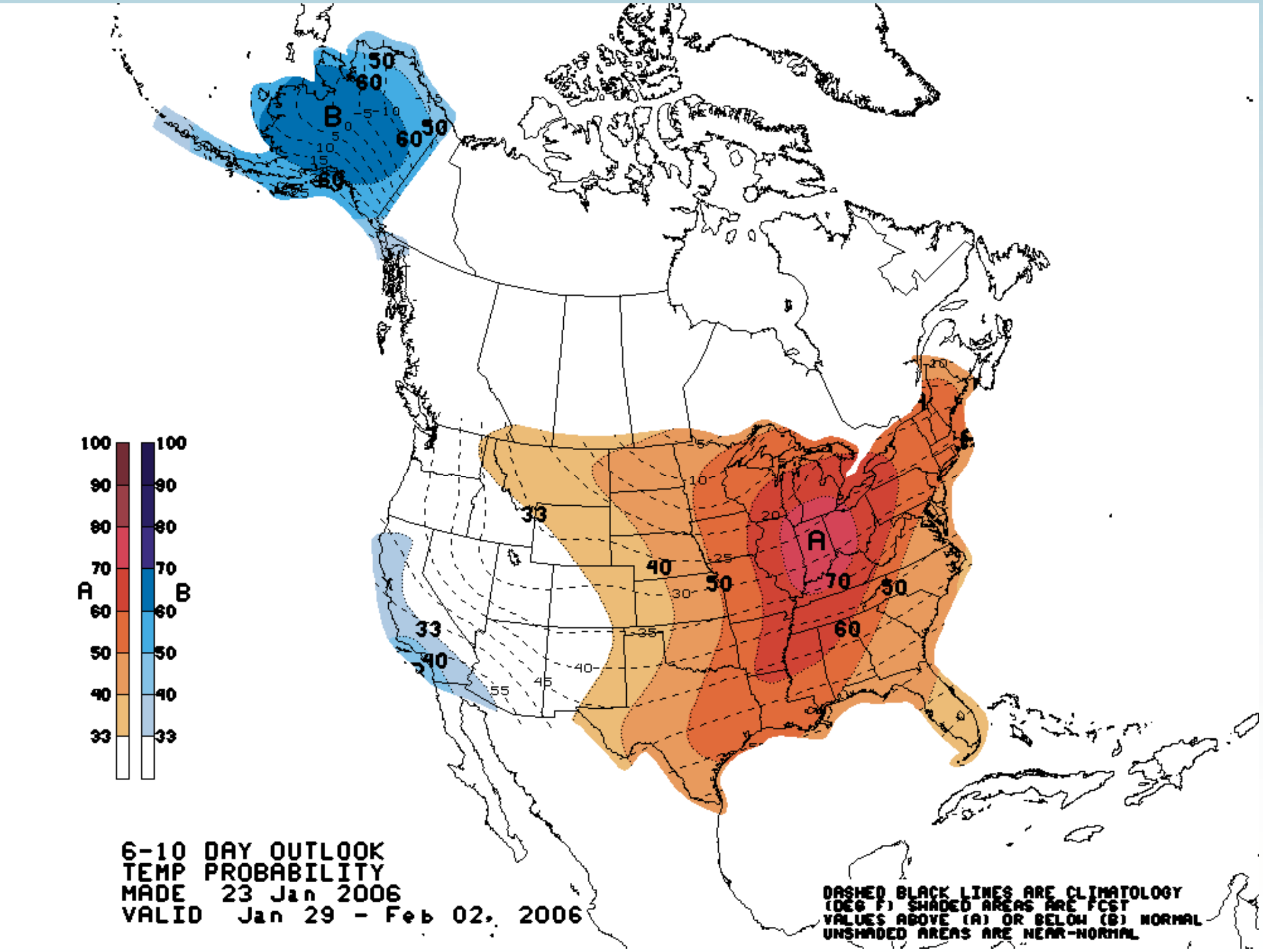
> *"One method that is sometimes used is to combine all the data into a single 2x2 table … this procedure is legitimate only if the probability* ***p*** *of an occurrence (on the null hypothesis) can be assumed to be the same in all the individual 2x2 tables. Consequently, if* ***p*** *obviously varies from table to table, or we suspect that it may vary, this procedure should not be used."*
>
> *W. G. Cochran, 1954, discussing ANOVA tests*

  – You will draw improper inferences due to "lurking variable" - i.e., the varying climatology should be a predictor.

  – Discerning real skill or skill difference gets tougher

Solutions ?

(1) Analyze events where climatological probabilities are the same at all locations, e.g., terciles.



6-10 DAY OUTLOOK
TEMP PROBABILITY
MADE 23 Jan 2006
VALID Jan 29 - Feb 02, 2006

DASHED BLACK LINES ARE CLIMATOLOGY
(DEG F) SHADED AREAS ARE FCST
VALUES ABOVE (A) OR BELOW (B) NORMAL
UNSHADED AREAS ARE NEAR-NORMAL

Solutions, continued

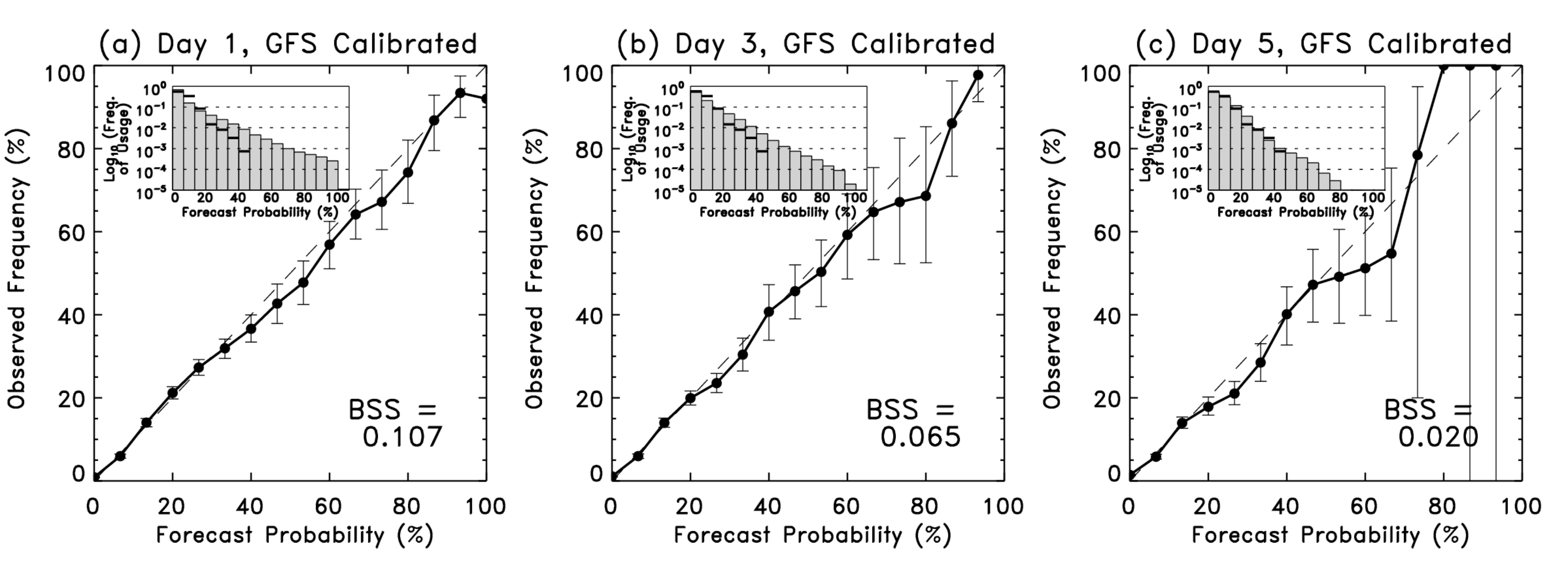
(2) Calculate metrics separately for different points with different climatologies. Form overall number using sample-weighted averages

$$BSS = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \left(1 - \frac{BS_f(k)}{BS_c(k)} \right)$$

ROC:
$$\overline{HR}_i = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \; HR_i(k) \qquad \overline{FAR}_i = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} \; FAR_i(k)$$

$$\overline{ETS} = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} ETS(k)$$

Real-world examples: (1) Why so little skill for so much reliability?

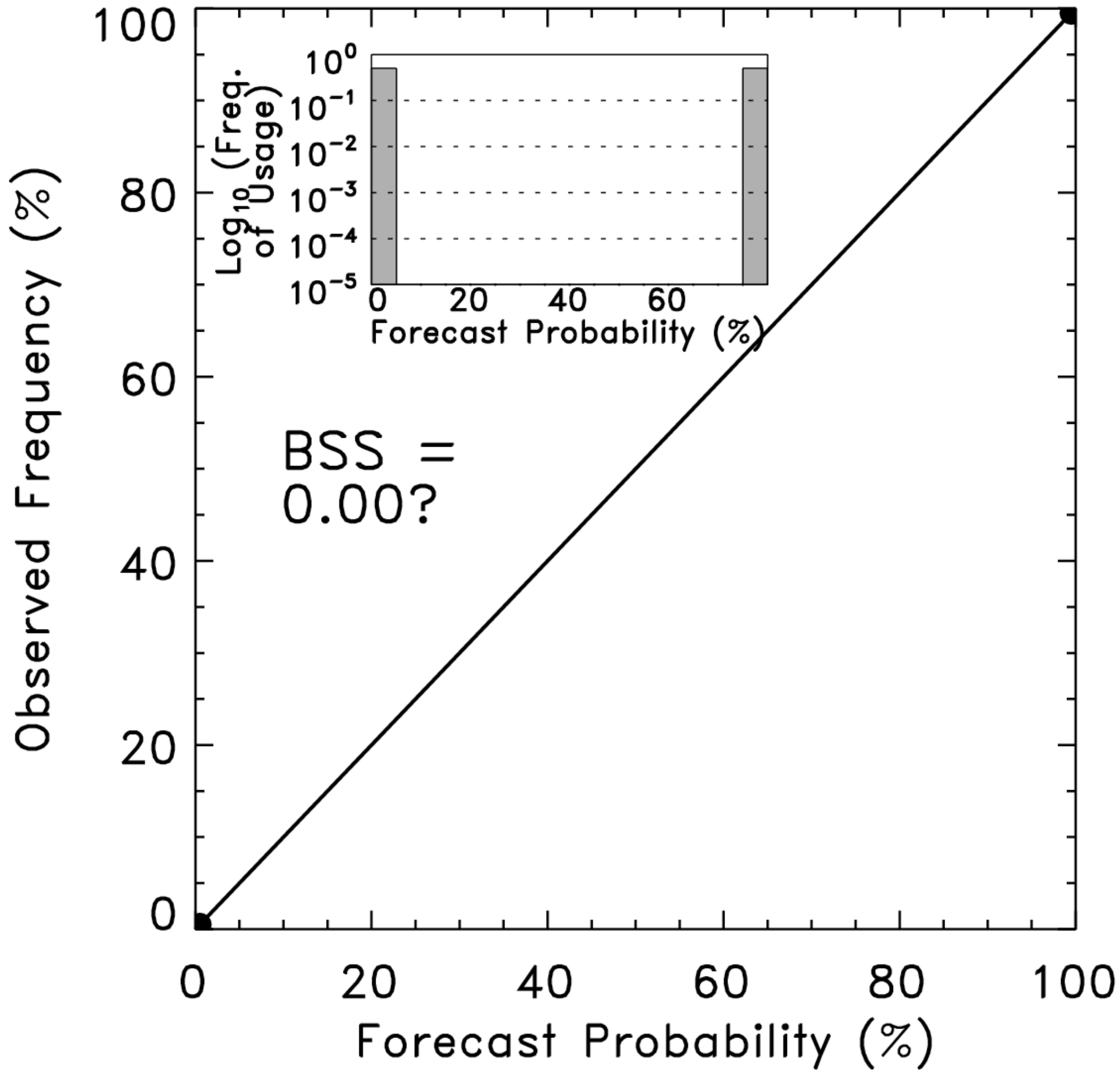

(a) Day 1, GFS Calibrated

(b) Day 3, GFS Calibrated

(c) Day 5, GFS Calibrated

These reliability diagrams formed from locations with different climatologies. Day-5 usage distribution not much different from climatological usage distribution (solid lines).

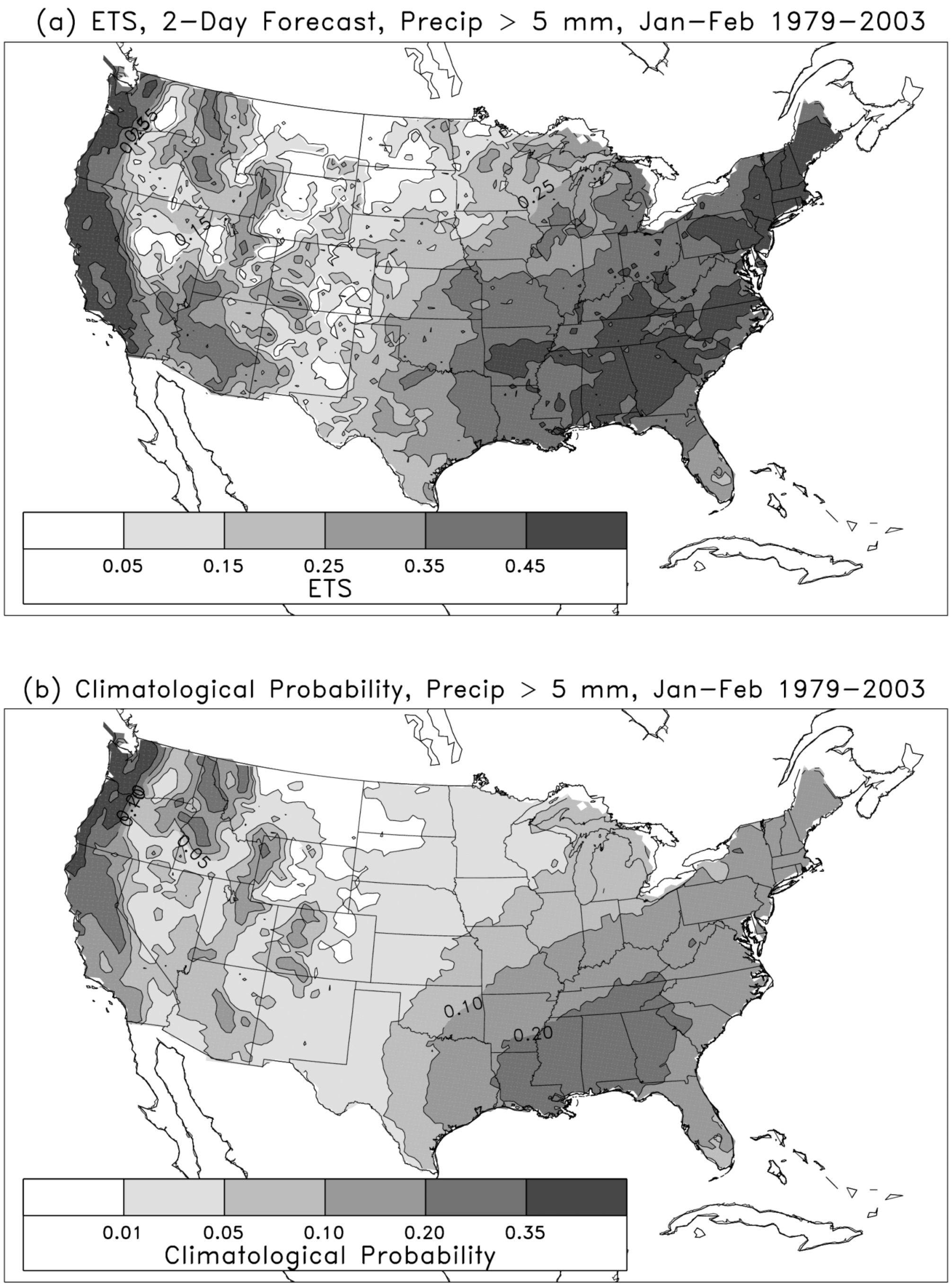
Perfectly Sharp, Perfect Reliability: Is BSS 1.0 or 0.0?



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.

(2) Consider Equitable Threat Scores…

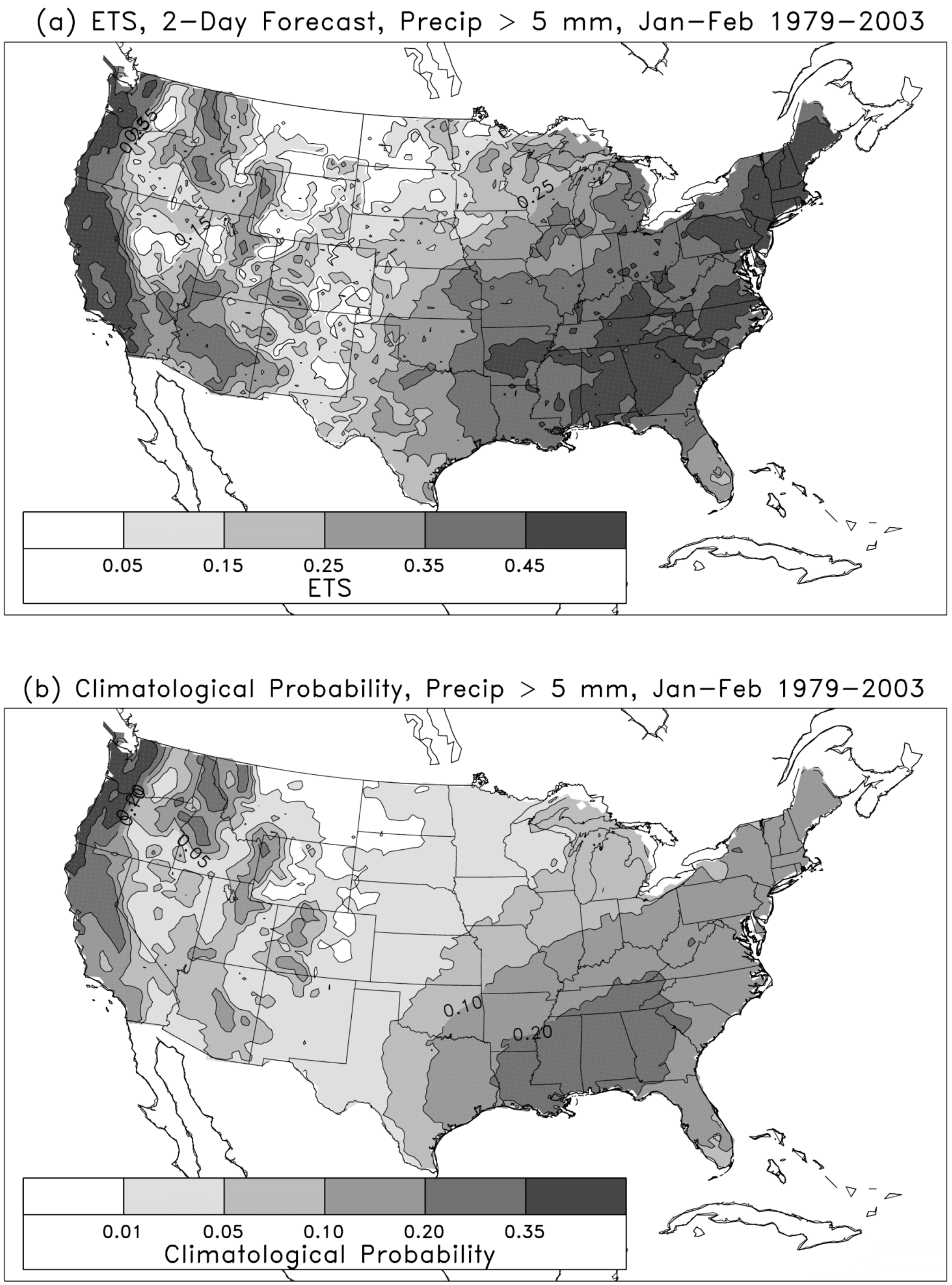


(a) ETS, 2-Day Forecast, Precip > 5 mm, Jan-Feb 1979-2003

(b) Climatological Probability, Precip > 5 mm, Jan-Feb 1979-2003

(2) Consider Equitable Threat Scores…

(1) ETS location-dependent, related to climatological probability.



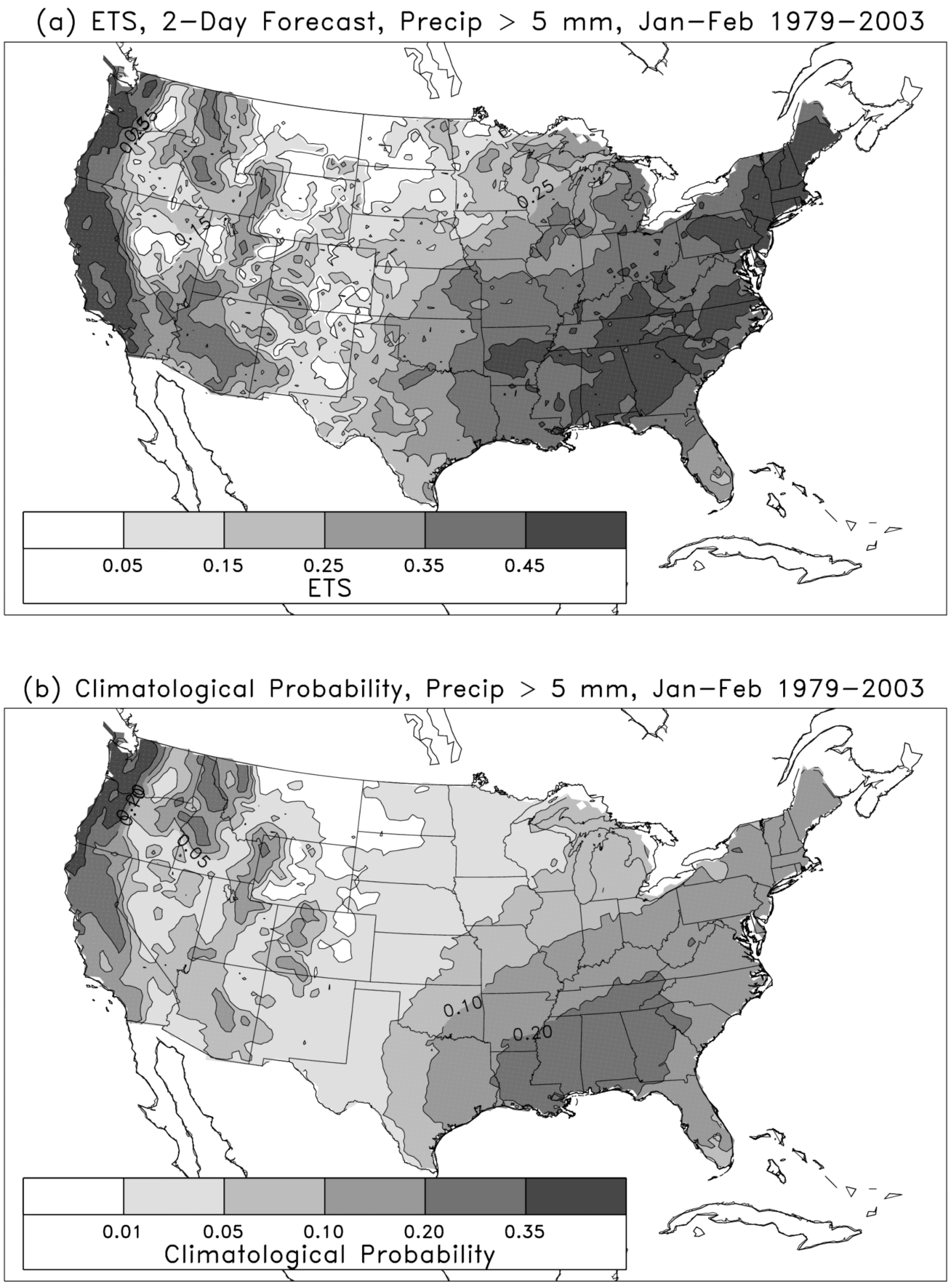
(a) ETS, 2-Day Forecast, Precip > 5 mm, Jan–Feb 1979–2003

(b) Climatological Probability, Precip > 5 mm, Jan–Feb 1979–2003

(2) Consider Equitable Threat Scores…

(1) ETS location-dependent, related to climatological probability.

(2) Average of ETS at individual grid points = 0.28


(a) ETS, 2-Day Forecast, Precip > 5 mm, Jan–Feb 1979–2003

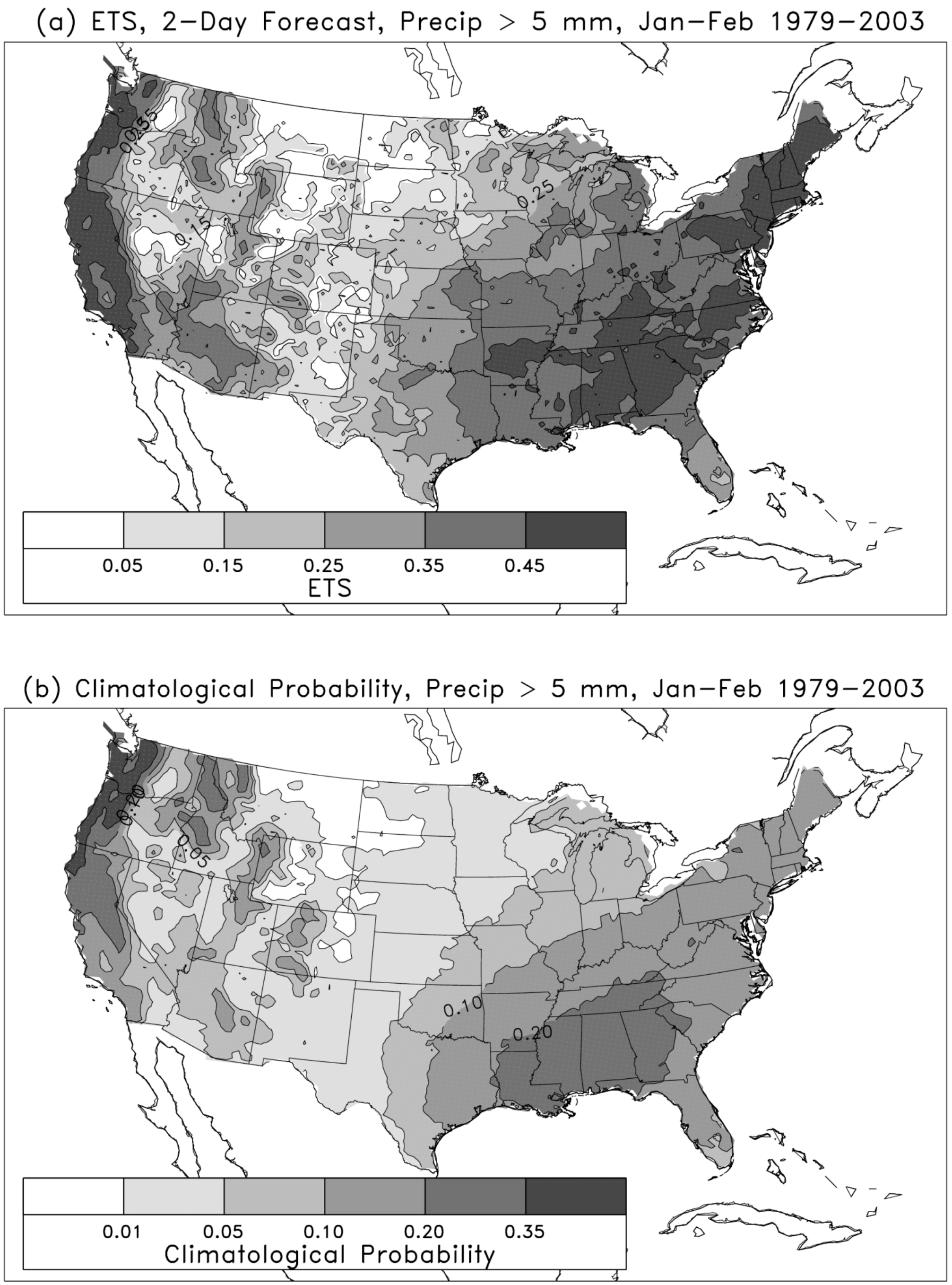
(b) Climatological Probability, Precip > 5 mm, Jan–Feb 1979–2003

(2) Consider Equitable Threat Scores…

(1) ETS location-dependent, related to climatological probability.

(2) Average of ETS at individual grid points = 0.28

(3) ETS after data lumped into one big table = 0.42



(a) ETS, 2-Day Forecast, Precip > 5 mm, Jan-Feb 1979-2003

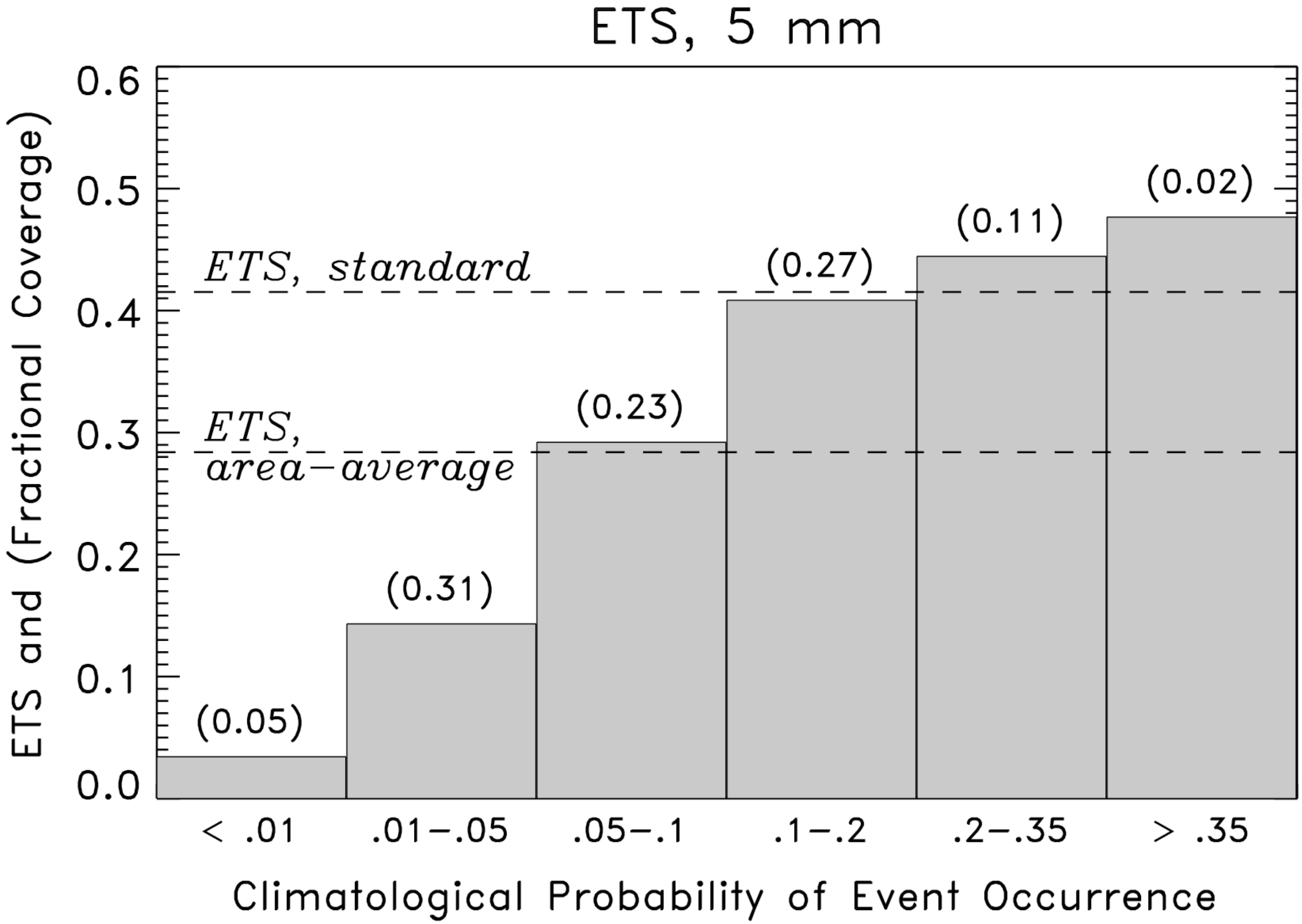
(b) Climatological Probability, Precip > 5 mm, Jan-Feb 1979-2003

Equitable Threat Score: alternative method of calculation

Consider the possibility of different regions with different climates. Assume n_c contingency tables, each associated with samples with a distinct climatological event frequency. $n_s(k)$ out of the m samples were used to populate the kth table. ETS calculated separately for each contingency table, and alternative, weighted-average ETS is calculated as

$$\overline{ETS} = \sum_{k=1}^{n_c} \frac{n_s(k)}{m} ETS(k)$$

ETS calculated two ways



Conclusions

- Many conventional verification metrics like BSS, RPSS, threat scores, ROC, potential economic value, etc. can be overestimated if climatology varies among samples.
  - results in false inferences: think there's skill where there's none.
  - complicates evaluation of model improvements; Model A better than Model B, but doesn't appear quite so since both inflated in skill.

- **Fixes**:

  (1) Consider events where climatology doesn't vary such as the exceedance of a quantile of the climatological distribution

  (2) Combine after calculating for distinct climatologies.

- **Please**: Document your method for calculating a score!

Acknowledgements: Matt Briggs, Dan Wilks, Craig Bishop, Beth Ebert, Steve Mullen, Simon Mason, Bob Glahn, Neill Bowler, Ken Mylne, Bill Gallus, Frederic Atger, Francois LaLaurette, Zoltan Toth, Jeff Whitaker.