Understanding Skill Scores

Michael K. Tippett

Columbia University

Climate Diagnostics and Prediction Workshop 2018

John von Neumann

- “Young man, in mathematics you don't *understand* things. You just *get used* to them.”
- “It is exceptional that one should be able to acquire the understanding of a process without having previously acquired a *deep familiarity* with running it, with using it, before one has assimilated it in an instinctive and empirical way”

Tony’s deep familiarity with forecasting, forecast verification and skill scores come from:

What:

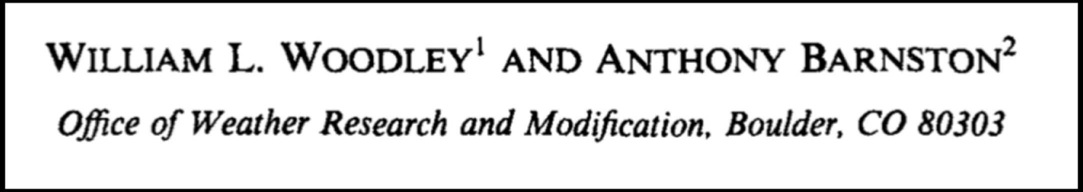
- ENSO
- 2m-temperature , precipitation
- Drought, heat waves
- Sea level
- Hospitalization, malaria
- Death on the Titanic
- “The effect of weather on mood, productivity, and frequency of emotional crisis …”

How:

- GCMs
- Analogues
- CCA
- QBO
- 11-year solar cycle
- Irrigation
- Cloud seeding!

Where:

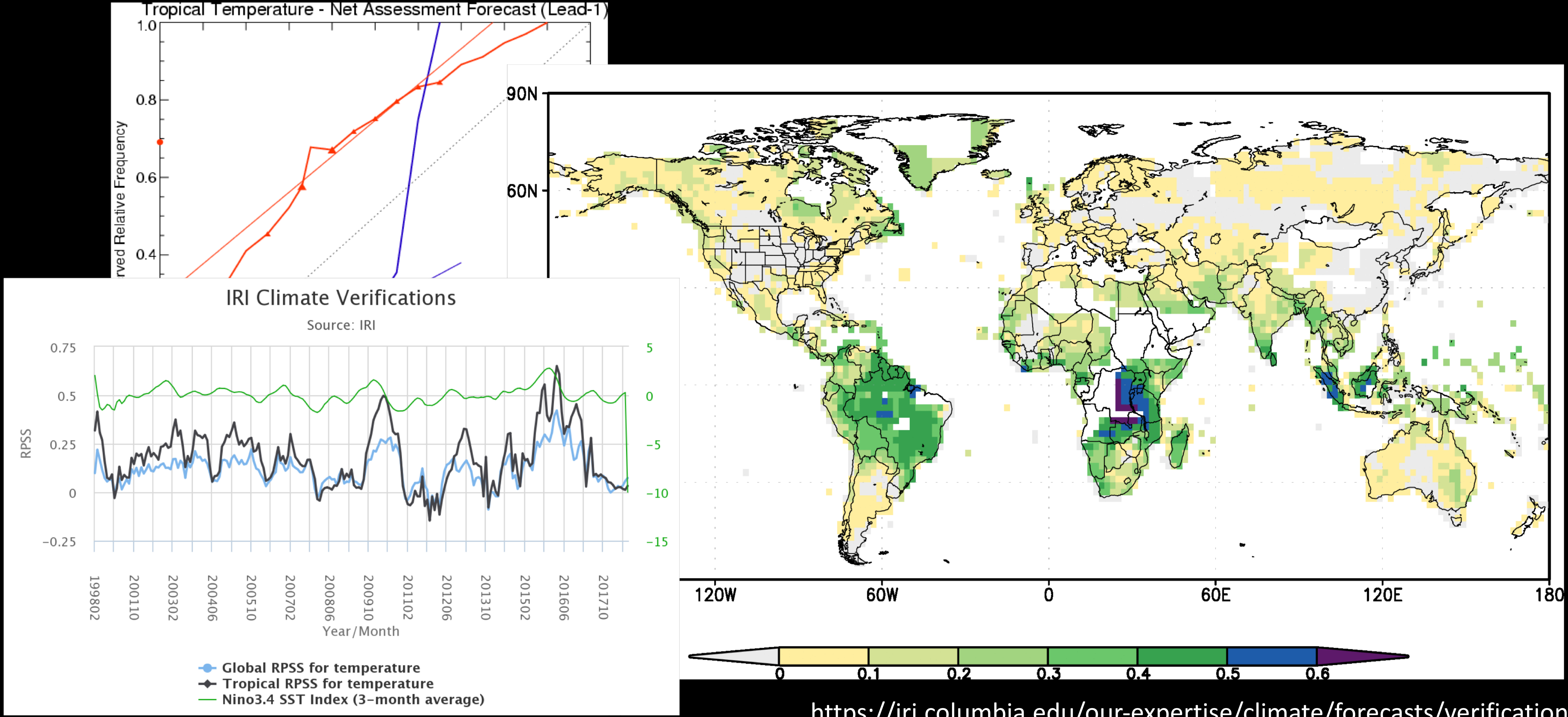
- “A highly ENSO-related region …”
- Northern Europe
- East, west, and central Asia
- Ecuador
- U.S. (and Hawaii, Alaska, Florida, and Great Plains)
- Africa, Ethiopia
- Tropical Pacific Islands



WILLIAM L. WOODLEY[1] AND ANTHONY BARNSTON[2]

Office of Weather Research and Modification, Boulder, CO 80303

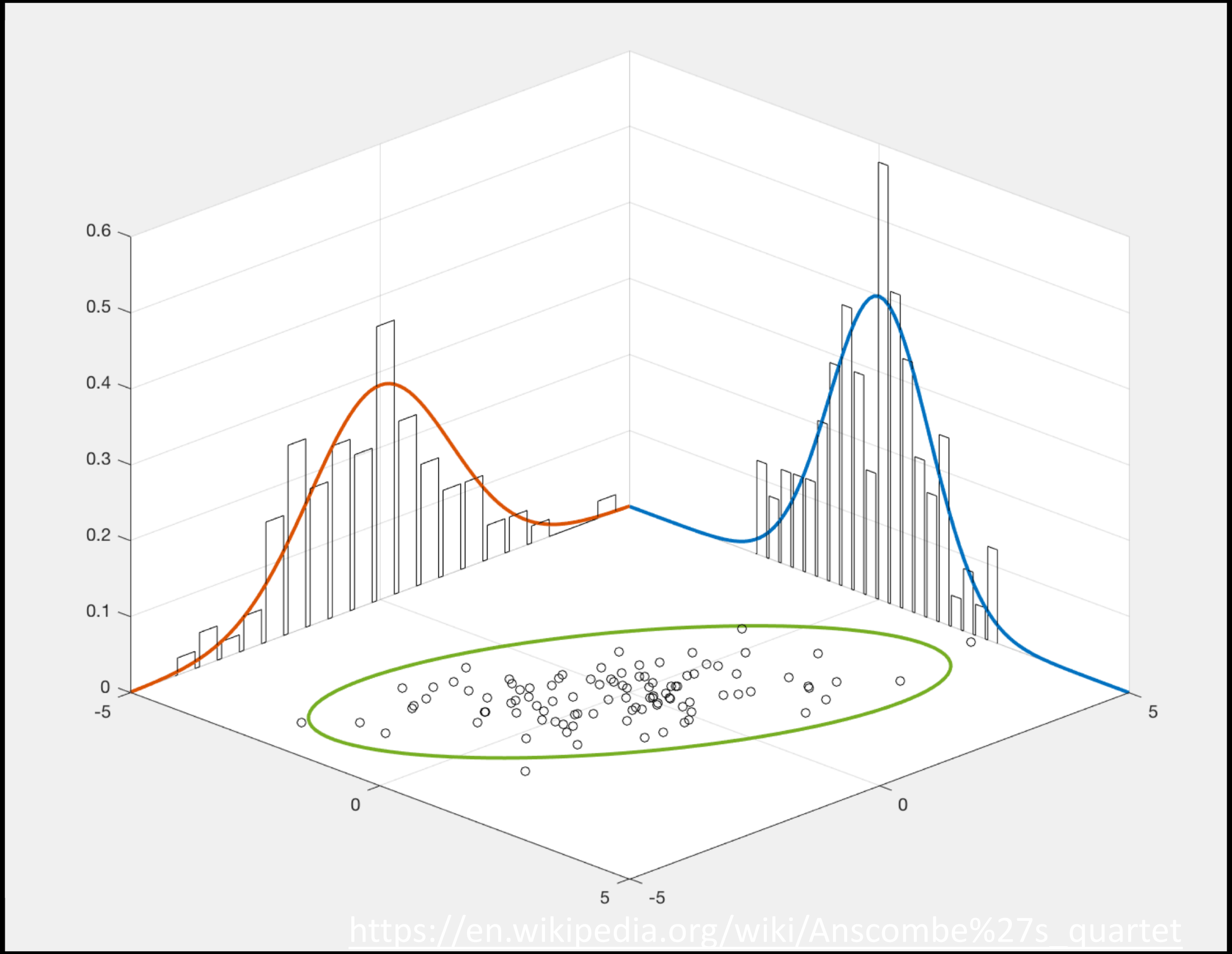
What about the rest of us?



https://iri.columbia.edu/our-expertise/climate/forecasts/verification/

What about the rest of us?

- Idea: Relate less familiar skill scores to more familiar ones
- Everyone understands correlation …
- Everyone understands Gaussian distributed forecast and observations



https://en.wikipedia.org/wiki/Anscombe%27s_quartet

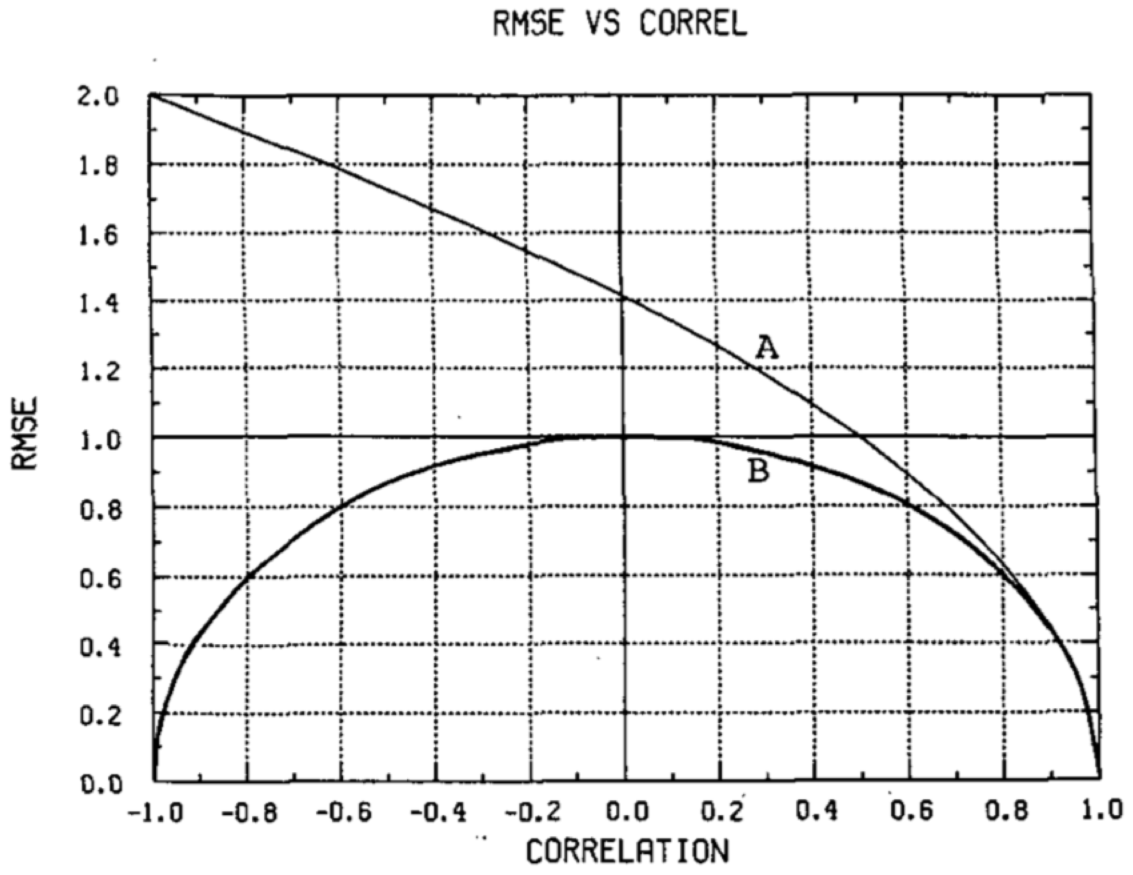
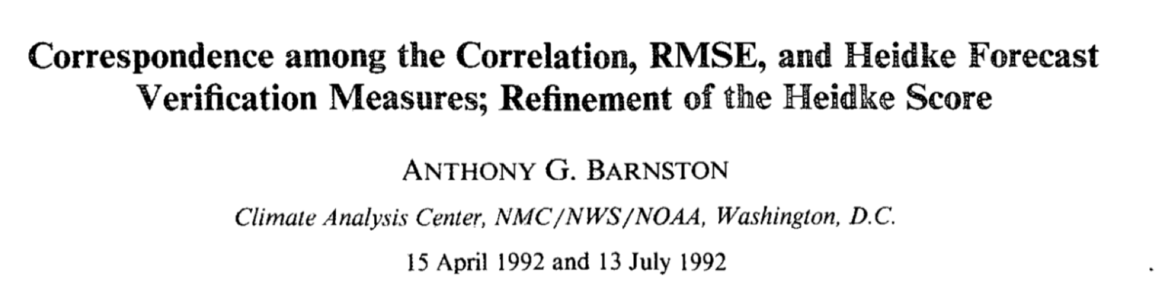


FIG. 1. Root-mean-square error (RMSE) as a function of correlation for standardized sets of forecasts and observations (curve A), and for same except that the forecasts have been damped and possibly sign reversed by multiplying by r_{fo}—i.e., the correlation between forecasts and observations (curve B).



Correspondence among the Correlation, RMSE, and Heidke Forecast Verification Measures; Refinement of the Heidke Score

ANTHONY G. BARNSTON

Climate Analysis Center, NMC/NWS/NOAA, Washington, D.C.

15 April 1992 and 13 July 1992

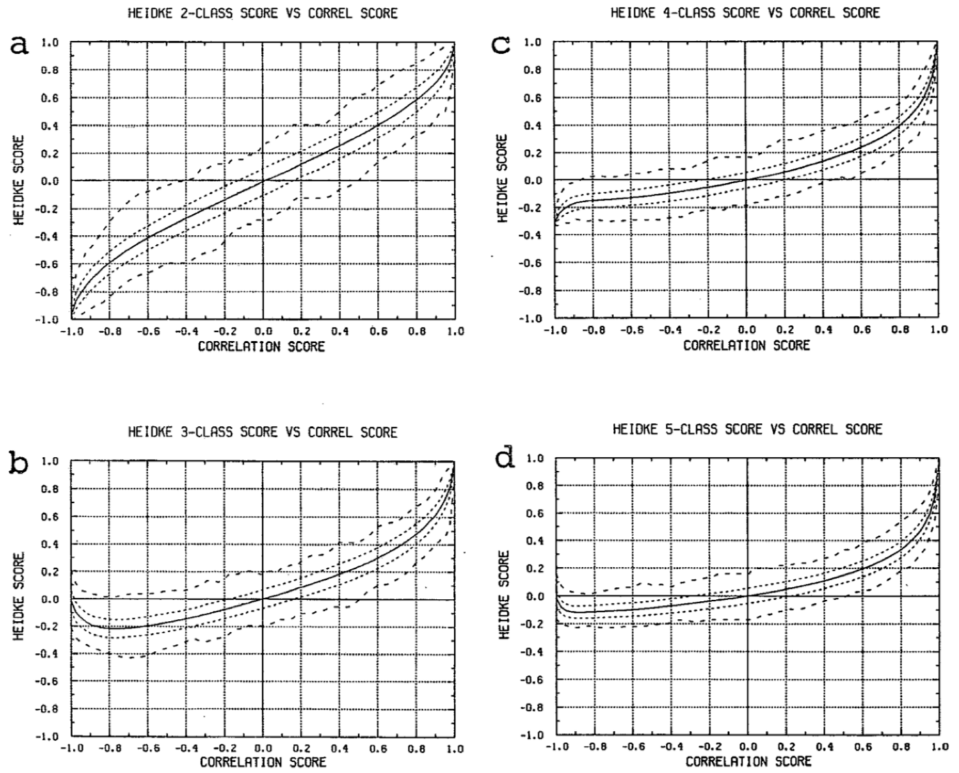
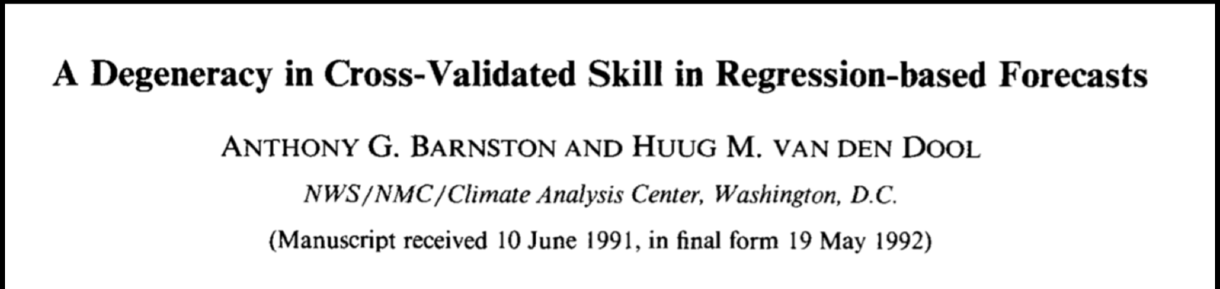


FIG. 3. Heidke score as a function of correlation score for (a) two, (b) three, (c) four, and (d) five equally likely categories, based on a large number of simulations using a random number generator. The solid curve represents mean results, the short-dashed curves the plus- and minus-one standard deviation interval, and the long-dashed curves the maximum and minimum results.

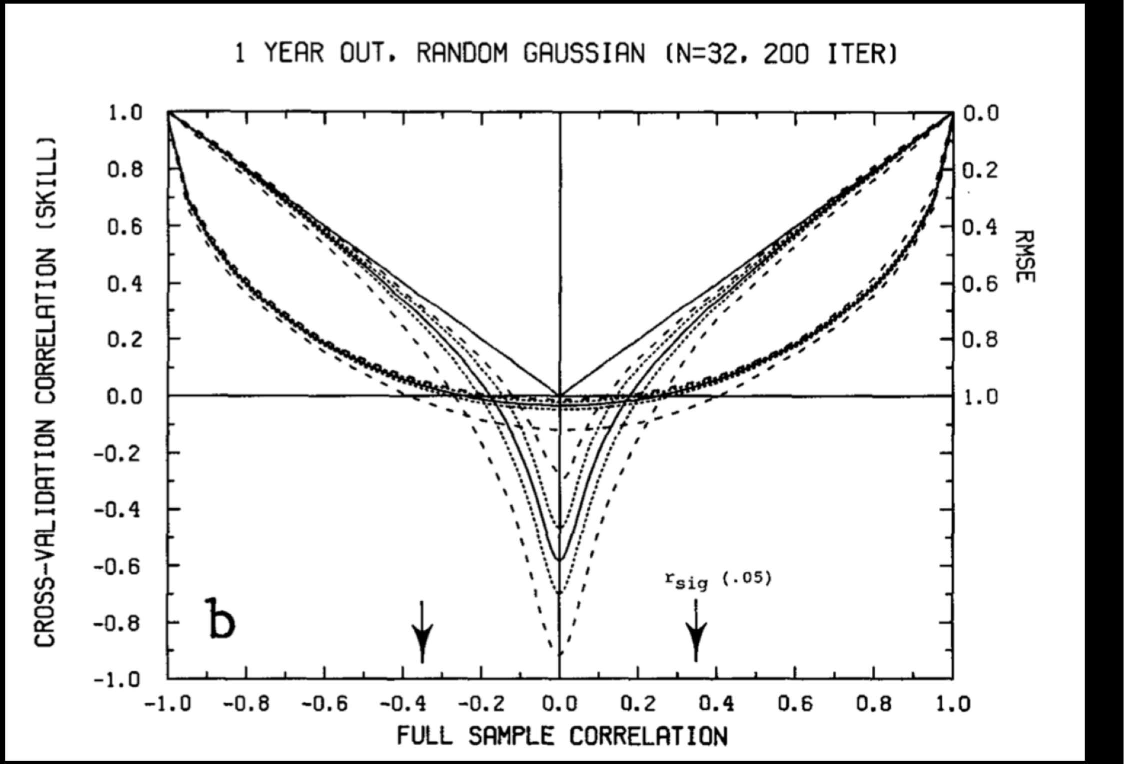


A Degeneracy in Cross-Validated Skill in Regression-based Forecasts

ANTHONY G. BARNSTON AND HUUG M. VAN DEN DOOL

NWS/NMC/Climate Analysis Center, Washington, D.C.

(Manuscript received 10 June 1991, in final form 19 May 1992)



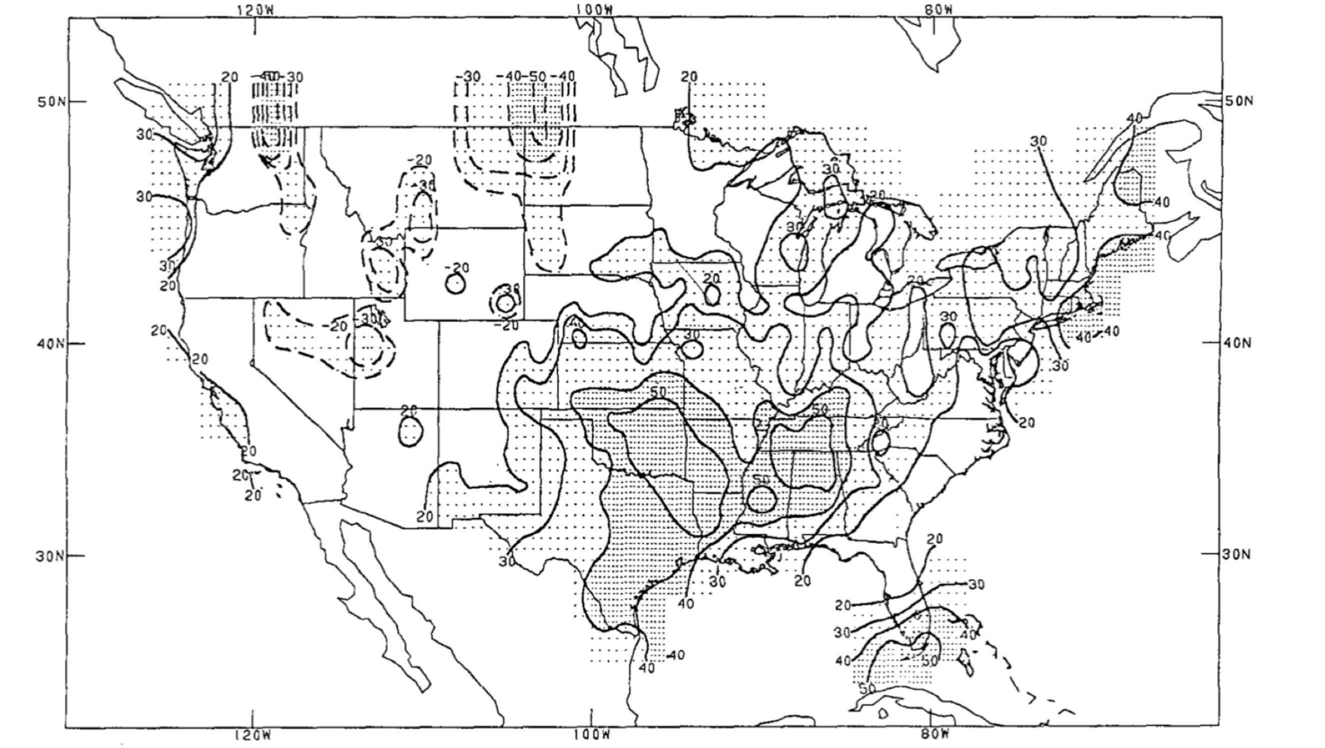


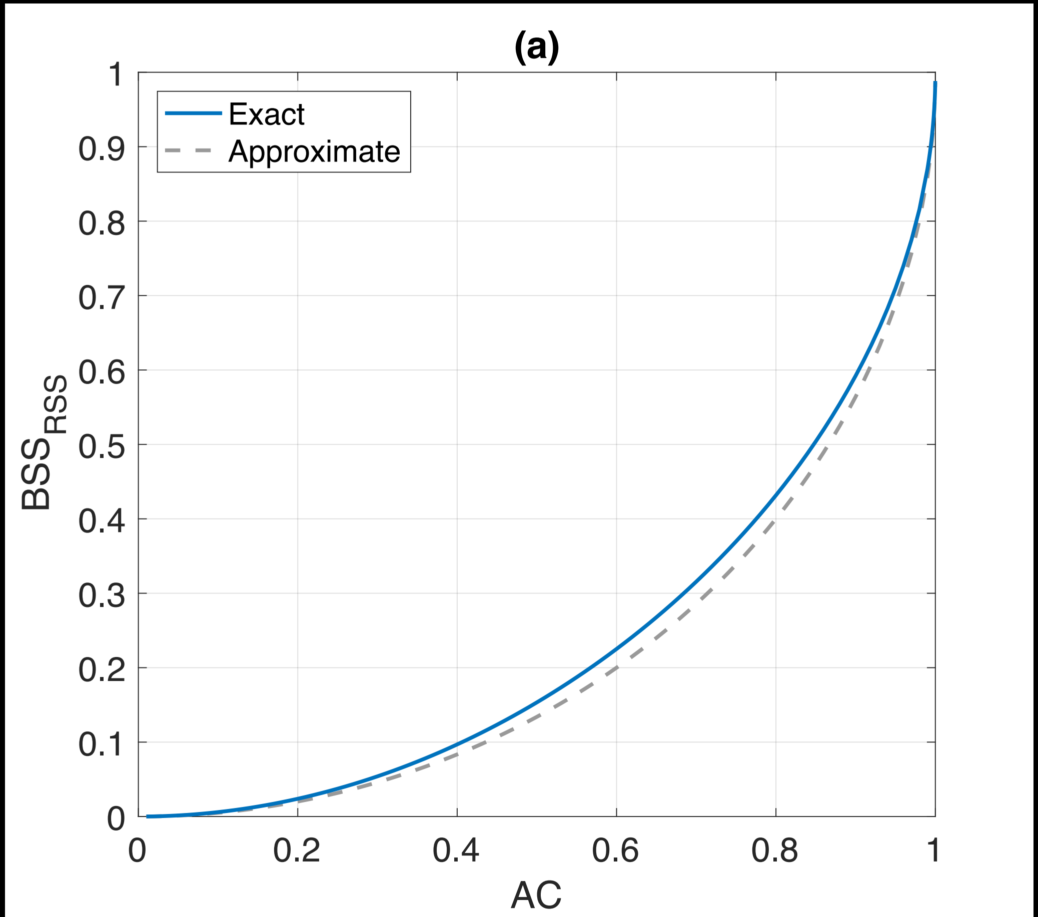
FIG. 5. Illustration of the subject degeneracy in a study using correlation-verified cross-validation, two-predictor multiple regression in prediction of monthly mean surface air temperature anomalies from previous month's temperature and precipitation anomalies (Huang and Van den Dool 1993). Here the geographic distribution of cross-validation skill is shown in predicting August temperature using July predictors. Each trial of cross-validation holds out 1 year as the forecast target and uses all remaining years to develop a regression equation. Units are correlation ×100. The −0.1, 0.0, and 0.1 contours are not shown. Areas of correlation skill score degeneracy are found in the northwestern United States with a minimum value of −0.51 in northwestern North Dakota and northeastern Washington.

Example: “Understanding” the Brier score

- Current probability forecast for El Nino (DJF) is **P=72%**.
- The Brier skill of that forecast is:
  - $(0\text{-}P)^2$=0.52 if El Nino does not occur. $(1\text{-}P)^2$ = 0.078 if El Nino does occur.
  - Smaller values are better.
- The average Brier score of many such forecasts (all P=72%) is
  - P(1-P)=0.2, if our forecasts are reliable
- What about the average Brier score for forecasts with different P’s?
  - Need a model for how forecast strength varies
  - Joint-Gaussian: strength depends on signal and correlation r
- For terciles, average Brier skill score is approximately $1-\sqrt{1-r^2}$

Tippett, Barnston, DelSole (2010)

Good approximation in theory

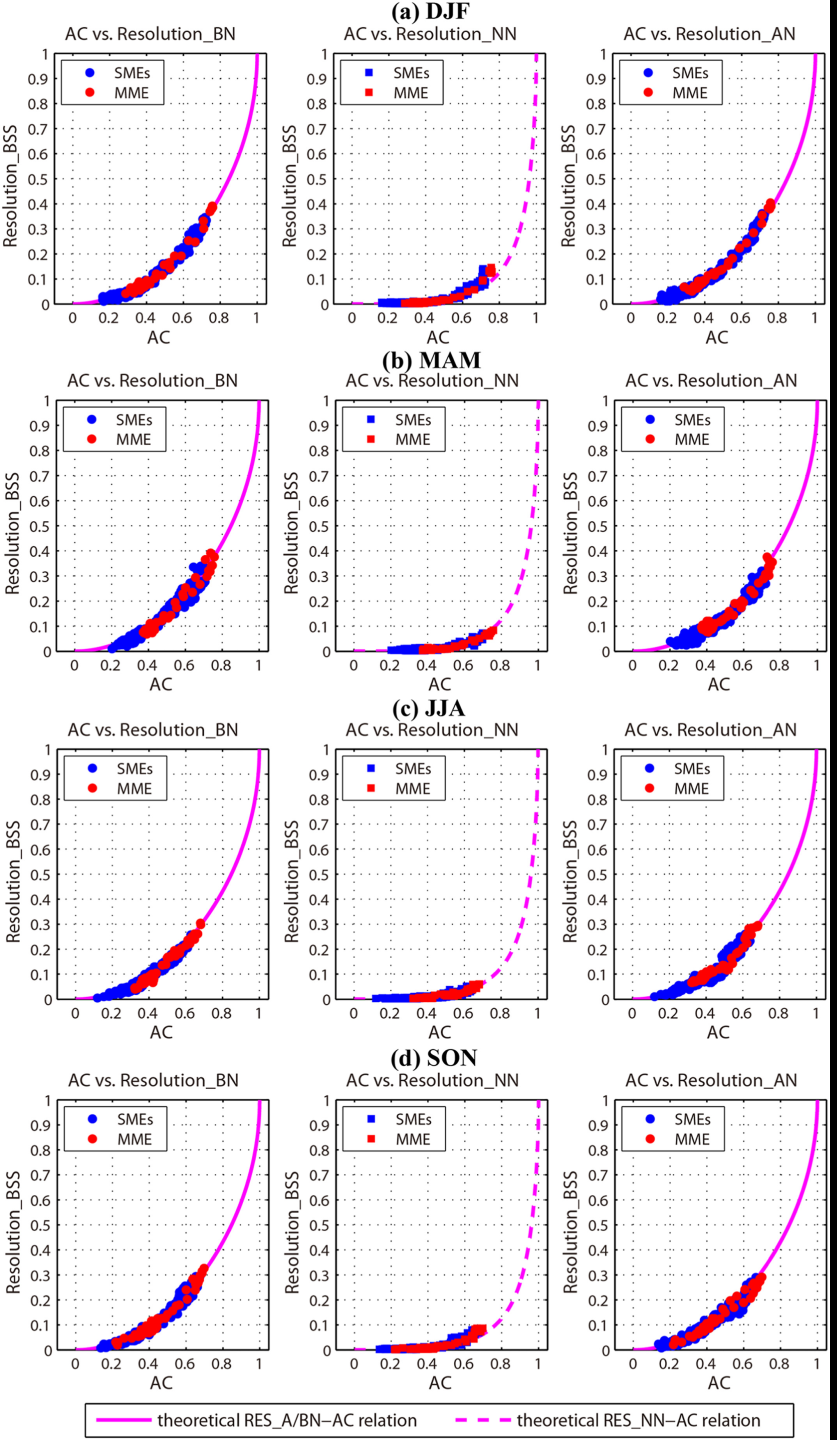


Why BSS and RPSS values are “small”

And practice

ENSEMBLES data

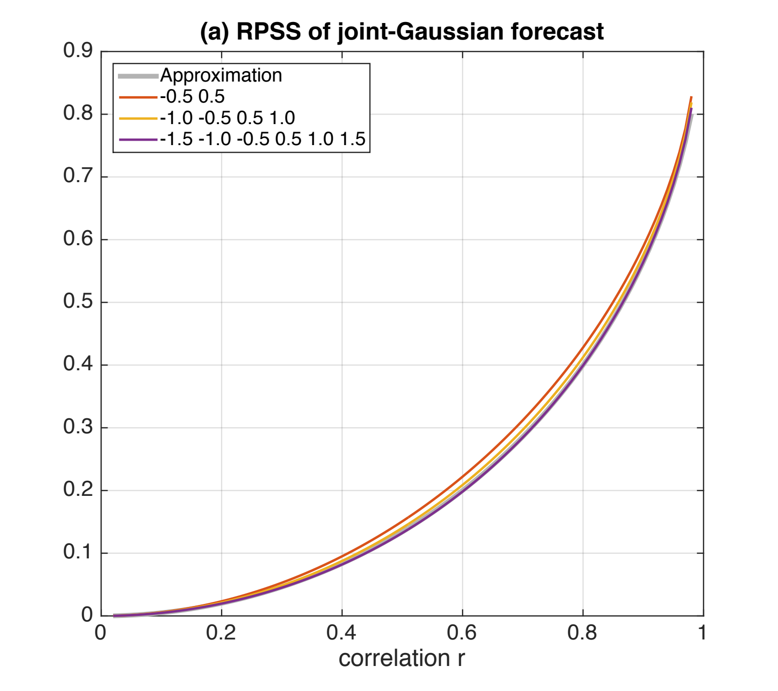
Yang et al. (2018). On the relationship between probabilistic and deterministic skills in dynamical seasonal climate prediction. *JGR*, 123, 5261–5283.



How do probability skill scores vary with number of categories?

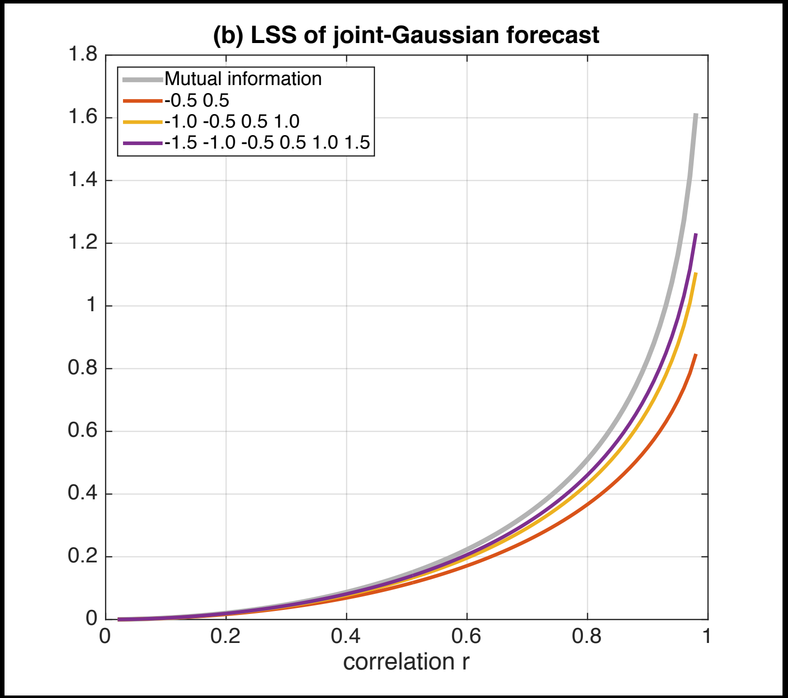
RPSS

- Not much



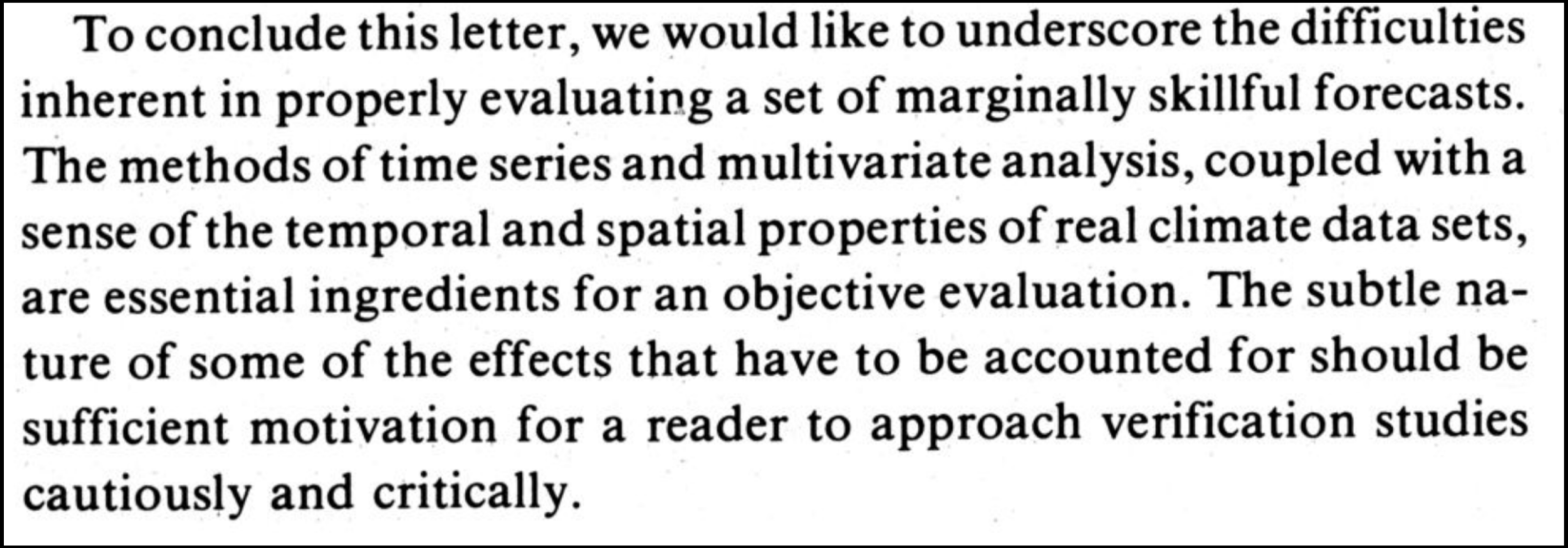
Log skill score

- More categories, more information, higher score
  - Limiting case $\mathrm{LSS} = -\frac{1}{2}\log(1 - r^2)$

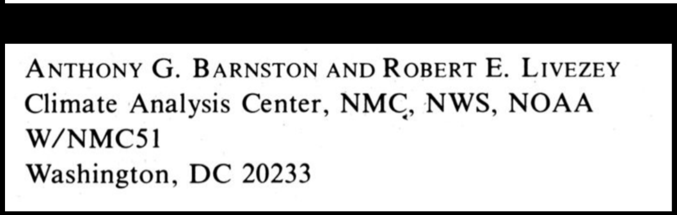


M. K. Tippett, M. Ranganathan, M. L'Heureux, A. G. Barnston, and T. DelSole. 2017.

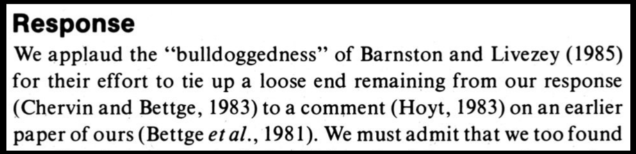
Trust, but verify (the verification)—Tony B.



To conclude this letter, we would like to underscore the difficulties inherent in properly evaluating a set of marginally skillful forecasts. The methods of time series and multivariate analysis, coupled with a sense of the temporal and spatial properties of real climate data sets, are essential ingredients for an objective evaluation. The subtle nature of some of the effects that have to be accounted for should be sufficient motivation for a reader to approach verification studies cautiously and critically.

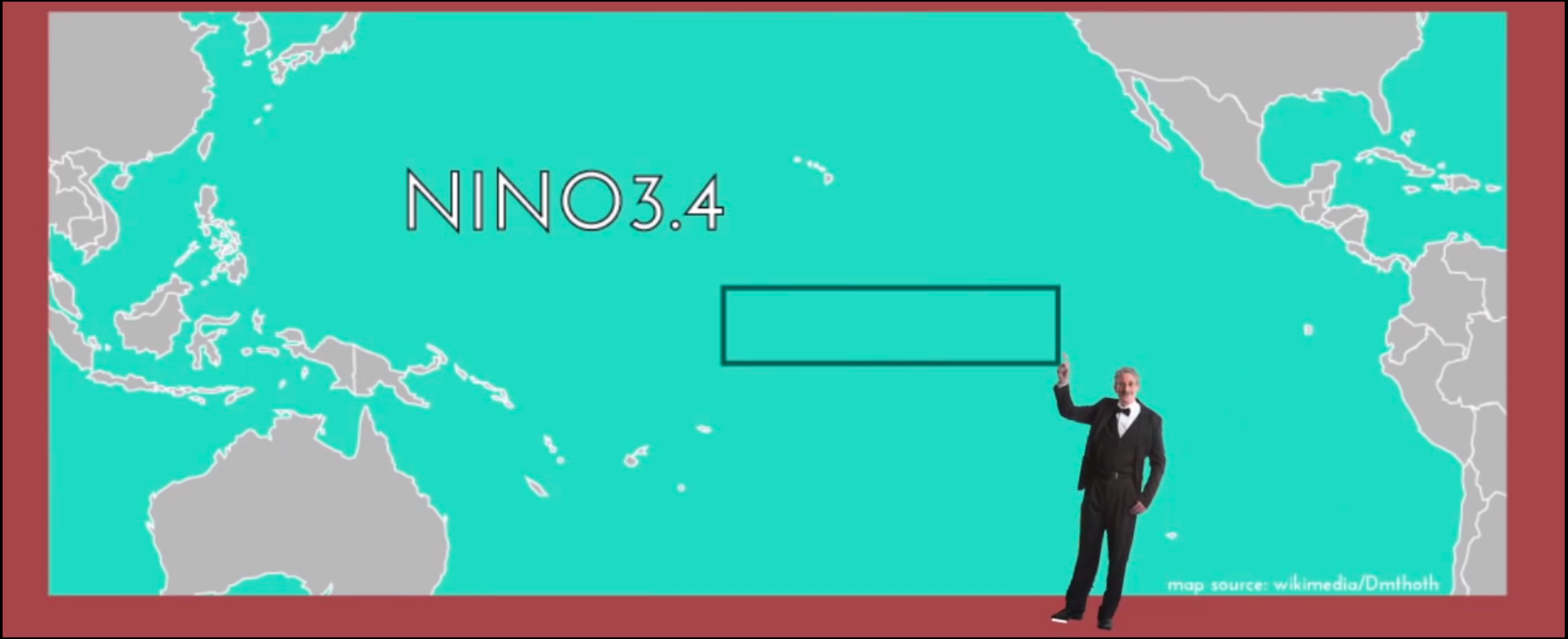


ANTHONY G. BARNSTON AND ROBERT E. LIVEZEY
Climate Analysis Center, NMC, NWS, NOAA
W/NMC51
Washington, DC 20233



Response

We applaud the "bulldoggedness" of Barnston and Livezey (1985) for their effort to tie up a loose end remaining from our response (Chervin and Bettge, 1983) to a comment (Hoyt, 1983) on an earlier paper of ours (Bettge *et al.*, 1981). We must admit that we too found



map source: wikimedia/Dmthoth