Verification and evaluation of a national probabilistic prediction system

Barbara Brown
NCAR
23 September 2009

Goals… (Questions)

- What are desirable attributes of verification capabilities for probabilistic forecasts?
- What should a national verification capability encompass?

Verification terminology and links

Verification is

"The process of assessing the *quality* of forecasts"

But – *quality* is closely linked to *value*

Other words…

Evaluation

Assessment

Verification links

Forecast verification is tightly linked (*integral*) to all aspects of forecast development, production, communication, and use

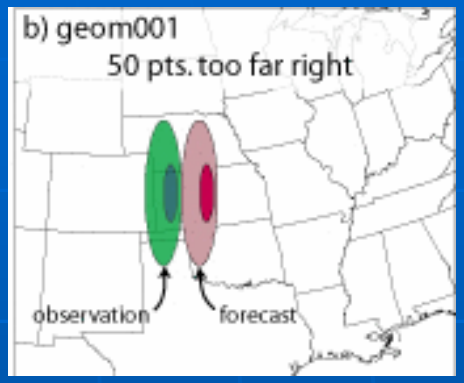
Verification has significant impacts on

- Forecast development
  - Optimization of forecast models and systems
    - *E.g., through choice of verification approaches, measures, and variables*
  - Post-processing
- Presentation of forecast information
- Decision making processes

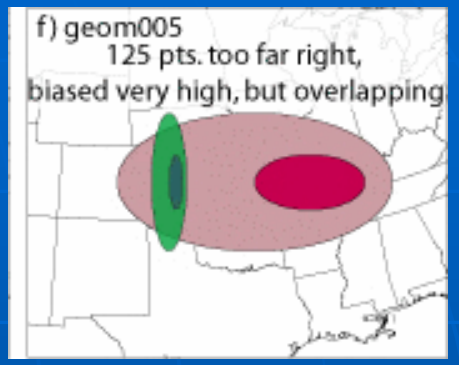
Verification challenges

- Verification results should be understandable by the users of the information

  Ex: What does a CSI of 0.28 really mean?

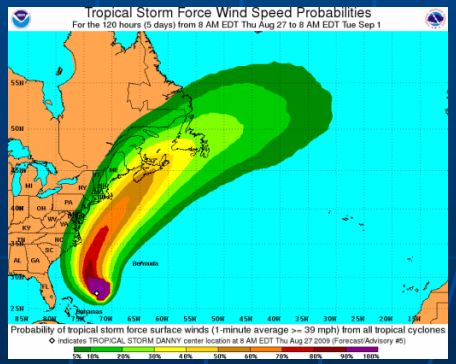


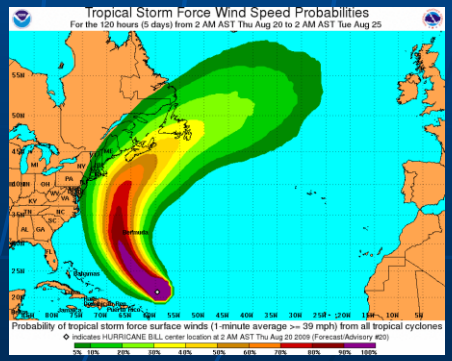
Correlation = -0.02
POD = 0.00
FAR = 1.00
GSS (ETS) = -0.01



Correlation = 0.2
POD = 0.88
FAR = 0.89
GSS (ETS) = 0.08

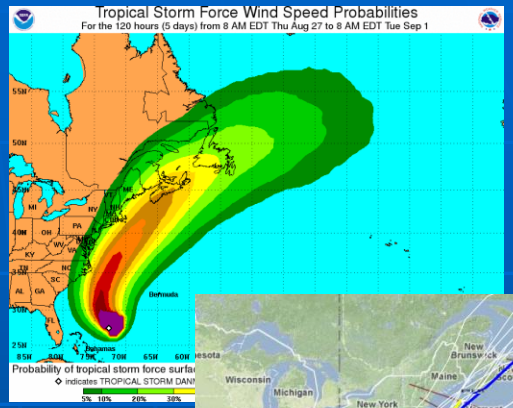
- Verification measures should appropriately distinguish and measure differences in performance

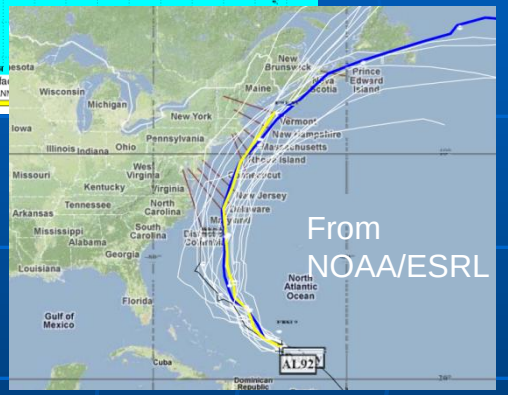


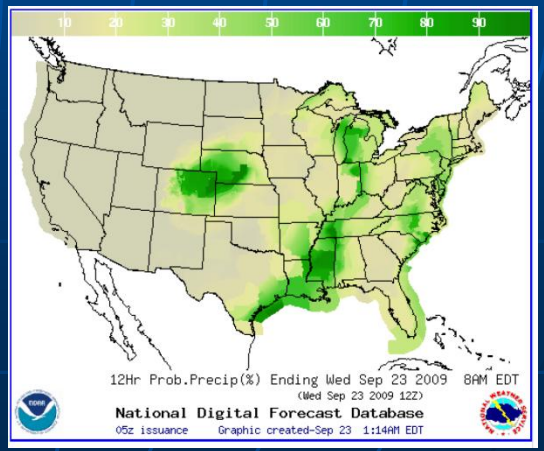


Factors impacting evaluation of probabilistic forecasts

- Variable of interest
- Forecast type:
  - Distribution
  - Probability
  - Point / space
- Verification attributes of interest
  - Reliability, Accuracy, Discrimination
  - Spread-skill
  - Timing errors; temporal changes
  - Spatial attributes (size, location, etc.)
- User of the verification information:
  - Forecast developer?
  - Forecaster?
  - End user or decision making system?

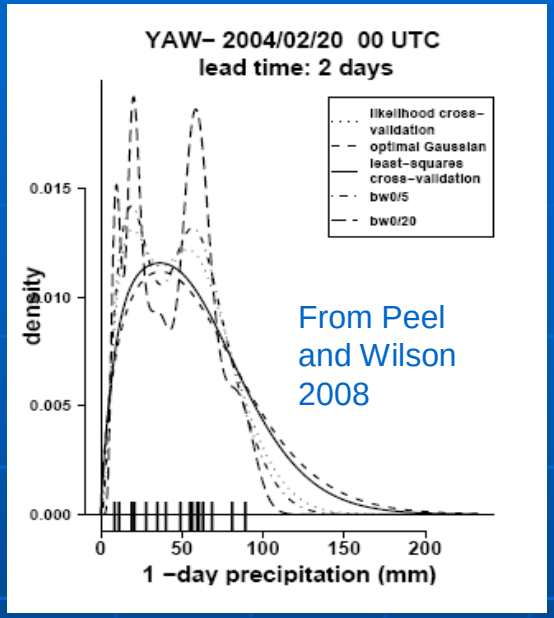




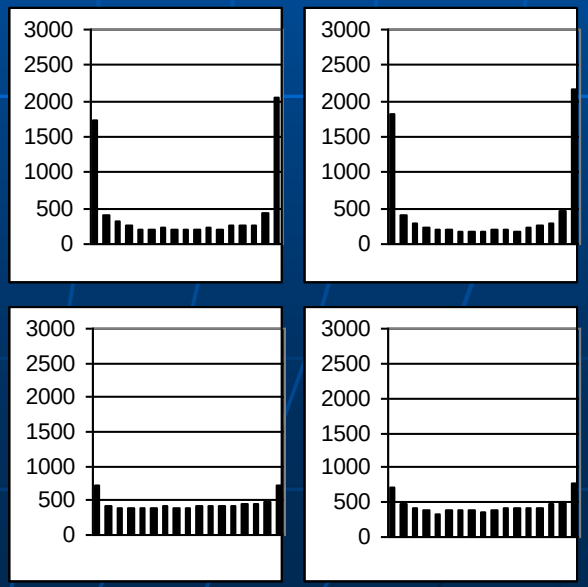


Evaluation of distributions

- *Ideal*: Interpret ensemble in terms of a distribution
  - Many approaches for distribution fitting
- Many measures:
  - CRPS, CRPSS
  - Rank Histogram (aka "Talagrand diagram")
  - Minimum spanning tree (e.g., for cyclone forecasts?)
  - Probability measure (Wilson 1999)
  - Ignorance score

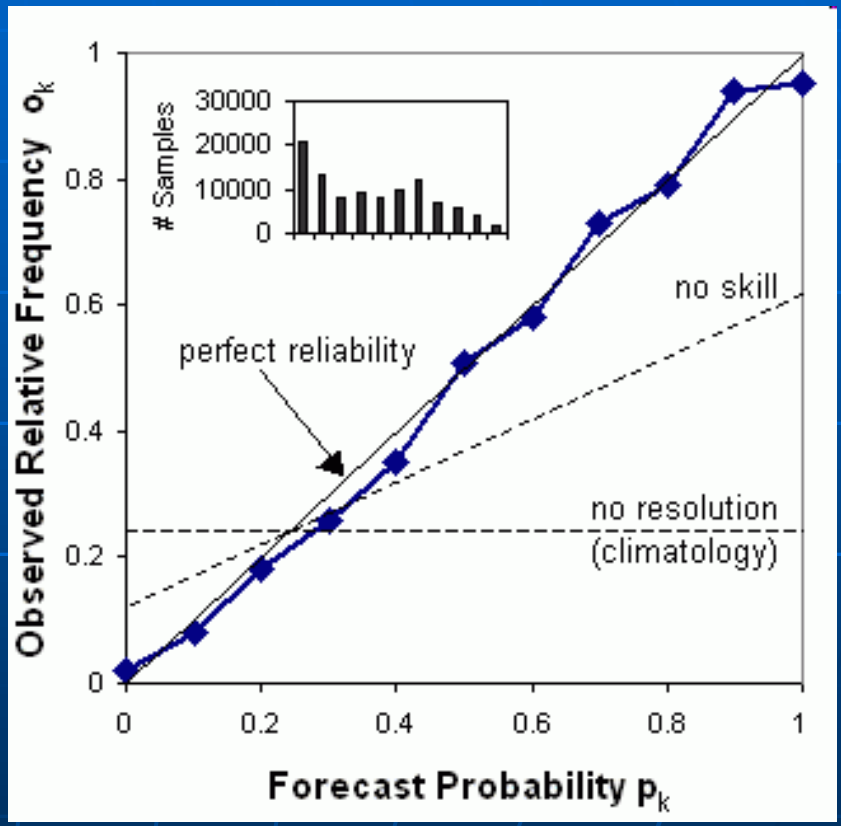


From Peel and Wilson 2008

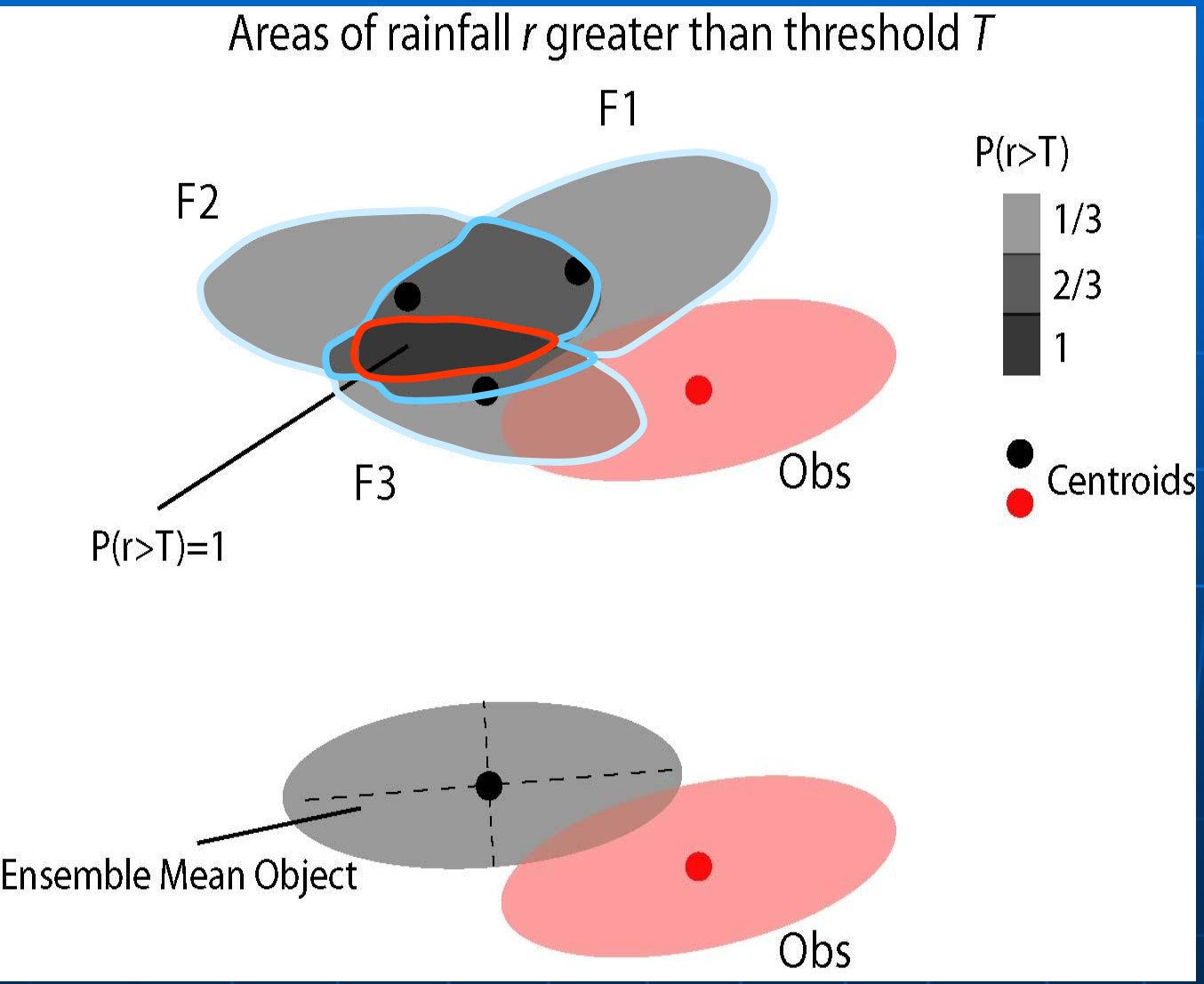


Evaluation of probabilities

- Based on selection of *meaningful threshold probabilities* or *events*
- Multiple measures provide information on accuracy, reliability, etc.
  - Brier score, ROC, reliability, discrimination
- Care is needed to appropriately select attributes of interest… and thresholds of interest



Treatment of Spatial Ensemble Forecasts



As probabilities:

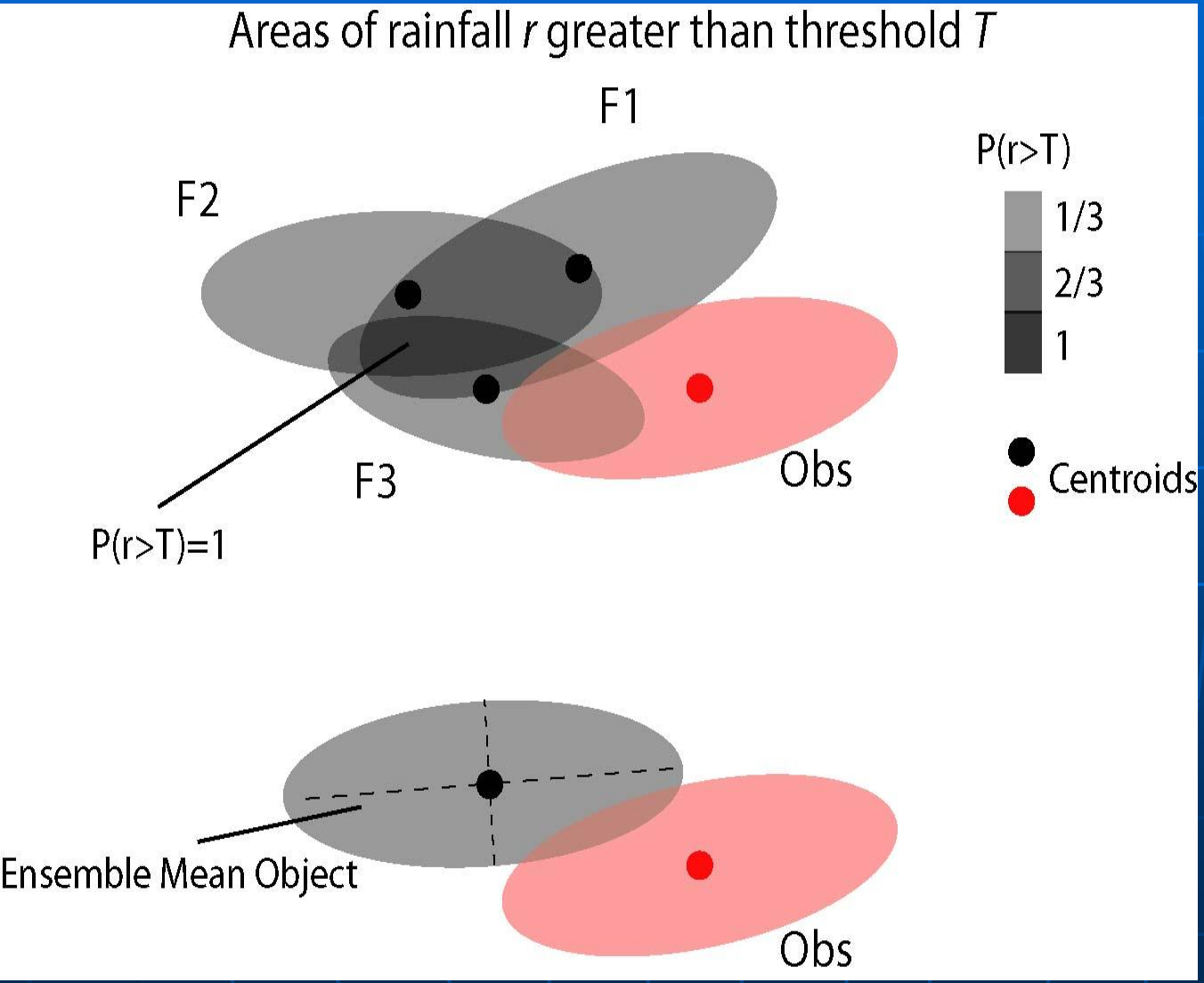
Areas do not have "shape" of convective storms

As mean:

Area is not equivalent to any of the underlying ensemble members

From C. Davis

Treatment of Spatial Ensemble Forecasts



Alternative:

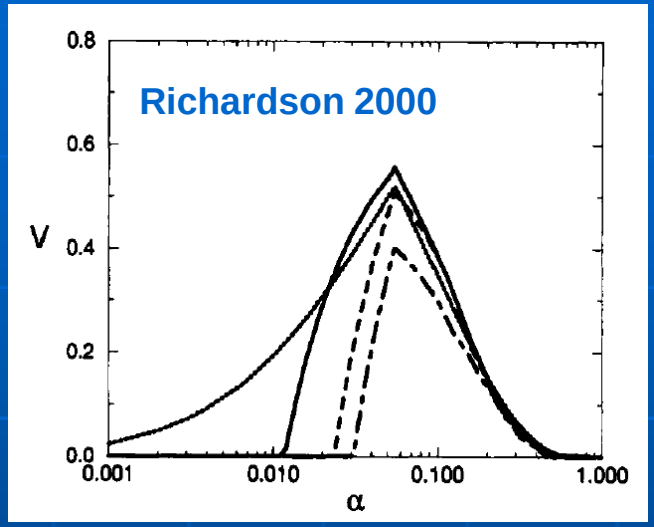
Consider ensembles of "*attributes*", such as…

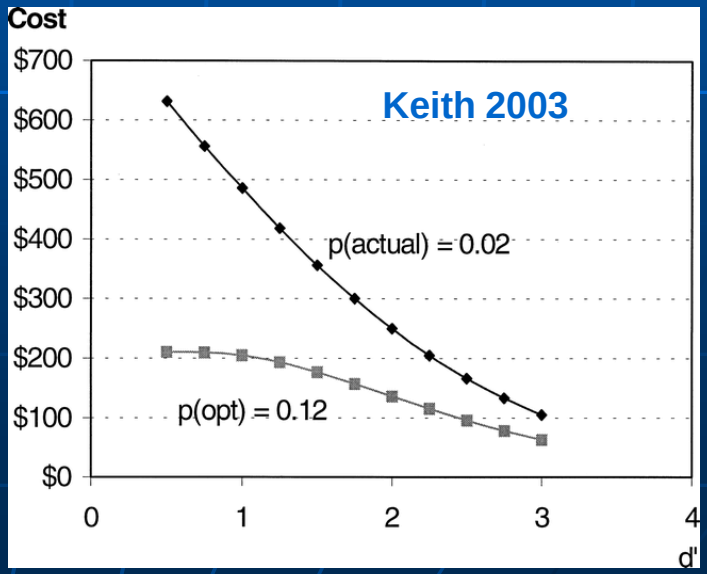
- Areas
- Intensities
- User-relevant measures (e.g., route blockages)

Collect and evaluate distributions of "*attribute*" errors

Links to value

- How do we measure value?
  - Richardson (2000): connect to ROC
  - Case study approach: e.g., Keith (2005)
  - Econometric approaches (e.g., Lazo 2009)
- Connection between *quality* and *value* is not transparent



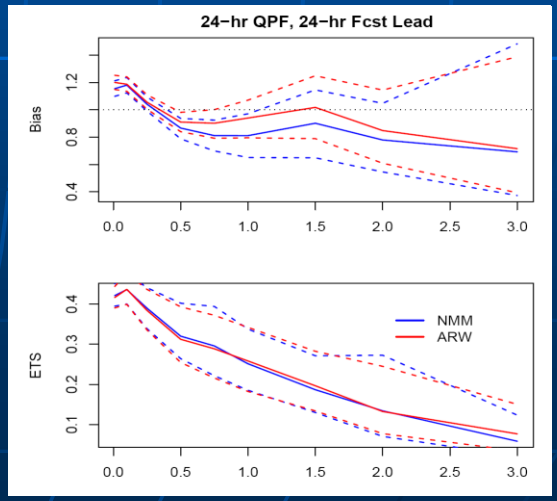


How do we meet the needs of diverse users?

- Provide information that is relevant to a wide spectrum of users
  - *Ex*: Multiple (user-selectable) thresholds
- Evaluate a ***wide variety*** of forecast attributes
- Focus on sensible weather elements
- Utilize diagnostic techniques
  - *Ex*: Distributions of statistics rather than (or in addition to) summary scores
- Provide access to "raw" forecasts and observations
- *Ideally* – strong interaction with users
  - Understand spectrum of applications of verification information
  - Engage users in discussions of use and value of forecasts
  - Need to work with social scientists to understand information and communication needs

Additional factors to consider…

- Sample size
  - Dimensionality of probabilistic verification typically requires many cases for robust evaluation
- Memory overload…
  - Ensembles require a lot of storage and a lot of memory to process…
- Uncertainty in verification measures…
  - MUST be shown
  - *How else do we make meaningful comparisons and measure progress in forecast development and improvement??? Or give honest appraisals of performance?*



Summary

Appropriate *measurement* and *reporting* of the quality of ensemble forecasts requires

- Understanding the integral nature of verification
  - Measure and present forecast quality at all stages
- Consideration of the spectrum of forecast users and their interests and requirements

Summary (cont.)

Appropriate *measurement* and *reporting* of the quality of ensemble forecasts requires (cont…)

- Provision of diagnostic information that
  - Appropriately reflects the form and characteristics of the forecasts and observations (e.g., distribution; multiple probability thresholds)
  - Answers a wide variety of questions of interest
  - Measures a meaningful breadth of forecast performance attributes
- Easy access to this information for the entire community

  (Don't keep this valuable information to ourselves)

  Contributions should be made by the entire community / enterprise

Recommendation 3.15: NWS should expand its verification systems for ensemble and other forecasts and make more explicit its choice of verification measures and rationale for those choices. Diagnostic and new verification approaches should be employed, and the verification should incorporate statistical standards such as stratification into homogeneous subgroups and estimation of uncertainty in verification measures. Verification information should be kept up to date and be easily accessible through the Web.

Recommendation 6. NWS should expand verification of its uncertainty products and make this information easily available to all users in near real time. A variety of verification measures and approaches (measuring multiple aspects of forecast quality that are relevant for users) should be used to appropriately represent the complexity and dimensionality of the verification problem. Verification statistics should be computed for meaningful subsets of the forecasts (e.g., by season, region) and should be presented in formats that are understandable by forecast users. Archival verification information on probabilistic forecasts, including model-generated and objectively generated forecasts and verifying observations, should be accessible so users can produce their own evaluation of the forecasts.