User-Oriented Forecast Verification
Some Issues

Laurence J. Wilson

MSC Dorval Quebec

Evaluation of forecasts

- Murphy's "goodness"
  - CONSISTENCY: forecasts agree with forecaster's true belief about the future weather [*strictly proper*]
  - QUALITY: correspondence between observations and forecasts [*verification*]
  - VALUE: increase or decrease in economic or other kind of value to someone as a result of using the forecast [*decision theory*]

Outline:

-Principles of verification

-Properness

-Verification issues

-Honesty in verification

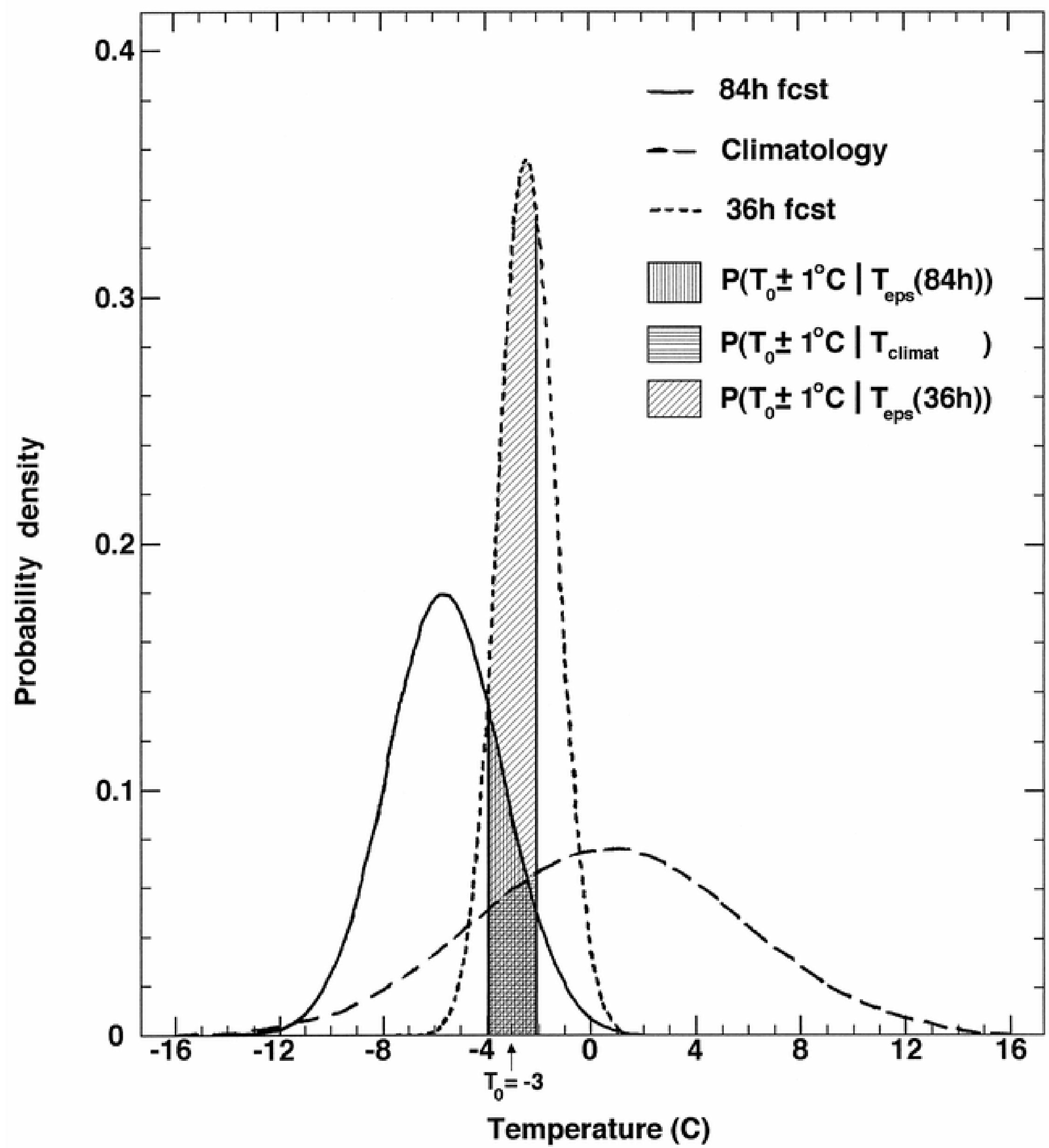
Principles of (Objective) Verification

- Verification activity has value only if the information generated leads to a decision about the forecast or system being verified
  - User of the information must be identified
  - Purpose of the verification must be known in advance: **Pose the verification question**
- No single verification measure provides complete information about the quality of a forecast product.
- Forecast must be stated in such a way that it can be verified
  - "chance" of showers
  - What does that gridpoint value really mean?
- Except for specific validation studies, verification should be carried out independently of the issuer of the product.

“Proper” verification

- Use of proper measures necessary to clearly separate forecaster and user role.
- Proper measures:
  - Most quadratic scoring rules, MAE, CRPS
- “Nearly” proper:
  - Skill scores
- Measures which weight the events of the verification sample may be improper
- Problem of designing measures which are both easy to understand AND proper.

Example



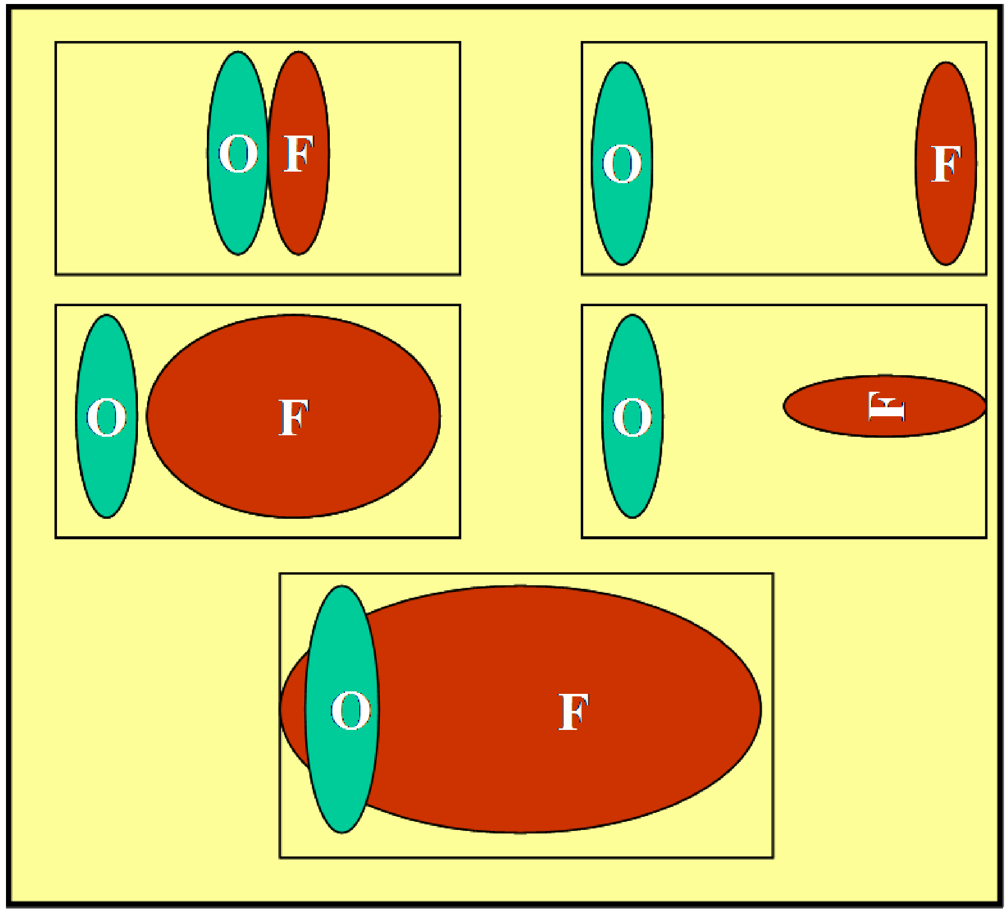
84h fcst
Climatology
36h fcst
P(T0± 1°C | Teps(84h))
P(T0± 1°C | Tclimat)
P(T0± 1°C | Teps(36h))
Probability density
0.4
0.3
0.2
0.1
0
-16
-12
-8
-4
0
4
8
12
16
T0= -3
Temperature (C)

Goals of Verification

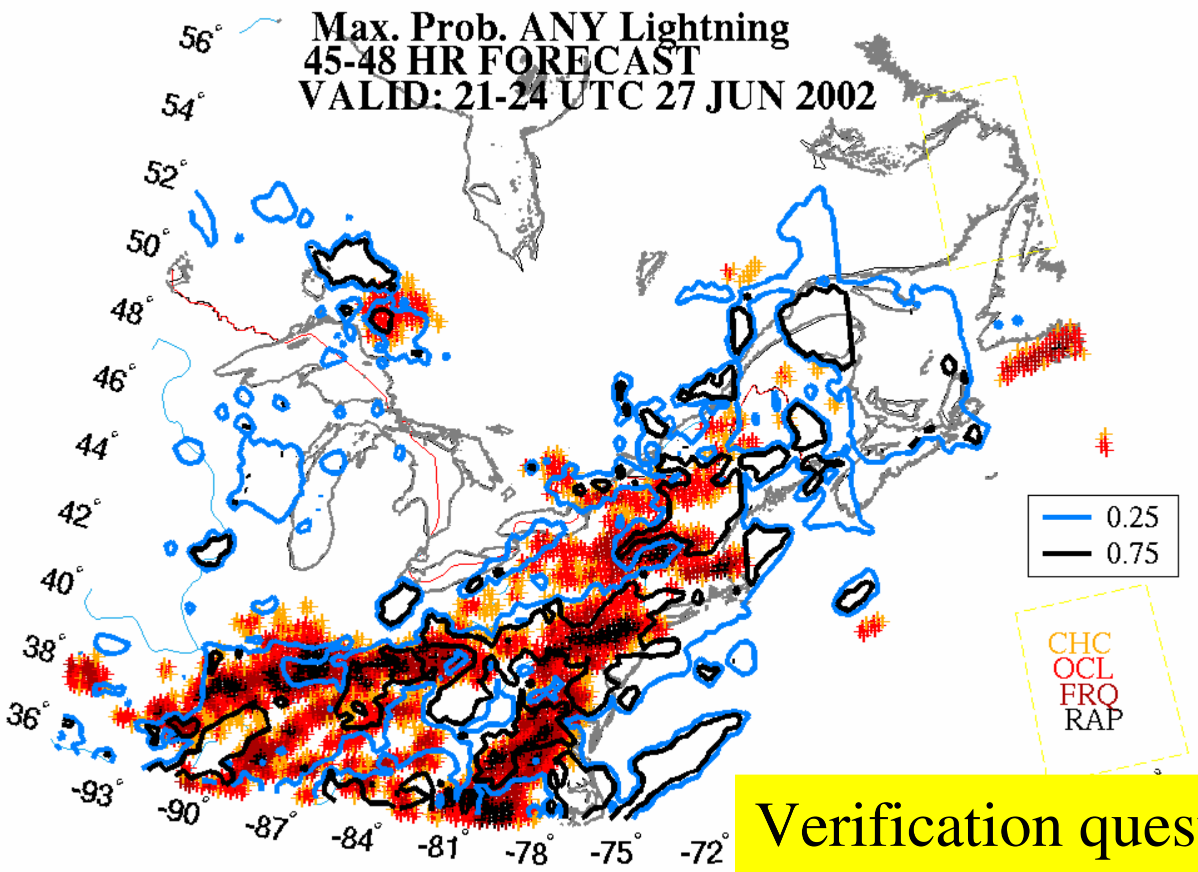
- Administrative ("Summary Verification")
  - Justify cost of provision of weather services
  - Justify additional or new equipment
  - Monitor the quality of forecasts and track changes
- Scientific ("Diagnostic Verification")
  - To identify the strengths and weaknesses of a forecast product in sufficient detail that actions can be specified that will lead to improvements in the product, ie to provide information to direct R&D.

COMMENT: Essential difference is the degree of summarizing over the dataset used and the depth of investigation of the data; there is overlap in the measures used – all objective verification uses "measures"

Object-oriented verification



Model Verification- Issues



- Resolution – scale separation
- Spatial verification – object-oriented

Verification question: What is the maximum resolution at which this forecast has skill?

Verification research needs (admittedly thinking of the modeling/forecasting community as user)

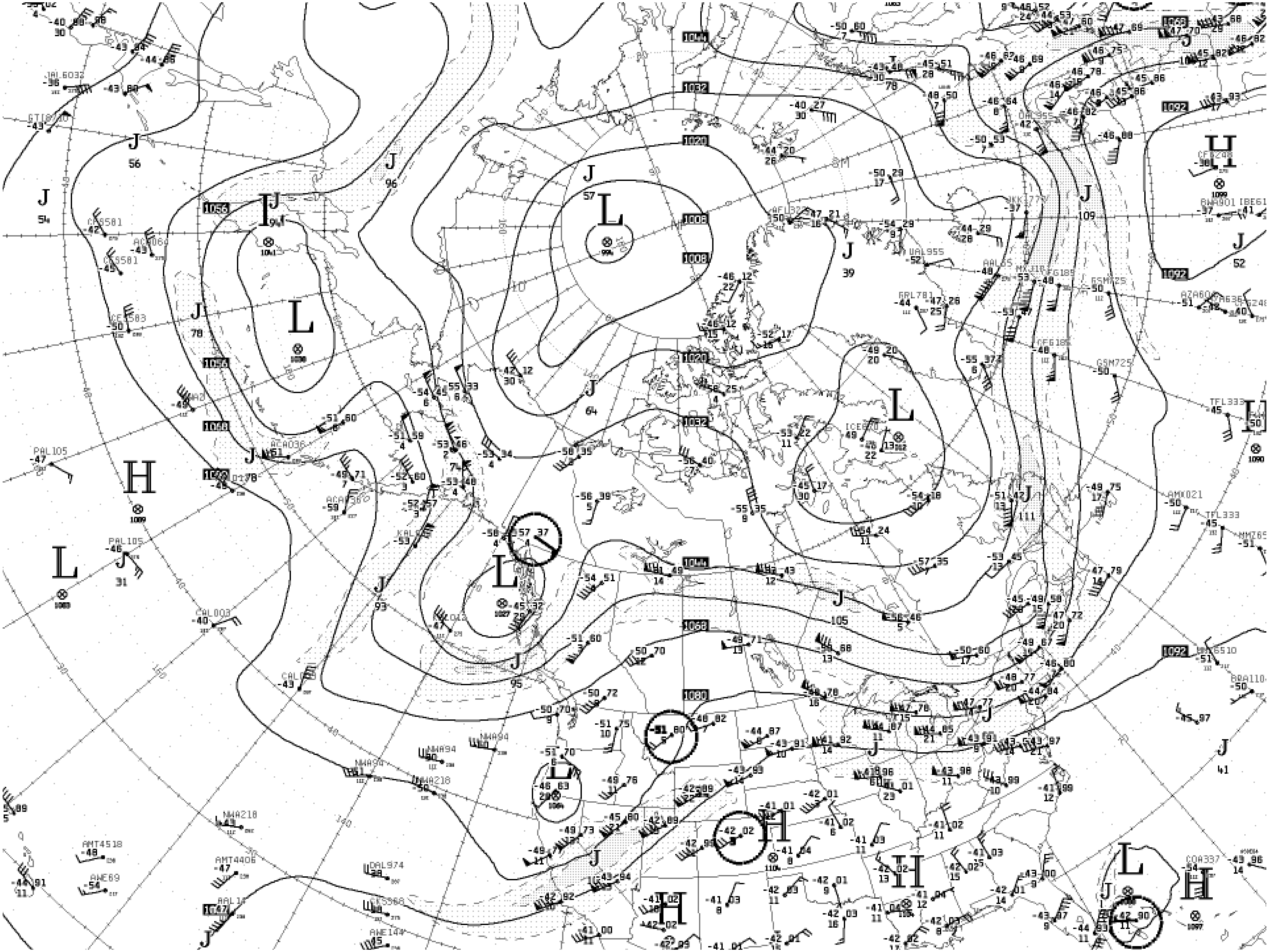
- Scale separation – verification conditional on scale component of forecast
- Object-oriented methods
- Verification of probability distributions (ensembles)
- Revisit "properness"
- Verification of high-spatial resolution forecasts
- Confidence limits on verification measures

User-oriented verification requires…

- Surface weather variables (not 500 mb)
- Using unprocessed quality controlled in situ observations if possible (completely independent)
- Selecting measures (or developing measures) which are sensitive to the aspect of forecast quality the user wants to know about
- (usually) pointwise rather than spatially-oriented
- More listening (to users) and less talking (about how good we think our verification is for them)
- Honesty in verification.

Tricks of the trade

- "How can I get a better (higher) number?"
  - Remove the bias before calculating scores (works really well for quadratic scoring rules) and don't tell anyone.
  - Claim that the model predicts grid box averages, even if it doesn't . Make the boxes as large as possible.
  - Never use observation data. It only contains a lot of "noise". As an alternative,:
    - Verify against an analysis that uses the model being verified as a trial field. Works best in data-sparse areas
    - Use a shorter range forecast from the model being verified and call it observation data.
    - If you must use observations, make sure they are quality-controlled against forecasts from the model you are verifying. Excellent way of getting rid of those troublesome extreme values in the dataset.
  - Design a new or modified score. Don't be bothered by restrictions such as strictly properness. Then the values can be as high as desired.
  - Stratify the verification data using posteriori rules. One can always get rid of pathological cases that bring down the average.
  - When comparing the performance of your model against others, make sure it is your analysis that is used as the verifying one.
  - Always insist on doing the verification of your own products….
- Remember, you already know you have a good product. The goal of the verification is to show it "objectively"



1044
1032
1020
1008
1056
1068
1080
1092

Concluding questions

- What should verification methodology research emphasize in THORPEX – SERA
  - Methods to diagnose model improvements resulting from other THORPEX projects?
  - Methods designed for specific users or user groups in support of SERA?
  - Both?
- Are we verification people there to tell the rest of THORPEX how to do verification to best support SERA?

Model Verification – Scale separation

- Mesoscale verification: Separating model errors due to resolution from other errors.
- IF high resolution spatial data is available:
  - Scale separation, wavelets or other method (Mike Baldwin)
  - Repeat verification on same dataset at different scales to get performance curve
  - Data combination: Use high resolution data to “inform” statistical estimates such as grid box averages.

Spatial Verification

- Object-oriented methods
- The calculation of displacement, size errors for specific objects (e.g. rain areas, fronts)
  - Hoffman, 1995; Ebert and McBride 2000 CRA method
    - Decomposition of errors into location, shape, size components
  - Others (Mike Baldwin)
  - Problem always is the matching of the forecast and observed object.

Verification of probability distributions

- Problem:
  - Comparison of distribution with a single outcome
- Solutions:
  - Verify density in vicinity of observation (Wilson, Burrows and Lanzinger, 1999)
  - Verify cdf against observation represented as cdf (CRPS, Hersbach 2000)
  - Extract probabilities from distribution and verify as probability forecasts (sample several thresholds)
  - Compare parameters of pooled distribution with sample climatology (Talagrand diagram)

References

- Wilks, D.S., 1995: Statistical methods in the atmospheric sciences. Academic Press. Chapter 7.
- Stanski, H.R., L.J. Wilson, and W.R. Burrows, 1990: Survey of common verification methods in meteorology. WMO WWW Technical Report No. 8. Also available on the web – see below.
- Jolliffe, I.T. and D.B. Stephenson, 2003: Forecast verification: A practitioner's guide in atmospheric science. Wiley.
- Murphy, A.H. and R.L. Winkler, 1987: A general framework for forecast verification. Mon. Wea. Rev. 115, 1330-1338.
- Murphy, A.H., 1993: What is a good forecast? An essay on the nature of goodness in weather forecasting. Wea. Forecasting 8, 281-293.

Web Page of the Joint WGNE/WWRP Verification Working Group:

http://www.bom.gov.au/bmrc/wefor/staff/eee/verif/verif_web_page.html

Has lots of links, and lots of information. (and is changing all the time)

ANNOUNCEMENT: This group is organizing a workshop on Verification methods in or near Montreal, Quebec, September 13 to 17, 2004.

Measuring value

- The cost-loss decision model
  - focus on maximizing gain or loss-avoidance
  - requires objective cost information from user
  - user specific, difficult to generalize
  - economic value to weather-sensitive operation only.
  - easy to evaluate relative value
- Contingent-valuation method
  - focuses on demand for service and "willingness to pay"
  - requires surveys of users to determine variations in demand as function of variations in price and/or quality of service
  - less user-specific; a larger crossection of users/industries can be evaluated in one study
  - measures in terms of perception rather than actual accuracy.
  - e.g. evaluation of ATADs, Rollins and Shaykewich, Met Apps Mar. 03