Forecast Evaluation and User-Focused Verification

Barbara Brown

Joint Numerical Testbed Program
Research Applications Laboratory
NCAR Boulder, Colorado USA

Sea-Ice Prediction Workshop
14 May 2014
Why forecast verification?

- Monitor performance
- Improve forecasts
- Communicate meaningful information to users
  - Requires identifying users’ information needs

Hence we need approaches that can do all of these things...

Different approaches for
- different purposes
- different types of forecasts
Tailoring verification approaches

Different types of forecasts
- Forecast “element” characteristics
  - Continuous (e.g., RMSE)
  - Categorical (e.g., Yes/No; POD, FAR)
  - Probabilistic
- Temporal characteristics
  - Time series?
- Spatial attributes
  - Gridded vs. Point
  - Spatial approaches

Different purposes
- Monitoring
  - Use basic easy-to-understand metrics
- Forecast improvement
  - Diagnostic approaches
- Users
  - Diagnostic
  - User-relevant
Identifying users’ needs

- Defining events:
  - What elements are needed? Time and space scales?
- What are the important decisions that are made relative to the events?
- What aspects are important?
  - Timing? Spatial location? Intensity?
- How do we measure the “quality of these aspects?"

Example events
- Decadal ice extent (building ships)
- Spatial extent of ice on a particular date (e.g., Sep 1) (seasonal prediction)
- Ice extent on specific dates and particular locations (ship movements)

Choices of events and metrics impact model optimization
Community Tools for Forecast Evaluation

- Traditional and new tools
- Initial version released in 2008
- Includes
  - Traditional approaches
  - Spatial methods (MODE, Scale, Neighborhood)
  - Confidence Intervals
  - Ensemble methods
- Supported to the community
  - More than 2,400 users (50% university)
  - Regular tutorials
  - Email help

Spatial distribution of Gilbert Skill Score

http://www.dtcenter.org/met/users/
Traditional spatial verification

- Requires an exact match between forecasts and observations at every grid point
  - Problem of "double penalty" - event predicted where it did not occur, no event predicted where it did occur
  - Traditional scores do not say very much about the source or nature of the errors

Hi res forecast
- RMS ~ 4.7
- POD=0, FAR=1
- TS=0

Low res forecast
- RMS ~ 2.7
- POD~1, FAR~0.7
- TS~0.3
Impacts of spatial variability

- Traditional approaches ignore spatial structure in the forecasts
  - Spatial correlations
- Small errors lead to poor scores (squared errors... smooth forecasts are rewarded)
- Methods for evaluation are not diagnostic
- Spatial methods can identify particular features of interest to evaluate

Grid-to-grid results:
POD = 0.40
FAR = 0.56
CSI = 0.27
(Poor Scores)
New Spatial Verification Approaches

Neighborhood
Successive smoothing of forecasts/obs
Gives credit to "close" forecasts

Scale separation
Measure scale-dependent error

Field deformation
Measure distortion and displacement (phase error) for whole field

Object- and feature-based
Evaluate attributes of identifiable features

How should the forecast be adjusted to make the best match with the observed field?

http://www.ral.ucar.edu/projects/icp/
Method for Object-based Diagnostic Evaluation (MODE)

Traditional verification results: 
*Forecast has very little skill*

MODE quantitative results:
- Most forecast areas too large
- Forecast areas slightly displaced
- Median and extreme intensities too large
- BUT – overall – forecast is pretty good
Applications to sea-ice and polar prediction problems

- Many tools exist for evaluation of time series (e.g., in MET)

- New spatial methods may be beneficial for evaluation of sea ice and other polar predictions to provide
  - Diagnostic information
  - More specific information tailored to evaluate meaningful events for users

From Arbetter 2012
Resources

- Model Evaluation Tools
- WMO verification Working Group
  - Connected to WWRP, WGNE, PPP, S2S, HIW
  - web page
- R verification package
- Verification discussion group

http://www.dtcenter.org/met/users/

Object/Feature-based

**Goals**: Measure and compare (user-) relevant features in the forecast and observed fields

**Examples**:  
- Contiguous Rain Area (CRA)  
- Method for Object-based Diagnostic Evaluation (MODE)  
- Procrustes  
- Cluster analysis  
- Structure Amplitude and Location (SAL)  
- Composite  
- Gaussian mixtures
Neighborhood methods

Goal: Examine forecast performance in a region; don’t require exact matches

- Also called “fuzzy” verification
- Example: Upscaling
  - Put observations and/or forecast on coarser grid
  - Calculate traditional metrics
- Provide information about scales where the forecasts have skill
- Examples: Roberts and Lean (2008) – Fractions Skill Score; Ebert (2008); Atger (2001); Marsigli et al. (2006)

From Mittermaier 2008
Scale separation methods

- **Goal:**
  Examine performance as a function of spatial scale

- **Examples:**
  - **Power spectra**
    - Does it look real?
    - Harris et al. (2001)
  - **Intensity-scale**
    - Casati et al. (2004)
  - **Multi-scale variability** (Zapeda-Arce et al. 2000; Harris et al. 2001; Mittermaier 2006)
  - **Variogram** (Marzban and Sandgathe 2009)

From Harris et al. 2001
**Field deformation**

**Goal:** Examine how much a forecast field needs to be transformed in order to match the observed field

**Examples:**
- Forecast Quality Index (*Venugopal et al.* 2005)
- Optical Flow (*Marzban et al.* 2009)

From *Keil and Craig* 2008