Verification Methods for High Resolution Model Forecasts

Barbara Brown (bgb@ucar.edu)

NCAR, Boulder, Colorado

Collaborators: Randy Bullock, John Halley Gotway, Chris Davis, David Ahijevych, Eric Gilleland, Beth Ebert

Hurricane Diagnostics and Verification Workshop
May 2009
Goals

Describe new approaches for evaluation of high-resolution spatial (gridded) forecasts

• Potentially useful for evaluating hurricane rainfall (and wind) patterns

The aim of these methods is to

• Provide more meaningful information about performance than traditional approaches
• Overcome some of the insensitivity of traditional approaches
Relevance for tropical cyclones

- Meaningful evaluation of precipitation evolution for land-falling hurricanes
  - Short-term predictions (e.g., extreme events; precipitation distribution; flooding)
  - Total storm precipitation predictions
- Meaningful comparisons of high- and low-resolution models
  - Low resolution tends to “win” using traditional measures even if pattern is not as good...
- Evaluation of oceanic precipitation using satellite
- Evaluation of wind pattern forecasts
- Other??
Hurricane Rita precipitation

Merged With Frontal Wave

Hurricane Rita
September 24-25, 2005
2977 sites
1
3
5
7
10
15

Maximum: 16.00"
Bunkie, LA

Track

5+
5+
7+
10+
15+
15+
5+
Hurricane precipitation verification

Swath-based precip verification approach (Marchok et al. 2007)

- QPF pattern matching
- Mean, volume, and distribution of rain values
- Production of extreme amounts
Hurricane Rita precipitation

Merged With Frontal Wave

Hurricane Rita
September 20-25, 2005
2977 sites

Track

Maximum: 16.00"
Bunkie, LA

9/20  9/21  9/22

9/23  9/24

9/25  9/26
Traditional approach

Consider gridded forecasts and observations of precipitation

Which is better?
Scores for Examples 1-4:
Correlation Coefficient = -0.02
Probability of Detection = 0.00
False Alarm Ratio = 1.00
Hanssen-Kuipers = -0.03
Gilbert Skill Score (ETS) = -0.01

Scores for Example 5:
Correlation Coefficient = 0.2
Probability of Detection = 0.88
False Alarm Ratio = 0.89
Hanssen-Kuipers = 0.69
Gilbert Skill Score (ETS) = 0.08

Forecast 5 is “Best”
Traditional approach

Some problems with the traditional approach:

(1) **Non-diagnostic** – doesn’t tell us what was wrong with the forecast – or what was right

(2) **Ultra-sensitive** to small errors in simulation of localized phenomena
Spatial forecasts

Weather variables (e.g., precipitation, wind fields) defined over spatial domains have **coherent structure and features**

**Spatial verification techniques aim to:**

- Account for
  - Uncertainties in timing and location
  - Spatial structure
- Provide information that is
  - Able to characterize error in physical terms
  - Diagnostic
  - Meaningful to forecast users
Spatial verification approaches

Filtering
1. Neighborhood
2. Scale separation

Displacement
3. Feature-based
4. Field deformation
New spatial verification approaches

- Field deformation approaches
  - Measure distortion and displacement (phase error) for whole field
  - *How should the forecast be adjusted to make the best match with the observed field?*

- Scale decomposition methods
  - Measure scale-dependent error

- Neighborhood verification methods
  - Give credit to "close" forecasts

- Object- and feature-based methods
  - Evaluate attributes of identifiable features

Keil and Craig, 2008
Intensity-scale method
Casati et al. (2004)

Evaluate forecast skill as a function of the precipitation intensity and the spatial scale of the error.
Scale → wavelet decomposition of binary error

\[ E_u = \sum_{l=1}^{L} E_{u,l} \]
\[ MSE_u = \sum_{l=1}^{L} MSE_{u,l} \]

From Ebert 2008
MSE skill score

\[
SS_{u,l} = \frac{MSE_{u,l} - MSE_{u,l,\text{random}}}{MSE_{u,l,\text{best}} - MSE_{u,l,\text{random}}} = 1 - \frac{MSE_{u,l}}{2\epsilon (1 - \epsilon) L}
\]

Sample climatology (base rate)

From Ebert 2008
Neighborhood verification

- Also called “fuzzy” verification
- Upscaling
  - Put observations and/or forecast on coarser grid
  - Calculate traditional metrics
- Provide information about scales where the forecasts have skill
Neighborhood methods

Fractional Skill Score (FSS)

From Mittermaier 2008
Feature-based verification

- Composite approach (Nachamkin)
- Contiguous rain area approach (CRA; Ebert and McBride, 2000; Gallus and others)
- Error components
  - displacement
  - volume
  - pattern
Spatial verification method: MODE

- **MODE**: Method for Object-based Diagnostic Evaluation

- **Goals**
  - Mimic how a human would identify storms and evaluate forecasts
  - Measure forecast “attributes” that are of interest to users

- **Steps**
  - **Identify objects**
  - **Measure attributes**
  - **Match forecast attributes**
  - **Measure differences in attributes**

**Example: Precipitation; 1 Jun 2005**

*User inputs:* Object identification; attribute selection and weighting
Object-based example: 1 June 2005

- **MODE quantitative results indicate**
  - Most forecast areas too large
  - Forecast areas slightly displaced
  - Median and extreme intensities too large
  - BUT – overall – forecast is pretty good

- **In contrast:**
  - POD = 0.40
  - FAR = 0.56
  - CSI = 0.27
MODE “Interest” measures overall ability of forecasts to match obs

Interest values provide more intuitive estimates of performance than the traditional measure (ETS)

But – even for spatial methods, Single measures don’t tell the whole story!
Spatial Method Intercomparison Project (ICP)

- **Goal**: Compare the various approaches using the same datasets *(real, geometric, known errors)*

- Includes all of the methods described here; international participants

- Collection of papers in preparation *(Weather and Forecasting)*

What do the new methods measure?

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Traditional</th>
<th>Feature-based</th>
<th>Neighborhood</th>
<th>Scale</th>
<th>Field Deformation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Perf at different scales</td>
<td>Indirectly</td>
<td>Indirectly</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Location errors</td>
<td>No</td>
<td>Yes</td>
<td>Indirectly</td>
<td>Indirectly</td>
<td>Yes</td>
</tr>
<tr>
<td>Intensity errors</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Structure errors</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Hits, etc.</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Indirectly</td>
<td>Yes</td>
</tr>
</tbody>
</table>
Applicability to ensemble forecasts

Areas of rainfall $r$ greater than threshold $T$

$P(r>T)$

- 1/3
- 2/3
- 1

F1
F2
F3

$P(r>T) = 1$

Obs

Centroids

Ensemble Mean Object

From C. Davis
Statistical inference

Confidence intervals are required to provide

- Meaningful evaluations of individual model performance
- Meaningful comparisons of model performance
Method availability

- Many methods available as part of the Model Evaluation Tools (MET)
  - MODE
  - Neighborhood
  - Intensity-scale
- MET is freely available
  - Strong user support
- Software for some others is available on the intercomparison website or from original developers

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

- New spatial methods provide great opportunities for more meaningful evaluation of precipitation forecasts – and other forecast fields
  - Feed back into forecast development
  - Provide information to users

- Each method is useful for particular types of situations and for answering particular types of questions
Topics for discussion

- Consider how new methods may be beneficial (and adaptable) for high-res NWP for hurricanes
  - Can these methods help?
  - How do they need to be adapted/altered?
  - Would they be useful for other fields (e.g., winds)?
  - Are other kinds of new methods needed?

- Use of aircraft observations – incomplete grids (reflectivity)

- Methods for evaluation of genesis? A global problem... Need to consider false alarms as well as misses

- Use of confidence intervals

From Bender et al. 2007