Methods for verifying spatial forecasts

Beth Ebert

Centre for Australian Weather and Climate Research (CAWCR)
Bureau of Meteorology, Melbourne, Australia

Acknowledgements: Barb Brown, Barbara Casati, Marion Mittermaier

4th Int'l Verification Methods Workshop, Helsinki, 4-6 June 2009
Spatial forecasts are made at many scales.
Visual ("eyeball") verification

Visually compare maps of forecast and observations

**Advantage:** "A picture tells a thousand words…"

**Disadvantages:** Labor intensive, not quantitative, subjective
Matching forecasts and observations

- Point-to-grid and grid-to-point

- Matching approach can impact the results of the verification
Matching forecasts and observations

- Grid to grid approach
  - Overlay forecast and observed grids
  - Match each forecast and observation
Traditional verification approaches

Compute statistics on forecast-observation pairs

- Continuous values (e.g., precipitation amount, temperature, NWP variables):
  - mean error, MSE, RMSE, correlation
  - anomaly correlation, S1 score

- Categorical values (e.g., precipitation occurrence):
  - Contingency table statistics (POD, FAR, Heidke skill score, equitable threat score, Hanssen-Kuipers statistic…)

Traditional spatial verification using categorical scores

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Observed</th>
<th>Predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>hits</td>
<td>false alarms</td>
</tr>
<tr>
<td>no</td>
<td>misses</td>
<td>correct negatives</td>
</tr>
</tbody>
</table>

$FBI = \frac{\text{hits} + \text{false alarms}}{\text{hits} + \text{misses}}$

$POD = \frac{\text{hits}}{\text{hits} + \text{misses}}$

$FAR = \frac{\text{false alarms}}{\text{hits} + \text{false alarms}}$

$TS = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{false alarms}}$

$ETS = \frac{\text{hits} - \text{hits}_{\text{random}}}{\text{hits} + \text{misses} + \text{false alarms} - \text{hits}_{\text{random}}}$
PODy = 0.39, FAR = 0.63, CSI = 0.24
High vs. low resolution

Which forecast would you rather use?

Mesoscale model (5 km) 21 Mar 2004

Global model (100 km) 21 Mar 2004

Observed 24h rain

RMS=13.0 mm

RMS=4.6 mm
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
What’s missing?

- Traditional approaches provide overall measures of skill but…

- They provide minimal *diagnostic* information about the forecast:
  - What went wrong? What went right?
  - Does the forecast look realistic?
  - How can I improve this forecast?
  - How can I use it to make a decision?

- Best performance for *smooth* forecasts

- Some scores are insensitive to the *size* of the errors…
Spatial forecasts

Weather variables defined over spatial domains have coherent spatial structure and features

New spatial verification techniques aim to:

- account for field spatial structure
- provide information on error in physical terms
- account for uncertainties in location (and timing)
New spatial verification approaches

- Neighborhood (fuzzy) verification methods
  - give credit to "close" forecasts

- Scale decomposition methods
  - measure scale-dependent error

- Object-oriented methods
  - evaluate attributes of identifiable features

- Field verification
  - evaluate phase errors
Spatial Verification Intercomparison Project

Begun February 2007

The main goals of this project are to:

- Obtain an inventory of the methods that are available and their capabilities

- Identify methods that
  - may be useful in operational settings
  - could provide automated feedback into forecasting systems
  - are particularly useful for specific applications (e.g., model diagnostics, hydrology, aviation)

- Identify where there may be holes in our capabilities and more research and development is needed
Spatial Verification Intercomparison Project

- Test cases
- Results
- Papers
- Code
Neighborhood (fuzzy) verification methods
→ give credit to "close" forecasts
Neighborhood verification methods

- Don't require an exact match between forecasts and observations
  - Unpredictable scales
  - Uncertainty in observations

- Look in a space / time neighborhood around the point of interest
  - Evaluate using categorical, continuous, probabilistic scores / methods
Neighborhood verification methods

Treatment of forecast data within a window:

- Mean value (upscaling)
- Occurrence of event* somewhere in window
- Frequency of events in window $\rightarrow$ probability
- Distribution of values within window

May also look in a neighborhood of observations

*Event defined as a value exceeding a given threshold, for example, rain exceeding 1 mm/hr
Oldest neighborhood verification method - upscaling

- Average the forecast and observations to successively larger grid resolutions, then verify using the usual metrics:
  - Continuous statistics – mean error, RMSE, correlation coefficient, etc.
  - Categorical statistics – POD, FAR, FBI, TS, ETS, etc.
Fractions skill score
(Roberts and Lean, MWR, 2008)

- We want to know
  - How forecast skill varies with neighborhood size
  - The smallest neighborhood size that can be used to give sufficiently accurate forecasts
  - Does higher resolution NWP provide more accurate forecasts on scales of interest (e.g., river catchments)

Compare forecast fractions with observed fractions (radar) in a *probabilistic* way over different sized neighbourhoods

\[
\text{FSS} = 1 - \frac{1}{N} \sum_{i=1}^{N} \left( P_{\text{fcst}} - P_{\text{obs}} \right)^2
\]

- Fraction = 6/25 = 0.24

observed

- Fraction = 6/25 = 0.24

forecast
Fractions skill score
(Roberts and Lean, *MWR*, 2008)

\[ f_{o} = \text{domain obs fraction} \]

\[ \frac{0.5 + f_{o}}{2} \]

Perfect 1 skill

Useful scales

Too much smoothing

asymptotes to value that depends on the frequency bias (1 if no bias)

0.5 + \( f_{o} \)/2

Present output on these scales

uniform target skill

Spatial scale (length of neighbourhood squares)

4th Int'l Verification Methods Workshop, Helsinki, 4-6 June 2009
Spatial multi-event contingency table

- Experienced forecasters interpret output from a high resolution deterministic forecast in a *probabilistic* way

\[ \text{"high probability of some heavy rain near Sydney"}, \quad \textit{not} \quad \text{"62 mm of rain will fall in Sydney"} \]

- The deterministic forecast is mentally "calibrated" according to how "close" the forecast is to the place / time / magnitude of interest.

\[
\text{Very close} \rightarrow \text{high probability} \\
\text{Not very close} \rightarrow \text{low probability}
\]
Spatial multi-event contingency table

- Verify using the Relative Operating Characteristic (ROC)

Measures how well the forecast can separate events from non-events based on some decision threshold

Decision thresholds to vary:
- magnitude (ex: 1 mm h\(^{-1}\) to 20 mm h\(^{-1}\))
- distance from point of interest (ex: within 10 km, .... , within 100 km)
- timing (ex: within 1 h, ... , within 12 h)
- anything else that may be important in interpreting the forecast
Different neighborhood verification methods have different decision models for what makes a *useful forecast*

<table>
<thead>
<tr>
<th>Neighborhood method</th>
<th>Matching strategy*</th>
<th>Decision model for useful forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Upscaling</strong> (Zepeda-Arce et al. 2000; Weygandt et al. 2004)</td>
<td>NO-NF</td>
<td>Resembles obs when averaged to coarser scales</td>
</tr>
<tr>
<td><strong>Minimum coverage</strong> (Damrath 2004)</td>
<td>NO-NF</td>
<td>Predicts event over minimum fraction of region</td>
</tr>
<tr>
<td><strong>Fuzzy logic</strong> (Damrath 2004), joint probability (Ebert 2002)</td>
<td>NO-NF</td>
<td>More correct than incorrect</td>
</tr>
<tr>
<td><strong>Fractions skill score</strong> (Roberts and Lean 2008)</td>
<td>NO-NF</td>
<td>Similar frequency of forecast and observed events</td>
</tr>
<tr>
<td><strong>Area-related RMSE</strong> (Rezacova et al. 2006)</td>
<td>NO-NF</td>
<td>Similar intensity distribution as observed</td>
</tr>
<tr>
<td><strong>Pragmatic</strong> (Theis et al. 2005)</td>
<td>SO-NF</td>
<td>Can distinguish events and non-events</td>
</tr>
<tr>
<td><strong>CSRR</strong> (Germann and Zawadzki 2004)</td>
<td>SO-NF</td>
<td>High probability of matching observed value</td>
</tr>
<tr>
<td><strong>Multi-event contingency table</strong> (Atger 2001)</td>
<td>SO-NF</td>
<td>Predicts at least one event close to observed event</td>
</tr>
<tr>
<td><strong>Practically perfect hindcast</strong> (Brooks et al. 1998)</td>
<td>SO-NF</td>
<td>Resembles forecast based on perfect knowledge of observations</td>
</tr>
</tbody>
</table>

*NO-NF = neighborhood observation-neighborhood forecast, SO-NF = single observation-neighborhood forecast

Moving windows

For each combination of neighborhood size and intensity threshold, accumulate scores as windows are moved through the domain.
Multi-scale, multi-intensity approach

- Forecast performance depends on the scale and intensity of the event

![Heatmap showing forecast performance depending on spatial scale and intensity](image-url)
Example: Neighborhood verification of precipitation forecast over USA

1. How does the average forecast precipitation improve with increasing scale?
2. At which scales does the forecast rain distribution resemble the observed distribution?
3. How far away do we have to look to find at least one forecast value similar to the observed value?
1. How does the average forecast precipitation improve with increasing scale?

- Upscaling method
2. At which scales does the forecast rain distribution resemble the observed distribution?

- Fractions skill score
3. How far away do we have to look to find at least one forecast value similar to the observed value?

■ Multi-event contingency table

![Multi-event contingency table](image)

\[ \text{KSS} = \text{POD} - \text{POFD} \]
Scale separation methods

→ scale-dependent error
Intensity-scale method
Casati et al., *Met. Apps.*, 2004

Evaluate the forecast skill as a function of the *intensity* and the *spatial scale* of the error

![Precipitation analysis](image1)
![Precipitation forecast](image2)
Intensity threshold $\rightarrow$ binary images

Binary analysis

Binary forecast

$E_u = I_{Y^* > u} - I_{X > u}$

$u=1 \text{ mm/h}$

Binary error
Scale → wavelet decomposition of binary error

mean (1280 km)

Scale l=8 (640 km)
Scale l=7 (320 km)

Scale l=6 (160 km)

Scale l=5 (80 km)

Scale l=4 (40 km)

Scale l=3 (20 km)

Scale l=2 (10 km)

Scale l=1 (5 km)

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

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

Sample climatology (base rate)
Example: Intensity-scale verification of precipitation forecast over USA

1. Which spatial scales are well represented and which scales have error?
2. How does the skill depend on the precipitation intensity?
Intensity-scale results

1. Which spatial scales are well represented and which scales have error?
2. How does the skill depend on the precipitation intensity?
What is the difference between neighborhood and scale decomposition approaches?

- Neighborhood (fuzzy) verification methods
  - Get scale information by filtering out higher resolution scales

- Scale decomposition methods
  - Get scale information by isolating scales of interest
Object-oriented methods
→ evaluate attributes of features
Feature-based approach (CRA)
Ebert and McBride, *J. Hydrol.*, 2000

- Define entities using threshold (Contiguous Rain Areas)
- Horizontally translate the forecast until a *pattern matching* criterion is met:
  - minimum total squared error between forecast and observations
  - maximum correlation
  - maximum overlap
- The displacement is the vector difference between the original and final locations of the forecast.
CRA error decomposition

Total mean squared error (MSE)

\[ MSE_{total} = MSE_{displacement} + MSE_{volume} + MSE_{pattern} \]

The *displacement error* is the difference between the mean square error before and after translation

\[ MSE_{displacement} = MSE_{total} - MSE_{shifted} \]

The *volume error* is the bias in mean intensity

\[ MSE_{volume} = (\bar{F} - \bar{X})^2 \]

where \( \bar{F} \) and \( \bar{X} \) are the mean forecast and observed values after shifting.

The *pattern error*, computed as a residual, accounts for differences in the fine structure,

\[ MSE_{pattern} = MSE_{shifted} - MSE_{volume} \]
Example: CRA verification of precipitation forecast over USA

1. What is the location error of the forecast?
2. How do the forecast and observed rain areas compare? Average values? Maximum values?
3. How do the displacement, volume, and pattern errors contribute to the total error?
wrf2 fcst 20050601 hour 00-24

Analysis 20050601

CRA 20050601

Predicted rainfall (shifted)

Analyzed rainfall

wrf2 24h fcst 20050601  n=8423
(33.49°,−132.28°) to (37.77°,−96.00°)
Verif. grid=0.042°  CRA threshold=1.0 mm/h

<table>
<thead>
<tr>
<th></th>
<th>Analysed</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td># gridpoints ≥1 mm/h</td>
<td>3304</td>
<td>3597</td>
</tr>
<tr>
<td>Average rainrate (mm/h)</td>
<td>3.58</td>
<td>3.61</td>
</tr>
<tr>
<td>Maximum rain (mm/h)</td>
<td>119.63</td>
<td>39.12</td>
</tr>
<tr>
<td>Rain volume (km³)</td>
<td>0.51</td>
<td>0.52</td>
</tr>
</tbody>
</table>

Displacement (E,N) = [2.20°,1.92°] max corr matching

RMS error (mm/d)    | 12.81     | 10.24    |
Correlation coefficient | −0.167    | 0.305    |

Error Decomposition:
Displacement error 36.1%
Volume error 0.0%
Pattern error 62.9%
wrf2 fcst 20050601 hour 00-24

Analysis 20050601

CRA 20050601

Predicted rainfall (shifted)

Analyzed rainfall

wrf2 24h fcst 20050601  n=11007
(37.52°, -131.29°) to (45.29°, -94.65°)
Verif. grid=0.042°  CRA threshold=1.0 mm/h

<table>
<thead>
<tr>
<th></th>
<th>Analysed</th>
<th>Forecast</th>
</tr>
</thead>
<tbody>
<tr>
<td># gridpoints ≥1 mm/h</td>
<td>4540</td>
<td>5699</td>
</tr>
<tr>
<td>Average rainrate (mm/h)</td>
<td>1.52</td>
<td>2.68</td>
</tr>
<tr>
<td>Maximum rain (mm/h)</td>
<td>21.08</td>
<td>27.69</td>
</tr>
<tr>
<td>Rain volume (km³)</td>
<td>0.26</td>
<td>0.46</td>
</tr>
</tbody>
</table>

Displacement (E,N) = [0.52°, -0.84°]  max.corr matching

RMS error (mm/d)                  5.11     4.65
Correlation coefficient           -0.040   0.193

Error Decomposition:
Displacement error               18.7%
Volume error                     4.3%
Pattern error                    76.4%
Sensitivity to rain threshold

1 mm h\(^{-1}\) 5 mm h\(^{-1}\) 10 mm h\(^{-1}\)
MODE – Method for Object-based Diagnostic Evaluation
Davis et al., MWR, 2006

Two parameters:
1. Convolution radius
2. Threshold
MODE object matching/merging

Compare attributes:
- centroid location
- intensity distribution
- area
- orientation
- etc.

When objects not matched:
- false alarms
- missed events
- rain volume
- etc.

24h forecast of 1h rainfall on 1 June 2005
MODE methodology

Identification

Measure Attributes

Merging

Matching

Comparison

Summarize

Convolution – threshold process

Fuzzy Logic Approach
- Compare forecast and observed attributes
- Merge single objects into clusters
- Compute *interest values*
- Identify matched pairs

Accumulate and examine comparisons across many cases
Example: MODE verification of precipitation forecast over USA

1. What is the location error of the forecast?
2. How do the forecast and observed rain areas compare? Average values? Maximum values? Shape?
3. What is the overall quality of the forecast as measured by the median of the maximum object interest values?
MODE applied to our US rain example

<table>
<thead>
<tr>
<th>Issue Time:</th>
<th>May 31, 2005 00:00:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid Time:</td>
<td>Jun 1, 2005 00:00:00</td>
</tr>
<tr>
<td>Lead Time:</td>
<td>24 hours</td>
</tr>
<tr>
<td>Accum Time:</td>
<td>1 hours</td>
</tr>
<tr>
<td>Fuzzy Engine Weights</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Raw Threat:</th>
<th>0.00 in/100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mask Bad:</td>
<td>off</td>
</tr>
<tr>
<td>Conv Radius:</td>
<td>15 gs</td>
</tr>
<tr>
<td>Conv Thresh:</td>
<td>5.00 in/100</td>
</tr>
</tbody>
</table>

Displacement errors

1. 25 km
2. 23 km
3. 30 km
Sensitivity to rain threshold and convolution radius

(Note: This is not for the same case)
Structure-Amplitude-Location (SAL)

For a chosen domain and precipitation threshold, compute:

**Amplitude error**

\[ A = \frac{(D(R_{fcst}) - D(R_{obs}))}{0.5*(D(R_{fcst}) + D(R_{obs}))} \]

- \( D(\ldots) \) denotes the area-mean value (e.g., catchment)
- \( A \in [-2, \ldots, 0, \ldots, +2] \)

**Location error**

\[ L = \frac{|r(R_{fcst}) - r(R_{obs})|}{dist_{max}} \]

- \( r(\ldots) \) denotes the centre of mass of the precipitation field in the area
- \( L \in [0, \ldots, 1] \)

**Structure error**

\[ S = \frac{(V(R_{fcst}^*) - V(R_{obs}^*))}{0.5*(V(R_{fcst}^*) + V(R_{obs}^*))} \]

- \( V(\ldots) \) denotes the weighted volume average of all scaled precipitation objects in considered area, \( R^* = R / R_{max} \)
- \( S \in [-2, \ldots, 0, \ldots, +2] \)
Example: SAL verification of precipitation forecast over USA

1. Is the domain average precipitation correctly forecast?
2. Is the mean location of the precipitation distribution in the domain correctly forecast?
3. Does the forecast capture the typical structure of the precipitation field (e.g., large broad objects vs. small peaked objects)?
SAL verification results

1. Is the domain average precipitation correctly forecast?  \( A = 0.21 \)
2. Is the mean location of the precipitation distribution in the domain correctly forecast?  \( L = 0.06 \)
3. Does the forecast capture the typical structure of the precipitation field (e.g., large broad objects vs. small peaked objects)?  \( S = 0.46 \)  
   (perfect=0)
Field verification
→ evaluate phase errors
Displacement and Amplitude Score (DAS)

Keil and Craig, *WAF*, 2009

Combines distance and amplitude measures by matching forecast $\rightarrow$ observation & observation $\rightarrow$ forecast

- Pyramidal image matching (optical flow) to get vector displacement field $\rightarrow$ DIS
- Intensity errors for morphed field $\rightarrow$ AMP
- Displacement-amplitude score

$$DAS = \frac{DIS}{D_{\text{max}}} + \frac{AMP}{I_0}$$
Example: DAS verification of precipitation forecast over USA

1. How much must the forecast be distorted in order to match the observations?
2. After morphing how much amplitude error remains in the forecast?
3. What is the overall quality of the forecast as measured by the distortion and amplitude errors together?
DAS applied to our US forecast

1. How much must the forecast be distorted in order to match the observations?
2. After morphing how much amplitude error remains in the forecast?
3. What is the overall quality of the forecast as measured by the distortion and amplitude errors together?
Conclusions

- What method should you use for spatial verification?
  - Depends what question(s) you would like to address

- Many spatial verification approaches
  - Neighborhood (fuzzy) – credit for "close" forecasts
  - Scale decomposition – scale-dependent error
  - Object-oriented – attributes of features
  - Field verification – phase and amplitude errors
What method(s) could you use to verify

Wind forecast (sea breeze)

Neighborhood (fuzzy) – credit for "close" forecasts
Scale decomposition – scale-dependent error
Object-oriented – attributes of features
Field verification – phase and amplitude errors
What method(s) could you use to verify

Cloud forecast

- **Neighborhood (fuzzy)** – credit for "close" forecasts
- **Scale decomposition** – scale-dependent error
- **Object-oriented** – attributes of features
- **Field verification** – phase and amplitude errors
What method(s) could you use to verify

Mean sea level pressure forecast

- Neighborhood (fuzzy) – credit for "close" forecasts
- Scale decomposition – scale-dependent error
- Object-oriented – attributes of features
- Field verification – phase and amplitude errors
What method(s) could you use to verify

Tropical cyclone forecast

- Neighborhood (fuzzy) – credit for "close" forecasts
- Scale decomposition – scale-dependent error
- Object-oriented – attributes of features
- Field verification – phase and amplitude errors

Observed

3-day forecast
That's it!