Verifying NARCCAP Models for Severe-Storm Environments

Eric Gilleland
Weather Systems Assessment Program
Research Applications Laboratory,
Seminar at: Statistical Sciences and Operations Research,
Virginia Commonwealth University, Richmond, Virginia
3 March 2017

Co-authors: Christopher L. Williams, Melissa Bukovsky, Seth McGinnis, Barb Brown, Linda Mearns, and Caspar Ammann

Support for this work provided by the NSF via the Weather and Climate Impacts Assessment Science Program (http://www.assessment.ucar.edu) and Earth System Modeling (EaSM) Grant number AGS-1243030.
Study and Visit Opportunities
https://www2.ucar.edu/opportunities

But, also talk to Montse about STATMOS!

Photo by Everett Nychka
Severe Storm Environments

Convective Available Potential Energy

\[ \text{CAPE} \times \text{Shear} \ (\text{J kg}^{-1} \times \text{m s}^{-1}) \]

0 – 6 km vertical wind shear

\[ W_{\text{max}} \times \text{Shear} \ (\text{WmSh, m}^2 \text{ s}^{-2}) \]

Maximum updraft velocity \( W_{\text{max}} \ (\text{ms}^{-1}) = (2 \times \text{CAPE})^{1/2} \)
All are interpolated to be on the same grid, which is $\approx 0.5^\circ$

<table>
<thead>
<tr>
<th>All</th>
<th>NCEP reanalysis</th>
<th>Community Climate System Model</th>
<th>3rd Generation Coupled Global Climate Model</th>
<th>Hadley Centre Coupled Model, v. 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>abbreviation</td>
<td>NCEP</td>
<td>CCSM3</td>
<td>CGCM3</td>
<td>HadCM3</td>
</tr>
<tr>
<td>Canadian Regional Climate Model (CRCM)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Hadley Regional Model 3 (HRM3)</td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Pennsylvania State University/NCAR mesoscale model (MM5I)</td>
<td></td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Weather Research and Forecasting model (WRFG)</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

http://www.narccap.ucar.edu/
http://www.emc.ncep.noaa.gov/mmb/rreanl/
Lingo

WmSh: As before, but set to zero if CAPE < 100 J kg\(^{-1}\)
or 5 \leq \text{Shear} \leq 50 \text{ ms}^{-1}

q75: Univariate time series giving the upper quartile of
CAPE or WmSh over space at each time point.

High “field energy”: when q75 > its 90\(^{\text{th}}\) percentile over
time.

\(\kappa\): Frequency of CAPE \geq 1000 \text{ J kg}^{-1} \text{ conditioned on the presence of high field energy.}

\(\omega\): Frequency of WmSh \geq 225 \text{ m}^2\text{s}^{-2} \text{ conditioned on the presence of high field energy.}
\[ K \]
Spatial Forecast Verification

All identical measures!

- Traditional Verification does not provide diagnostic information
- Often favors coarser scale models
  - double penalty
  - aggregation of small-scale errors

Fig. 1 and Table 2 from Ahijevych et al. (2009, WAF, 24, 1485 – 1497)
Spatial Forecast Verification

List of papers:
http://www.ral.ucar.edu/projects/icp/references.html

- Numerous papers rapidly introduced new methods
  - image analysis
  - computer vision
  - shape analysis
  - spatial statistics
- ICP invoked to get a handle on the methods
  - geometric and real cases
  - precipitation over central United States
  - Most methods fall into one of 4 categories
  - MesoVICT continuation of ICP
    - complex terrain
    - More variables
    - Ensembles (obs and model)

Fig. 2 from G. et al. (2010, BAMS, 91 (10), 1365 – 1373)
Mean Error Distance

MED(A, B) is the average distance from points in the set B to points in the set A.

\[ MED(A, B) = \frac{\sum_x d(x, B \mid x \in A)}{N} \]

MED(A, B) is the average distance from points in the set A to points in the set B.

\[ MED(A, B) = \frac{\sum_x d(x, A \mid x \in B)}{N} \]

N is the size of the domain.

\[ N = 80 \]

Centroid distance = 80.
Baddeley’s $\Delta$ Metric

Distance maps for $A$ and $B$. Note dependence on location within the domain.
Baddeley’s $\Delta$ Metric

$$T = |d(x, A) - d(x, B)|$$

- $p = 1$ gives the arithmetic average of $T$
- $p = 2$ is the usual choice
- $p = \infty$ gives the max of $T$ (Hausdorff distance)

$\Delta$ is the $L_p$ norm of $T$

d($x, A$) and d($x, B$) are first transformed by a function $\omega$.

Usually, $\omega(x) = \max(x, \text{constant})$, but all results here use $\infty$ for the constant term.

$$\Delta(B, A) = \Delta(B, A) = \left[\sum_{x \in \text{Domain}} |d(x, A) - d(x, B)|^p\right]^{1/p} / N$$

$N$ is the size of the domain
Contrived Examples: Circles

All circles have radius = 20 grid squares

Domain size is 200 by 200

Touching the edge of the domain
## Contrived Examples: Circles

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>MED(A, B)</th>
<th>rank</th>
<th>MED(B, A)</th>
<th>rank</th>
<th>Δ(A, B)</th>
<th>rank</th>
<th>cent. dist.</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2</td>
<td>22</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td>29</td>
<td>2</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>62</td>
<td>4</td>
<td>62</td>
<td>3</td>
<td>57</td>
<td>6</td>
<td>80</td>
<td>4</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>38</td>
<td>3</td>
<td>38</td>
<td>2</td>
<td>41</td>
<td>5</td>
<td>57</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>22</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td>31</td>
<td>3</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>22</td>
<td>2</td>
<td>22</td>
<td>1</td>
<td>28</td>
<td>1</td>
<td>40</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1, 3, 4</td>
<td>11</td>
<td>1</td>
<td>22</td>
<td>1</td>
<td>29</td>
<td>2</td>
<td>13</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>38</td>
<td>3</td>
<td>38</td>
<td>2</td>
<td>38</td>
<td>4</td>
<td>57</td>
<td>3</td>
</tr>
</tbody>
</table>

![Diagram](image)
Circle and a Ring

MED(A, B) = 32
MED(B, A) = 28
Δ(A, B) = 38
centroid distance = 0
Mean Error Distance

MED(ST2, ARW) ≈ 15.42 is much smaller than MED(ARW, ST2) ≈ 66.16

High sensitivity to small changes in the field!

Good or bad quality depending on user need.

Fig. 2 from G. (2016 submitted to WAF, available at: http://www.ral.ucar.edu/staff/ericg/Gilleland2016.pdf)
# Geometric ICP Cases

Values rounded to zero decimal places

Table from part of Table 1 in G. (2016, submitted to WAF)

Fig. 1 from Ahijevych et al. (2009, WAF, 24, 1485 – 1497)

<table>
<thead>
<tr>
<th>Case</th>
<th>MED(A, Obs)</th>
<th>rank</th>
<th>MED(Obs, A)</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29</td>
<td>2</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>180</td>
<td>5</td>
<td>180</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>3</td>
<td>104</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>52</td>
<td>4</td>
<td>101</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>114</td>
<td>4</td>
</tr>
</tbody>
</table>
Geometric ICP Cases

<table>
<thead>
<tr>
<th>Case</th>
<th>MED(A, Obs)</th>
<th>rank</th>
<th>MED(Obs, A)</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>29</td>
<td>2</td>
<td>29</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>180</td>
<td>5</td>
<td>180</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>36</td>
<td>3</td>
<td>104</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>52</td>
<td>4</td>
<td>101</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>1</td>
<td>114</td>
<td>4</td>
</tr>
</tbody>
</table>

Values rounded to zero decimal places

Table from part of Table 1 in G. (2016, submitted to WAF)
Fig. 1 from Ahijevych et al. (2009, WAF, 24, 1485 – 1497)
Mean Error Distance

- Magnitude of MED tells how good or bad the “misses/false alarms” are.
- Miss = Average distance of observed non-zero grid points from forecast.
  - Perfect score: MED(Forecast, Observation) = zero (no misses at all)
    - All observations are within forecasted non-zero grid point sets.
  - Good score = Small values of MED(Forecast, Observation)
    - all observations are near forecasted non-zero grid points, on average.
- False alarm = Average distance of forecast non-zero grid points from observations.
  - Perfect score: MED(Observation, Forecast) = zero (no false alarms at all)
    - All forecasted non-zero grid points fall overlap completely with observations.
  - Good score = Small values of MED(Observation, Forecast)
    - all forecasts are near observations, on average.
- Hit/Correct Negative
  - Perfect Score: MED(both directions) = 0
  - Good Value = Small values of MED(both directions)
Note the Scales

Most models are closer to the NARR on average than the NARR is to them (more “misses” than “false alarms”). HRM3-HadCM3 and CRCM-NCEP are exceptions, but both have very small average distances in both directions.
0.95 quantile threshold

Misses
0.9 quantile threshold

Misses

False Alarms
MED Summary

• Mean Error Distance
  ▪ Useful summary when applied in both directions
  ▪ New idea of false alarms and misses (spatial context)
  ▪ Computationally efficient and easy to interpret

• Properties
  ▪ High sensitivity to small changes in one or both fields
  ▪ Does not inform about bias per se
    • Could hedge results by over forecasting, but only if over forecasts are in the vicinity of observations!
  ▪ No edge or position effects (unless part of object goes outside the domain)
  ▪ Does not inform about patterns of errors
  ▪ Does not directly account for intensity errors (only location)
  ▪ Fast and easy to compute and interpret

• Complementary Methods include (but not limited to)
  ▪ Frequency bias (traditional)
  ▪ Geometric indices (AghaKouchak et al 2011, doi:10.1175/2010JHM1298.1)
Baddeley’s $\Delta$ Metric Summary

- Sensitive to differences in size, shape, and location
- A proper mathematical metric (therefore, amenable to ranking)
  - positivity ($\Delta(A, B) \geq 0$ for all $A$ and $B$)
  - identity ($\Delta(A, A) = 0$ and $\Delta(A, B) > 0$ if $A \neq B$)
  - symmetry ($\Delta(A, B) = \Delta(B, A)$)
  - triangle inequality ($\Delta(A, C) \leq \Delta(A, B) + \Delta(B, C)$)
- Sensitive to position within the domain, edge effects, and orientation between two objects (so, when ranking, need to be careful if values are close)
  - For single object comparisons, perhaps could be overcome by centering and rotating (the pair of objects together) and calculating within a bounding box. Future work!
- Unbounded upper limit! (i.e., $\Delta(A, B)$ in $[0, \infty)$)
  - Can be alleviated by proper normalization (as is done here).
  - Need to take care when ranking anyway because of above issues.
Centroid Distance Summary

- Is a true mathematical metric. So, conducive to rankings.
- Not sensitive to position within a field (or orientation of A to B; i.e., if A and B are rotated as a pair, the distance does not change)
- No edge effects
- Gives useful information for translation errors between objects that are similar in size, shape and orientation.
- Not as useful otherwise.
- Should be combined with other information.
Spatial Forecast Verification

Image Warping

Forecast Image \((F(s))\)

Observed Image \((O(s))\)

Warped Image \((F(W(s)))\)

Graphic by Johan Lindström
Image Warping

Pair of thin-plate spline transformations

$$\Phi(s) = (\Phi_1(s), \Phi_2(s))^T = a + Gs + W^T \Psi(s - p_0)$$

- x-coordinate
- y-coordinate
- affine transformation

$$\Psi(h) = ||h||^2 \log ||h||$$

Nonlinear transformations

Columns of coefficients in $W$ and the sum of products of $W$ times $p_0$ both constrained to sum to 0.
Image Warping

Pair of thin-plate spline transformations

$$\Phi(s) = (\Phi_1(s), \Phi_2(s))^T = a + Gs + W^T \Psi(s - p_0)$$

$$LA = \begin{bmatrix}
\Psi & 1_k & p_0 \\
1_k^T & 0 & 0 \\
p_0^T & 0 & 0
\end{bmatrix} \begin{bmatrix}
w \\
a^T \\
g^T
\end{bmatrix} = \begin{bmatrix}
p_1 \\
0 \\
0
\end{bmatrix}$$

Want $L^{-1}$. The upper $k \times k$ matrix of $L^{-1}$, call it $L^{11}$, gives the bending energy matrix. And $W = L^{11}p_1$. The bending energy is given by trace($p_1^T L^{11}p_1$).
Image Warping

Ideally, want to find the optimal deformation without hand-selecting control points!

$k$ parameters of interest are the locations $p_1$.

Found by numerically optimizing the objective function:

$$Q(p_1) = \frac{1}{N\sigma^2} \sum_{s=1}^{N} \left( \hat{Z}(W(s)) - Z(s) \right)^2 +$$

$$\beta \left[ \begin{array}{c} \left( p_{1,x} - p_0 \right)^T L_{11} \left( p_{1,x} - p_0 \right) + \left( p_{1,y} - p_0 \right)^T L_{11} \left( p_{1,y} - p_0 \right) \end{array} \right]$$

RMSE of deformed Forecast against observation

Penalty for too much warping and too much bending

User-chosen penalty parameter. Controls how much bending and deformation can happen
Image Warping
Image Warping
Image Warping

MM5I-CCSM3

\( K \)

\( \text{RMSE}_0 = 0.2665 \)

\( \text{RMSE}_1 = 0.1605 \)

% error reduction \( \approx 40\% \)

minimum bending energy = 2.0042

Important for later
# Image Warping

<table>
<thead>
<tr>
<th></th>
<th>$\text{RMSE}_0$</th>
<th>$\text{RMSE}_1$</th>
<th>$\text{RMSE Reduction}$</th>
<th>$\text{Minimum Bending Energy}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRCM-CCSM3</td>
<td>0.214</td>
<td>0.139</td>
<td>35%</td>
<td>0.96</td>
</tr>
<tr>
<td>CRCM-CGCM3</td>
<td>0.147</td>
<td>0.103</td>
<td>30%</td>
<td>1.07</td>
</tr>
<tr>
<td>HRM3-HadCM3</td>
<td>0.157</td>
<td>0.110</td>
<td>30%</td>
<td>0.25</td>
</tr>
<tr>
<td>MM5I-CCSM3</td>
<td>0.267</td>
<td>0.161</td>
<td>40%</td>
<td>2.00</td>
</tr>
<tr>
<td>MM5I-HadCM3</td>
<td>0.148</td>
<td>0.084</td>
<td>43%</td>
<td>0.69</td>
</tr>
<tr>
<td>WRFG-CCSM3</td>
<td>0.249</td>
<td>0.096</td>
<td>61%</td>
<td>3.27</td>
</tr>
<tr>
<td>WRFG-CGCM3</td>
<td>0.241</td>
<td>0.092</td>
<td>62%</td>
<td>3.32</td>
</tr>
<tr>
<td>CRCM-NCEP</td>
<td>0.214</td>
<td>0.173</td>
<td>19%</td>
<td>0.25</td>
</tr>
<tr>
<td>WRFG-NCEP</td>
<td>0.171</td>
<td>0.092</td>
<td>46%</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Spatial Prediction Comparison Test

No significant results for these verification sets using standard SPCT.

AE + distance map loss
AE + deformation loss

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>SPCT Statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CRCM-CCSM3</td>
<td>CRCM-CGCM3</td>
<td>-1.24</td>
<td>0.21</td>
<td></td>
</tr>
<tr>
<td>CRCM-CCSM3</td>
<td>HRM3-HadCM3</td>
<td>1.15</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>CRCM-CGCM3</td>
<td>HRM3-HadCM3</td>
<td>1.66</td>
<td>0.10</td>
<td></td>
</tr>
<tr>
<td>HRM3-HadCM3</td>
<td>MM5I-CCSM3</td>
<td>-1.71</td>
<td>0.09</td>
<td></td>
</tr>
<tr>
<td>HRM3-HadCM3</td>
<td>WRFG-CCSM3</td>
<td>-1.45</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>HRM3-HadCM3</td>
<td>WRFG-CGCM3</td>
<td>-3.06</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>MM5I-CCSM3</td>
<td>WRFG-CGCM3</td>
<td>-2.12</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>MM5I-HadCM3</td>
<td>WRFG-CGCM3</td>
<td>-1.45</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>WRFG-CGCM3</td>
<td>WRFG-NCEP</td>
<td>1.42</td>
<td>0.16</td>
<td></td>
</tr>
</tbody>
</table>
Conclusions

- Models generally agree with NARR about spatial location and overall pattern of high severe storm frequencies ($\kappa$ and $\omega$).
- They tend to under-project the spatial extent of high frequency areas compared to NARR.
- HRM3-HadCM3 is by far the closest to NARR for both $\kappa$ and $\omega$.
- WRFG configurations not coupled with NCEP (i.e., “observations”) have the least agreement with NARR.
- Climate models should reproduce observed distributional properties for the current-period climate, making spatial forecast verification methods particularly useful, and easy to implement in this context.
- Full analysis including many other spatial methods in G. et al. (submitted to ASCMO, available at http://www.ral.ucar.edu/staff/ericg/GillelandEtAl2016.pdf)
Thank you. Questions?

- [http://www.ral.ucar.edu/staff/ericg](http://www.ral.ucar.edu/staff/ericg)
- Test cases for part 2 of ICP (MesoVICT)
  - [http://www.ral.ucar.edu/projects/icp](http://www.ral.ucar.edu/projects/icp)
  - Ensembles of models
  - Ensembles of observations
  - Precipitation, wind
  - complex terrain
  - point observations + re-analysis product