Evaluation of “non-standard” variables

Barbara Brown (bgb@ucar.edu)
National Center for Atmospheric Research (NCAR)
Boulder, Colorado, USA

Verification Methods Tutorial
Berlin, Germany
What makes variables “non-standard”

• Not focused on commonly measured weather variables (e.g., T, Td, Wind speed, u, v, etc.)
• ???
• Perhaps...
  – Not observed well or require special observations
  – Forecasts of things that are difficult to measure
  – Predictions directly serve specific users
    • Particular events are forecasted for particular decision-making situations (e.g., C&V for determining if planes can land)
    • The stakes can be high! (i.e., the decisions can have major safety and/or economic impacts)
Topics

- Tropical cyclones
- Wildfires and fire weather
- Sea ice
- Aviation
- Resources
TROPICAL CYCLONE FORECAST VERIFICATION

TC Inigo 2003

TC Gillian 2013
What makes TC forecast verification “special”? 

- **High impact weather and High impact weather Forecasts** 
  - TC weather impacts affect large populations and have major economic impacts 
  - TC weather forecasts impact disaster management decisions 

- **TC forecasts are given intense attention by the media and public – in the public “eye”, so to speak** 

- **Observations** are generally inferred and limited 

---

**What is not different?**

- Information is needed by Managers, Forecasters, Model developers and End users 
- Basic verification methods are applicable, in general (i.e., continuous, categorical, probabilistic)
What attributes of TC forecasts?

Deterministic

• TC track
  – Overall error
  – Cross-track error
  – Along-track error
  – Landfall timing and location

• Intensity
  – maximum wind
  – central pressure
  – temporal trend (rapid intensification)

• Wind field
• Size / radii
• Precipitation
• Temporal consistency
• Storm surge
• Waves
What Attributes of TC forecasts?

Ensemble

- Track distribution
- Strike probability
- Intensity distribution
  - mean / median
  - spread
  - 90$^{\text{th}}$ percentile
- Prob (wind $>$ threshold)
- Prob (precip $>$ threshold)
- Storm surge
- Landfall timing
What verification methods are appropriate?

Since we are evaluating a variety of variables and attributes...

A variety of methods are used

- **Categorical**
  - Rapid intensification / weakening

- **Continuous**
  - Intensity, track location, wind, size, winds, precipitation, ...

- **Probabilistic / ensemble**
  - Track and intensity, location ellipses, exceedance probabilities, precipitation, winds, size, strike probability, ...

- **Spatial**
  - Wind structure, precipitation, ...
What about observations?

Many hurricane observations are inferred...
As usual there is no such thing as “truth” – but maybe more so for tropical cyclones than other phenomena

Track and intensity
• Identified in “Best track” - Subjective analysis
  – Track: Latitude, longitude
  – Intensity: Minimum sea level pressure, maximum 1-min surface wind speed
• Best track is an analysis of all of the latest information about a storm in post-analysis
  – Uses satellite and reconnaissance information
  – Smoothed version of the track
  – Intensity often subjectively inferred from flight level winds or satellite information (Dvorak technique)

Precipitation and wind fields
Over oceans limited to satellite based information + data from reconnaissance
Forecast characteristics

- **Forecast types**
  - *Human-generated* tracks and intensity
  - *NWP Models*: Cyclone tracks are analyzed from gridded model output using a “tracker” algorithm
  - *Statistical models*: Especially useful for predicting intensity

- **Model interpolation**
  - Needed to adjust “late” models with current track information

- **Reference forecasts**
  - Statistical forecast or climate/persistence
Quality of deterministic TC Track forecasts

Example questions:

- What are the track errors (along-track, cross-track)?
- What are the intensity errors?
- Are temporal intensity trends correctly predicted?
- What is the error in timing of landfall?
- What is the error in forecast maximum wind (rain)?
  - Multi-day total precipitation
- Is the spatial distribution of wind (rain) correct?

Others?
Total, Along-, and Cross-Track Errors

Cross-track measures error in direction of movement.

Along-track measures error in speed of movement.

Track error
Cross-Track Error (Forecast too far to the right)
Along-Track Error (Forecast too slow)

Courtesy, J. Franklin
Track error summary

Track error is typically summarized as Average error (always positive)
Verification methods for deterministic TC forecasts

- Example: Along-track and cross-track errors

- “Along-track” measures errors in “Speed”
- “Cross-track” measures errors in “Direction”

Courtesy James Franklin, NHC
Intensity error is typically summarized as (1) mean error (bias) or (2) mean absolute error (always positive).
Alternative:
Examine distributions of errors

And differences in errors
Paired comparisons: Track and intensity
% improvement and p-value

<table>
<thead>
<tr>
<th>Track and water)</th>
<th>rack</th>
<th>HMI</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.0</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-15%</td>
<td>-3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.999</td>
<td>0.979</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-17%</td>
<td>-5.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.999</td>
<td>0.977</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-17%</td>
<td>-5.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.999</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-17%</td>
<td>-9.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.999</td>
<td>0.989</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-12%</td>
<td>0.927</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4%</td>
<td>0.551</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.989</td>
<td>0.661</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7%</td>
<td>0.508</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7%</td>
<td>0.836</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-4%</td>
<td>0.836</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-7%</td>
<td>0.836</td>
<td></td>
</tr>
</tbody>
</table>

Target forecasts significantly improve on standard of comparison for intensity forecasts.
Paired comparisons: Track and intensity % improvement and p-value

<table>
<thead>
<tr>
<th>Track</th>
<th>% improvement</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>HMI</td>
<td>-3.1</td>
<td>-11.3</td>
</tr>
<tr>
<td>HMI</td>
<td>-5.1</td>
<td>-19.8</td>
</tr>
<tr>
<td>HMI</td>
<td>-11.5</td>
<td>-10.3</td>
</tr>
<tr>
<td>HMI</td>
<td>-15.0</td>
<td>-28.0</td>
</tr>
<tr>
<td>HMI</td>
<td>-35.4</td>
<td>-47.0</td>
</tr>
<tr>
<td>rack</td>
<td>-15%</td>
<td>-16%</td>
</tr>
<tr>
<td>rack</td>
<td>-17%</td>
<td>-18%</td>
</tr>
<tr>
<td>rack</td>
<td>-19%</td>
<td>-12%</td>
</tr>
<tr>
<td>rack</td>
<td>-19%</td>
<td>-19%</td>
</tr>
<tr>
<td>rack</td>
<td>-23%</td>
<td></td>
</tr>
<tr>
<td>and water</td>
<td>0.999</td>
<td>0.999</td>
</tr>
<tr>
<td>and water</td>
<td>0.990</td>
<td>0.950</td>
</tr>
<tr>
<td>and water</td>
<td>0.929</td>
<td>0.971</td>
</tr>
<tr>
<td>and water</td>
<td>0.999</td>
<td>0.987</td>
</tr>
<tr>
<td>and water</td>
<td>0.927</td>
<td>0.989</td>
</tr>
<tr>
<td>and water</td>
<td>0.927</td>
<td>0.551</td>
</tr>
<tr>
<td>and water</td>
<td>0.950</td>
<td>0.661</td>
</tr>
<tr>
<td>and water</td>
<td>0.971</td>
<td>0.508</td>
</tr>
<tr>
<td>and water</td>
<td>0.987</td>
<td>0.836</td>
</tr>
</tbody>
</table>

Target forecasts significantly reduce performance relative to the standard of comparison for track forecasts and some intensity forecasts.
Using a flexible definition of Rapid Intensification / Rapid Weakening events

**Standard Definition:** NHC Definition 30m/s in 24 hours

**Stricter Definition:** 30m/s in 12 hours

**“Fuzzy” Definition:** Adjustable window to give credit even if there is a timing error

Categorical statistics for RI/RW events can then be calculated: POD, FAR, CSI, etc.
Evaluating features: TC precipitation evaluation

Storm-following masking with range rings

Shifted forecast precipitation to account for track error, with range rings around the best track

Accumulated storm precipitation distributions for Model, Satellite, and Radar by range ring
WILDFIRES AND FIRE WEATHER
Fire weather verification

- Wildfire conditions and associated weather can be predicted by humans, spread simulators, or coupled weather-fire models

- Variables of interest:
  - Fire perimeter
  - Fire rate-of-spread
  - Underlying wind and other weather variables
  - Significant fire behavior (flame length, pyrocumulus, etc.)

Many complications with evaluation of these variables
Meeting the users’ needs

Australia BOM process

• Focus on process to identify and document stakeholders’ goals
• Different users have different needs
  – Management (Which model/simulator is best?)
  – Fire behavior analysts (How accurate are fire predictions?)
  – Simulator / Model developers (quantify uncertainty in weather inputs to identify simulator improvements needed)
Observation issues

• Fire perimeter
  – Observed from the air?
  – Satellite?
  – Obs are infrequent at best...

• Only rare observations of significant phenomena (flame height, heat release, pyrocumulus, etc.)

• Weather observations very limited...
  – Poor coverage in complex terrain
Verification approaches

- Spatial methods
  - MODE? CRA?
- Contingency table statistics (TS, Bias)
- Area measures

**Issue:** What about the impact of fire suppression efforts?

From Ebert presentation Monday
SEA ICE
The Challenges

• Arctic sea ice is changing dramatically and quickly
• Climate, seasonal, and other models depend on good estimates of sea ice extent – and other characteristics
• Many users interested in impacts of changes in ice (shipping, mining, etc.)
• Observations are limited...
  – Mainly satellite-based
  – Ice extent is best observed; other properties (thickness, concentration) more limited
Possible verification methods

• Spatial
  – MODE
  – CRA
  – Image warping

• Distance metrics:
  – Baddeley, Hausdorff (see methods in R package)
  – See references by Gilleland and others at https://ral.ucar.edu/projects/icp/

From Arbetter 2012
AVIATION WEATHER
• Main issue: Observations!!
  – Limited in space and time
  – Biased in space and time, and by event (e.g., around airports, on flight routes; where weather is good!)
Example: Icing PIREPs

Notable biases in location, time, intensity

Potentially systematic in areas near airports?

From Brown et al. 1997 (Weather and Forecasting)
EDR (turbulence) example: Automated observations

Spatial biases and highly skewed distribution

- Difficult to tune forecasts to predict “positive” events
- Turbulence forecasts may not be representative of areas where planes don’t fly

From Sharman et al. 2014 (*J. Appl. Climate and Weather*)
TAIWIN: Terminal Area Icing Weather Information for NextGen (TAIWIN)

- **Goal**: Improve NWP forecasts of precip type (especially freezing rain/drizzle) to predict super-cooled liquid

- **First step**: Identify appropriate observations
  - METARs
  - Radar/Satellite
  - Crowd-sourced (MPING)

*Courtesy J. Wolff*
Perfect score
Overforecast
Underforecast

Bias
Critical Success Index

Credit: J. Wolff, NCAR
Observation implications...

- Observation characteristics often *limit the kinds of verification* that can be done
  - *Ex:* In-flight icing, turbulence; gridded C&V
- Observation characteristics can *bias the verification results*
- Improvement of aviation weather observations would greatly help improve development and evaluation of aviation weather forecasts
SUMMARY
Summary

• Standard verification methods apply to most variables
  But:
  – Focus is needed on what aspects users care about
    • Good news: Typically easier to understand who the users are
  – Care needed to understand implications of observation biases
    • May limit what verification approaches are reasonable to apply

• Observation issues – availability, uncertainty – are even more important than for more standard variables!
RESOURCES
Many resources on the WMO website:

https://www.wmo.int/pages/prog/arep/wwrp/new/Forecast_Verification.html

- Guidance documents
- Links to past tutorials
- Information about upcoming meetings, etc.
Resources: Verification methods and FAQ

- Website maintained by WMO verification working group (JWGFVR)
- Includes
  - Issues
  - Methods (brief definitions)
  - FAQs
  - Links and references
- Verification discussion group:
  http://mail.rap.ucar.edu/mailman/listinfo/vx-discuss
Resources: Overview papers


• Ebert et al. 2013: *Progress and challenges in forecast verification* 
  Meteorological Applications, 20, 130-139.
Information about spatial methods

MesoVICT website:

- Includes references and some software
- New results will be provided as available

http://www.ral.ucar.edu/projects/icp/
Resources - Books


Software: R verification packages

- R is a flexible, open source statistical computing package
  - Available from https://www.r-project.org/
  - Works on all platforms
- “Verification” and “SpatialVx” packages are available from the contributed packages list on the “CRAN” website: http://cran.repo.bppt.go.id/ in Indonesia

- “Verification” package
  - Includes all standard verification metrics for
    - Contingency tables
    - Continuous variables
    - Probability forecasts
    - Ensemble forecasts
  - Many graphical tools (e.g., attribute diagrams)
- “SpatialVx” package
  - Includes many of the new spatial verification methods
  - New methods added as available
Software: Model Evaluation Tools (MET)

- MET is a freely available software package
- Supported to the community and well-documented
- Highly configurable and flexible
- Tutorials (on-line and in person) are available

Includes
- Traditional methods (contingency table, continuous, probabilistic)
- Ensemble approaches
- Spatial methods
- Package for Tropical Cyclones (MET-TC)

MET is available at: www.dtcenter.org/met/users