Evaluation of Weather and Climate Forecasts: A 2016 perspective

Barbara Brown and Eric Gilleland

NCAR, Boulder, Colorado, USA

AMS Annual Meeting
14 January 2015
Goals

To understand where we are going, it’s helpful to understand where we have been...

• Examine the evolution of verification over the last few decades
  ▪ Where have we come from?
  ▪ Where are we going?
  ▪ Raise questions regarding verification practices along the way...

• What is left to do?
  ▪ What are the opportunities?
  ▪ What are the problems/impediments?
Early verification

- Finley period... (see Murphy paper on “The Finley Affair”; WAF, 11, 1996)
- Focused on contingency table statistics
- Development of many of the common measures still used today:
  - Gilbert (ETS)
  - Peirce (Hanssen-Kuipers)
  - Heidke
  - Etc...
- These methods are still the backbone of many verification efforts – for warnings and events
  - Are they used too often?
  - Should everything be categorized? What is lost by categorizing things into Yes/No?
  - How can this categorization be made more user-friendly/relevant?

<table>
<thead>
<tr>
<th>Observed</th>
<th></th>
<th></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>
Early years continued: Continuous measures

- Focus on squared error statistics
  - Mean-squared error
  - Correlation
  - Bias
  - Note: Little recognition before Murphy of the non-independence of these measures

- Extension to probabilistic forecasts
  - Brier Score (1950) – well before prevalence of probability forecasts!

- Development of “NWP” measures
  - S1 score
  - Anomaly correlation
  - Still relied on for monitoring and comparing performance of NWP systems (Are these still the best measures for this purpose?)
The “Renaissance”: The Allan Murphy era

- Expanded methods for probabilistic forecasts
  - Decompositions of scores led to more meaningful interpretations of verification results
  - Attribute diagram

- Initiation of ideas of meta verification: Equitability, Propriety

- Statistical framework for forecast verification
  - Clarified multiple perspectives for evaluating forecasts (joint distribution factorizations)
  - Placed verification in a statistical context, where it belongs

- Connections between forecast “quality” and “value”
  - Evaluation of cost-loss decision-making situations in the context of improved forecast quality
  - Non-linear nature of quality-value relationships
“Forecasts contain no intrinsic value. They acquire value through their ability to influence the decisions made by users of the forecasts.”

“Forecast quality is inherently multifaceted in nature... however, forecast verification has tended to focus on one or two aspects of overall forecasting performance such as accuracy and skill.”

Allan H. Murphy, *Weather and Forecasting*, 8, 1993: “What is a good forecast: An essay on the nature of goodness in forecasting”
Murphy era cont.

Development of the idea of “diagnostic” verification

- Also called “distribution-oriented” verification
- Focus on measuring or representing attributes of performance rather than relying on summary measures
- A revolutionary idea: Instead of relying on a single measure of “overall” performance, ask questions about performance and measure attributes that are able to answer those questions

Example: Use of conditional quantile plots to examine conditional biases in forecasts
The “Modern” era

• New focus on evaluation of ensemble forecasts
  ▪ Development of new methods specific to ensembles (rank histogram, CRPS)
• Greater understanding of limitations of methods
  ▪ “Meta” verification
• Evaluation of sampling uncertainty in verification measures
• Approaches to evaluate multiple attributes simultaneously (**note**: this is actually an extension of Murphy’s attribute diagram idea to other types of measures)
  • **Ex**: Performance diagrams, Taylor diagrams
The “Modern” era cont.

- Development of an international Verification Community
  - Workshops, textbooks...

- Evaluation approaches for special kinds of forecasts
  - Extreme events (Extremal Dependency Scores)
  - “NWP” measures

- Extension of diagnostic verification ideas
  - Spatial verification methods
  - Feature-based evaluations (e.g., of time series)

- Movement toward “User-relevant” approaches

From Ferro and Stephenson 2011 (Wx and Forecasting)
Spatial verification methods

Inspired by the limited *diagnostic* information available from traditional approaches for evaluating NWP predictions

• Difficult to distinguish differences between forecasts

• The double penalty problem
  ▪ Forecasts that appear good by the eye test fail by traditional measures... often due to small offsets in spatial location
  ▪ Smoother forecasts often “win” even if less useful

• Traditional scores don’t say what went wrong or was good about a forecast

• Many new approaches developed over the last 15 years

• Starting to also be applied in climate model evaluation
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
How should the forecast be adjusted to make the best match with the observed field?


**Table 11. Spatial prediction comparison test for frequency of CAPE > 1000 m/s\(^{-1}\) conditioned on q75’s being greater than its 90-th percentile. Two-sided p-value is given inside (·).**

<table>
<thead>
<tr>
<th>Threshold</th>
<th>CRCM-CGCM3</th>
<th>HRM3-HadCM3</th>
<th>MM5I-CCSM</th>
<th>MM5I-HadCM3</th>
<th>WRFG-CCSM</th>
<th>WRFG-CGCM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>0.011 (0.991)</td>
<td>0.005 (0.996)</td>
<td>-0.028 (0.978)</td>
<td>0.001 (0.999)</td>
<td>0.000 (≈ 1)</td>
<td>-0.002 (0.998)</td>
</tr>
<tr>
<td>0.5</td>
<td>0.287 (0.774)</td>
<td>0.120 (0.904)</td>
<td>-0.083 (0.934)</td>
<td>0.066 (0.948)</td>
<td>-0.164 (0.870)</td>
<td>-0.061 (0.951)</td>
</tr>
<tr>
<td>None</td>
<td>0.000 (≈ 1)</td>
<td>-0.014 (0.989)</td>
<td>2.781 × 10(^{-15}) (1)</td>
<td>-0.009 (0.993)</td>
<td>-0.006 (0.995)</td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>-0.007 (0.995)</td>
<td>-0.291 (0.771)</td>
<td>-0.162 (0.871)</td>
<td>-0.394 (0.694)</td>
<td>-0.284 (0.776)</td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>-0.007 (0.994)</td>
<td>0.000 (≈ 1)</td>
<td>-0.006 (0.995)</td>
<td>-0.004 (0.997)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>-0.143 (0.886)</td>
<td>-0.093 (0.926)</td>
<td>-0.204 (0.840)</td>
<td>-0.168 (0.866)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0.001 (0.999)</td>
<td>0.025 (0.980)</td>
<td>0.017 (0.987)</td>
<td>0.017 (0.986)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.142 (0.887)</td>
<td>-0.056 (0.956)</td>
<td>-0.001 (0.999)</td>
<td>-0.001 (0.999)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>None</td>
<td>0.005 (0.999)</td>
<td>-0.208 (0.835)</td>
<td>-0.173 (0.863)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.5</td>
<td>0.098 (0.922)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Meta-evaluation of spatial methods: What are the capabilities of the new methods?

- MesoVICT incorporates more variables and more realistic forecast verification cases
  - precipitation, wind, pressure, etc.
  - complex terrain
  - cases over multiple time points (and lead times)
  - forecast ensembles
  - ensembles of observations
  - point observations and analysis product
- New spatial verification methods since ICP
- Evaluation of climate models
  - lack of temporal compatibility in most cases, but not a problem (actually makes things simpler)
  - especially well suited to spatial verification methods
User-relevant verification: The Weather Information Value Chain

In recent years, there has been a new focus on User-relevant verification – moving toward Murphy’s vision of diagnostic verification and the connections between forecast quality and value.

Courtesy of Jeff Lazo
User-relevant verification

Levels of user-relevance

1. Making traditional verification methods useful for a range of users (e.g., variety of thresholds)
2. Developing and applying specific methods for particular users [Ex: Particular statistics; user-relevant variables]
3. Applying meaningful diagnostic (e.g., spatial) methods that are relevant for a particular users’ question
4. Connecting economic and other value directly with forecast performance

Most verification studies are at Levels 1 and 2, with some approaching 3, and very few actually at Level 4

Some examples....
### Solar Power forecasts: Adapting standard metrics to meet user needs

#### Preferences for end users:
- Raw Values (Non-normalized)
- Normalized by Capacity/Clear Sky
- Normalized by Actuals

#### Sum of Errors over Forecast Period

<table>
<thead>
<tr>
<th>Component</th>
<th>Intra Hour</th>
<th>Day Ahead</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp1</td>
<td>960 (-76 / 1036)</td>
<td>858 (-79 / 937)</td>
</tr>
<tr>
<td>Comp2</td>
<td>1418 (-43 / 1461)</td>
<td>1295 (-43 / 1339)</td>
</tr>
<tr>
<td>Comp3</td>
<td>1433 (-33 / 1466)</td>
<td>1106 (-5 / 1111)</td>
</tr>
<tr>
<td>Comp4</td>
<td>510 (-84 / 594)</td>
<td>n/a</td>
</tr>
<tr>
<td>Comp5</td>
<td>1748 (-12 / 1760)</td>
<td>n/a</td>
</tr>
<tr>
<td>Comp6</td>
<td>1004 (-40 / 1044)</td>
<td>n/a</td>
</tr>
<tr>
<td>Blended Model</td>
<td>226 (-92 / 318)</td>
<td>178 (-110 / 288)</td>
</tr>
</tbody>
</table>

**User:** Energy Trader

Sum of Errors over entire day more informative

---

**Courtesy T. Jensen**
Capturing Energy Ramps

Frequency Bias = \( \frac{\# \text{ of Fcst Events}}{\# \text{ Observed Events}} \)

Baseline

Credit: T. Jensen
Applications of object–based approaches

Example: Evaluation of jet cores, highs, lows (using MODE object based approach) for model acceptance testing

Courtesy Marion Mittermaier, UK Met Office
Comments on user-relevant verification

• Moving toward user relevant verification will increase both the usefulness and quality of forecasts, and will benefit developers as well as users

• Many of the steps toward user relevance (e.g., user-specified stratifications & thresholds) are easy to achieve
  ▪ Others require major multi-disciplinary efforts

• Verification practitioners – people who do verification – should endeavor as much as possible to understand the needs of the forecast users

• Much is left to be explored by creative minds!
**Challenge: Develop best new user-relevant verification method**

- **Sponsored by WMO/WWRP**
  - Joint Working Group on Forecast Verification Research;
  - High Impact Weather (HIW), Sub-seasonal to seasonal (S2S), and Polar Prediction (PPP) projects

- **Focus**
  - All applications of meteorological and hydrological forecasts (including TCs, seasonal and sub-seasonal, polar, ...)
  - Metrics can be quantitative scores or diagnostics (e.g., diagrams)

- **Criteria for being selected as “best”**
  - Originality, user relevance, intuitiveness, simplicity and ease of computing, robustness, and resistance to hedging.
  - Desirable characteristics:
    1. Clear statistical foundation;
    2. Applicability to a broader set of problems
Challenge: Develop best new user-relevant verification method

• Dates:
  • Formal announcement: Sept 2015
  • Deadline for submission: 31 Oct 2016

• Prize: Invited keynote talk at the 7th International Verification Methods Workshop in 2017

• Contact verifchallenge@ucar.edu for more information

• See website at
  http://www.wmo.int/pages/prog/arep/wwrp/new/FcstVerChallenge.html
Summary

• Much progress has been made in the last few decades, advancing the capabilities and potential impacts of forecast evaluation

• Many new approaches have been developed, examined, and applied, and are providing opportunities for more meaningful evaluations of both weather and climate forecasts

• But still more challenges ahead...
Remaining challenges (some examples)

• Expansion of user-relevant metrics – moving to Levels 3 and 4, expanding Levels 1 and 2
  ▪ Providing a breadth of information to users

• Sorting out how to incorporate uncertainty appropriately
  ▪ Spatial / Temporal
  ▪ Measurement / Observation
  ▪ Sampling

• Improving communication
  ▪ How do we communicate forecast quality information to the general public? To specific users?