Verification for High Impact Weather

Beth Ebert
Bureau of Meteorology, Australia

Acknowledgements: Barb Brown, Michael Sharpe, Manfred Dorninger, Helen Titley, Joanne Robbins, Lesley Allison, Brian Golding
What is high impact weather?

• Affects people
• Involves making important decisions

Natural hazard related deaths in the US 1970-2004

A fire weather warning for Friday is current in the Northern Country, Wimmera, Mallee, North Central and Northeast forecast districts. Temperatures up to 41 degrees, relative humidity down to 9% and winds to 25 km/h will cause extreme fire danger. CFA advises people living in areas at risk of fire to activate their bush fire plan. The next warning will be issued by 11:00 pm EDT Thursday.

---

Warnings
Outlooks
NWP

... Starting to link to hazard impact models
Challenges in modelling high impact weather

- Models may not capture the intensity of high impact events
  - Sub grid scale processes
  - Coarse resolution
  - Difficulty representing processes
- May be a mismatch between what models can provide and what warnings need to be made for
  - Lightning, hail, wind gusts, fog, ...
- Large uncertainty with extreme events
  - Ensemble / probabilistic forecasts
  - Extreme forecast index (EFI) and anomaly forecasts (ANF) measure relative "extremeness"
Verification for high impact weather

• How should we do it?
• What recent research can assist?
• What are some of the challenges requiring further research?
Useful verification of HIW events

Guides users in making better decisions based on forecasts

• How reliable is the forecast at capturing events?
• What are typical errors in timing / location / intensity of events?
• Are the forecasts biased?

Informs modellers / forecast system developers on how to improve forecasts

• Do the forecasts show the right behaviour?
• What is the nature of the errors?

Assists managers in monitoring forecast performance
Historical perspective


<table>
<thead>
<tr>
<th>Month</th>
<th>Predictions for</th>
<th>Total number</th>
<th>Number predictions favorable</th>
<th>Fully verified</th>
<th>Number predictions unfavorable</th>
<th>Fully verified</th>
<th>Total number made</th>
<th>Total number fully verified</th>
</tr>
</thead>
<tbody>
<tr>
<td>March</td>
<td>8 hours</td>
<td>771</td>
<td>43</td>
<td>6</td>
<td>728</td>
<td>731</td>
<td>771</td>
<td>737</td>
</tr>
<tr>
<td>April</td>
<td>8 hours</td>
<td>934</td>
<td>25</td>
<td>11</td>
<td>909</td>
<td>906</td>
<td>934</td>
<td>917</td>
</tr>
<tr>
<td>May</td>
<td>8 hours</td>
<td>558</td>
<td>10</td>
<td>8</td>
<td>548</td>
<td>542</td>
<td>558</td>
<td>550</td>
</tr>
<tr>
<td>May</td>
<td>16 hours</td>
<td>549</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

May, 1917.

Abnormally low temperature prevailed in the States to the west and north. Frost warnings were issued for Illinois, Missouri, Iowa, Wisconsin, and extreme eastern Kansas, and were generally verified. Warnings were issued each day from the 5th to 13th, inclusive, for some portion of the upper Mississippi Valley or western Lakes region, and were partially verified. No further warnings
Modern perspective – case studies

Service Assessment

The Historic South Carolina Floods of October 1–5, 2015

U.S. DEPARTMENT OF COMMERCE
National Oceanic and Atmospheric Administration
National Weather Service
Silver Spring, Maryland
Modern perspective – systematic verification

MODE verification performed every day at NWS Weather Prediction Center
Challenges in observing high impact weather

• Rare events
• Sampling error (timing, location, magnitude)
• Measurement error (gauge undercatch, radar attenuation, etc.)
• Non-reports
• May not match the forecast space & time scales (representativeness "error")

El Reno, TX, weather station post-tornado
Photo: Cliff Mass weather blog
3rd party observations and crowd sourcing

Weather Observations Website (WOW)

Mobile Precipitation Identification Near the Ground (mPING)
Observation uncertainty in verification

• As models improve, we can no longer ignore observation error!

• What are the effects of ignoring the observation error?
  – Forecasts may actually be better than they seem
  – Should users of verification results be advised?

• What are the effects of including the observation error?
  – “Noise” leads to poorer scores for deterministic forecasts
  – Probabilistic/ensemble forecasts have poorer reliability & ROC

➢ 7IVMW session on observation uncertainty
How does observation uncertainty compare to forecast uncertainty in verification?

- 6h forecasts of hourly precipitation, 11th June – 26th August 2015
- Observation (VPR) uncertainty – UKV vs radar ensemble (13 members)
- Forecast uncertainty – MOGREPS-UK ensemble (12 members) vs radar
- Fractions skill score for 51km neighbourhood

![Graphs showing comparison of UKV vs radar ensemble and MOGREPS-UK ensemble against unperturbed radar analysis.](image)

Courtesy Lesley Allison, Met Office
Dealing with observation uncertainty

• Strategies for reducing observation error
  – Quality control on measurements, correction of systematic errors
  – Averaging / analysis to larger space and time scales
  – Multiple observation sources

• Some approaches estimate the "true" verification scores, i.e., what would be computed if there were no observation error
  – Obs error distribution must be very well known and spatially uncorrelated

• "Tolerant" verification approaches
  – Distributions-based diagnostics including binning, quantiles, error bars
  – Object-based methods
  – Neighbourhood verification methods
  – Probabilistic observations → probabilistic scores
Simple verification approaches suit some users

- Easy to understand
- Can guide decision making

Contingency table

<table>
<thead>
<tr>
<th>Forecast events</th>
<th>Observed events</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>yes</td>
<td>yes</td>
<td>59</td>
<td>22</td>
</tr>
<tr>
<td>no</td>
<td>yes</td>
<td>34</td>
<td>674</td>
</tr>
<tr>
<td></td>
<td>no</td>
<td>93</td>
<td>696</td>
</tr>
</tbody>
</table>

- Total number of forecasts: 789
- Total number of observed events: 708
- Total number of forecasted events: 93

Q1: Given that an event is forecast, what is the chance that the event will actually occur?

\[
\frac{59}{93} \times 100 = 63\%
\]

Q2: When events occur, how often is the forecast correct?

\[
\frac{59}{81} \times 100 = 73\%
\]

Q3: Do the forecasts predict events too often / not often enough?

\[
\frac{93-81}{81} \times 100 = 15\% \quad \text{(too frequent)}
\]
Simple verification approaches

**Deterministic limit** – how long does it take until the forecast is more wrong than right?

- Can be used to set appropriate targets for warning provision
- Provides guidance on when to switch from deterministic forecasts to probabilistic ones

(Hewson 2007)
Verifying rare extreme values

Scoring categorical forecasts
– Metrics should reward hits, penalise misses and false alarms
– For rare events, traditional categorical scores like ETS \[ 0 \]
– Symmetric extremal dependency index:

\[ SEDI = \frac{\log F - \log H - \log(1 - F) + \log(1 - H)}{\log F + \log H + \log(1 - F) + \log(1 - H)} \]
Verifying probability forecasts

- Cannot verify an individual probability forecast
- Probabilistic verification requires a large sample of forecasts
- Difficult to explain to many people
- Continuous Ranked Probability Score (CRPS) emerging as score of choice for model verification

![Likelihood diagram](image1)

- Observed non-events
- Observed events

![Reliability diagram](image2)

- Forecast probability
- Observed relative frequency

![Relative Operating Characteristic](image3)

- Hit rate
- False alarm rate

simple \[\longrightarrow\] complex
Verification of extreme events

Summer day-time max temperatures over UK, 2014-2015

How much better at predicting relative-extremes was the forecast compared with the climatology?

Score > 0.5 means the forecast was better than the climatology

Skill decreases with increasing forecast range

Even the forecast on day 9 is better than the climatology

Courtesy Michael Sharpe, Met Office
Other modifications of CRPS

• Rare/extreme values are in the tails of the climatological distribution

• Possible strategies
  – Weighted scoring rules
  – Extreme value theory
  – Quantile verification

Talks this session by
Petra Friederichs,
Maxime Taillardat,
Sebastian Lerch, Hong Guan
Generalized Discrimination Score (GDS)

Two-alternative forced choice:

- **Observation 1** vs. **Forecast 1**
- **Observation 2** vs. **Forecast 2**
- **Observation 3** vs. **Forecast 3**
- **Observation N-1** vs. **Forecast N-1**
- **Observation N** vs. **Forecast N**

- Obs correctly discriminated?
  - YES / NO

GDS = proportion of successful rankings
(no skill = 50%)

Talks by Roger Harbord, Alexander Jordan

Mason & Weigel, MWR, 2009
Seamless verification to span scales

Zhu et al. 2014
Spatial verification

Object-oriented

Scale separation

Neighborhood

Field distortion

Distance metrics – watch this space…
Object-based vs. traditional verification

- Traditional scores suggest the forecast was very poor
- MODE provides much more information about performance than traditional scores
- MODE defines and quantifies the flaws and good qualities of the forecast:
  - Many misses and false alarms (small objects/areas)
  - Significant storm area somewhat too large and too intense, but placed well
  - Less significant storm area (SE) too small and not intense enough

<table>
<thead>
<tr>
<th>TRADITIONAL SCORES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>POD</td>
<td>0.22</td>
</tr>
<tr>
<td>FAR</td>
<td>0.86</td>
</tr>
<tr>
<td>CSI</td>
<td>0.09</td>
</tr>
<tr>
<td>GILBERT (ETS)</td>
<td>0.08</td>
</tr>
<tr>
<td>BIAS</td>
<td>1.6</td>
</tr>
</tbody>
</table>
Neighborhood verification credits "close" forecasts

**Fractions skill score** compares forecast and observed fractional coverage (Roberts and Lean 2008)

**Multi-event contingency table** measures whether a forecast event is close to an observed event (Atger 2001)

Knowing which scales have skill suggests the scales at which the forecast should be presented and trusted.
Flexible verification of warnings

Sharpe, Met Apps 2016
MesoVICT: Mesoscale Verification Intercomparison over Complex Terrain

- How well do spatial verification methods work in complex terrain?
- Can they be used effectively to verify other parameters besides precipitation, e.g., wind?
- Can spatial verification methods be applied to ensemble forecasts?
- Can they account for uncertainty in observations?

http://www.ral.ucar.edu/projects/icp/
MesoVICT experiment design

- Ensemble forecasts
- Ensemble analyses to explore observation uncertainty
Spatial verification and ensembles

• Neighborhood verification is easily extended to ensembles
• Adapting existing scores for comparing probabilistic forecasts and probabilistic observations

\[
CRPS = \int_{-\infty}^{\infty} (P_{\text{fcst}}(x) - P_{\text{obs}}(x))^2 \, dx
\]

• SAL also applies well to ensembles
  ➢ Talks by Craig Schwartz, Marion Mittermaier, Helge Goessling, Sabine Radanovics
Weather forecasts → impact forecasts

- Floods
- Tourism
- Sports
- Emergency management
- Energy
- Roads
- Agriculture
- Air travel
Summarise risk of high-impact weather across the globe in the next 7 days using global multi-model ensemble forecasts

- Precipitation / wind / snow
- Tropical cyclones
- Heatwave and coldwave

Web Map Service

Symbol-based summary map

Drill down to particular variables / days / models / areas of interest

Overlay vulnerability and exposure layers

- Population density
- Fragile State Index
- Soil moisture
- Recent earthquakes
GHM forecast layers and identifying high-impact weather events

ECMWF ENS; MOGREPS-UK; Multi-Model

Day 3 forecast from 00Z 09/03/2016
Day 4 forecast from 00Z 19/01/2016
Day 4 forecast from 00Z 25/03/2016
Day 5 forecast from 12Z 15/06/2015
Day 6 forecast from 00Z 15/06/2015

24hr Precipitation Accum.
24hr Snowfall Accum.
24hr Max. Wind Gust
Excess Heat Factor (EHF)
Excess Cold Factor (EHF)

Probability of exceeding the 99th centile of forecast climatology

Summary polygons, coloured by lead time, show areas where probabilities are significant (≥0.4) for that lead time and hazard
Global Hazard Map: evaluation of precipitation forecasts

How does GHM perform in meeting its key aim “to summarise the risk of high-impact weather for the week ahead”?

(1) Did the forecast weather at a certain level of severity occur?

*Traditional ensemble-based verification against weather observations*

Comparing gridded hazard forecasts against station-based weather observations to create contingency based verification statistics as to whether or not the weather event occurred

(2) Did the forecast weather result in a high-impact event?

*Newly developed impact-based evaluation method*

Aims to evaluate how well the Global Hazard Map summary polygons relate to records of community impacts (e.g. fatalities, injuries, displacement, evacuation, receipt of aid, disruption, denial of access, hardship)
GHM: (1) Verification against precipitation observations

- Verification against global station-based observations (3315 sites) from Feb-Dec 2015

- **Forecast event**: probability of 24-hour precipitation exceeding the 99th percentile in the forecast climatology

- **Observed event**: 24-hour precipitation exceeding the 99th percentile in the observed climatology at that site

- Calculated contingency based statistics (reliability, ROC diagram, Brier skill score, etc.) for each of the three model precipitation layers (ECMWF ENS, MOGREPS-G and the multi-model ensemble)

- Skill (area under ROC curve) greatest for multi-model at all lead times

- Good skill shown throughout forecast period
Socio-economic Impact Databases

<table>
<thead>
<tr>
<th>Heavy Rainfall Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spatial_ID (entry ID)</td>
</tr>
<tr>
<td>Event_ID (hazard event ID)</td>
</tr>
<tr>
<td>Record Date</td>
</tr>
<tr>
<td>Start Date</td>
</tr>
<tr>
<td>End Date</td>
</tr>
<tr>
<td>Hazard_Type (Heavy rainfall)</td>
</tr>
<tr>
<td>Trigger/Cause</td>
</tr>
<tr>
<td>Secondary Hazards</td>
</tr>
<tr>
<td>Hazard Notes</td>
</tr>
<tr>
<td>Country Name</td>
</tr>
<tr>
<td>Region/State/Province Name</td>
</tr>
<tr>
<td>Region/State/Province Latitude</td>
</tr>
<tr>
<td>Region/State/Province Longitude</td>
</tr>
<tr>
<td>Settlement Name</td>
</tr>
<tr>
<td>Settlement Latitude</td>
</tr>
<tr>
<td>Settlement Longitude</td>
</tr>
<tr>
<td>Impact Information</td>
</tr>
<tr>
<td>Impact Categorisation</td>
</tr>
<tr>
<td>References</td>
</tr>
</tbody>
</table>

Location of heavy rainfall impacts (February – December 2015)

Between February 8th and December 31st 2015 a total of 261 heavy rainfall events were recorded, resulting in 853 database entries.
GHM: (2) Evaluation against rainfall impact observations

- Forecast heavy rainfall events compared to heavy rainfall impacts, Feb-Dec 2015
- **Forecast event**: GHM summary polygon features from multi-model ensemble representing the area where forecast probabilities exceed 0.4.
- **Observed event**: polygon features representing the location of observed community impacts. Heavy rainfall impact database contains 853 entries, split into impact severity categories (low, moderate, high and disastrous)

Measures intersects between impact polygons and GHM forecast summary polygons
WWRP High Impact Weather Project

Aim: Improve forecasts on timescales of minutes to weeks and enhance their utility in social, economic and environmental applications.
Example: Flood hazard and its impacts

Rain → Snow Melt → Run-off → Drains → River → Breach/Overtopping

Landslide → Dirty water in the “wrong” place

Depth → Velocity → Dirt/Trash → Pollutants → Contamination

Buildings damaged → Contents destroyed → Transport blocked → Drowning → Water/energy interrupted

Occupants displaced → Food/fuel shortage

Injury → Distress → Illness

Death → Recovery cost

Pressure → Storm surge → Ocean Waves

Wind → Breach/Overtopping

Courtesy Brian Golding
Weather information value chain

COMMUNICATION PROCESSES

Service Production

Weather Climate Water

VALUE-ADDING PROCESSES

Basic & Specialised Services
NMHS & Commercial Providers

User Decisions & Actions

Outcomes

After WMO 2015
HIWeather challenges for user-oriented evaluation

- Appropriate verification methods for temporal and spatial high impact weather forecasts (high resolution ensembles, extremes, nowcasts, warnings, downstream hazards, etc.)
- Use social media and non-standard data to evaluate hazards, impact, response
- Build users’ trust by informing about good and bad forecasts, and user-focused verification approaches
- Entrain social scientists to help understand the decisions made in response to high impact weather and associated hazards
- Evaluation of the weather information value chain
- Quantify the socio-economic benefits of high impact weather forecasts, including identifying avoided losses
Final remarks

• Enormous progress in recent years in improving methods for verifying high impact weather
  – Spatial / diagnostic verification approaches now mainstream
  – New methods for verifying rare extreme events
  – Simple approaches appropriate for communicating with some users
  – Need more work on timing verification

• Observations of high impact weather remains a challenge
  – Unconventional observations getting more uses
  – Methods for dealing with observation uncertainty are in development

• WWRP High Impact Weather project is encouraging user-oriented evaluation of impacts and whole value chain
Levels of user focus

**Level 0**: Conventional *measures-based approaches*

- Best for administrative purposes

**Level 1**: Broad *diagnostic approaches*

- Evaluate variables of interest to users
- User-selectable information (stratifications, thresholds)
- Often graphical
- Confidence intervals

Courtesy Barb Brown
Levels of user focus

**Level 2: Features-based and enhanced diagnostic approaches applied**
- Evaluation of **multiple attributes** of broad interest to users

**Level 3: User-specific approaches and measures**
- Interact closely with users to determine meaningful approaches and measures
- May include specialized datasets that are user-specific

**Level 4: Forecast value estimated**, making use of user-focused verification information
- Close interaction with users
- Deep understanding of users’ decision-making and applications of forecasts

Courtesy Barb Brown