Verification and evaluation of weather and weather-impact forecasts for aviation applications: Some principles and resources

Herbert Puempel
hpuempel@gmail.com

Barbara Brown
bgb@ucar.edu

AvRDP Training Workshop October 2018
Hong Kong, China

Outline

- Goals
- Definition and purposes of verification
- User-relevant verification:
  - Translating weather information into potential impact for verification/validation
  - Identifying questions to be answered
- Relevant methods of verification/evaluation for aviation applications
- Stratification
- Resources and tools

Goals

- To discuss appropriate approaches for evaluating forecasts of aviation *weather* and aviation-*potential impact* forecasts
- To consider various nuances of verification including
  - Translation of weather information to potential impact
  - Methods appropriate for different types of forecasts
  - Requirement for “good” observations of weather and impacts
  - Value of stratification of forecasts to obtain useful information about forecast performance in different scenarios
- To provide guidance on additional resources on verification methods and approaches

What is verification?

Verify: **ver·i·fy**
Pronunciation: 'ver-&-"fI
1 : to confirm or substantiate in law by oath
2 : to establish the truth, accuracy, or reality of <*verify* the claim>
synonym see **CONFIRM**

- Verification is the process of comparing forecasts to relevant observations
  - Verification is one aspect of measuring forecast ***goodness***
- Verification measures the ***quality*** of forecasts (as opposed to their ***value***) by quantifying differences between fcst and obs
- For many purposes a more appropriate term is "***evaluation***"

Why verify?

- Administrative purpose
  - Monitoring performance and building trust
  - Choice of model or model configuration (has the model improved in relevant criteria?)
- Scientific purpose
  - Identifying and correcting model flaws >
  - Forecast improvement
- Economic purpose
  - Improved and objective decision making process
  - “Feeding” decision models or decision support systems

Why verify? … some other examples

- Help operational forecasters understand/mitigate model biases and select appropriate models for use in different conditions
- Help "users" interpret forecasts (e.g., "What does a wind speed forecast of 20 knots really mean?")
- Help users to select the most appropriate and reliable content of forecast information
- Identify forecast weaknesses, strengths, differences

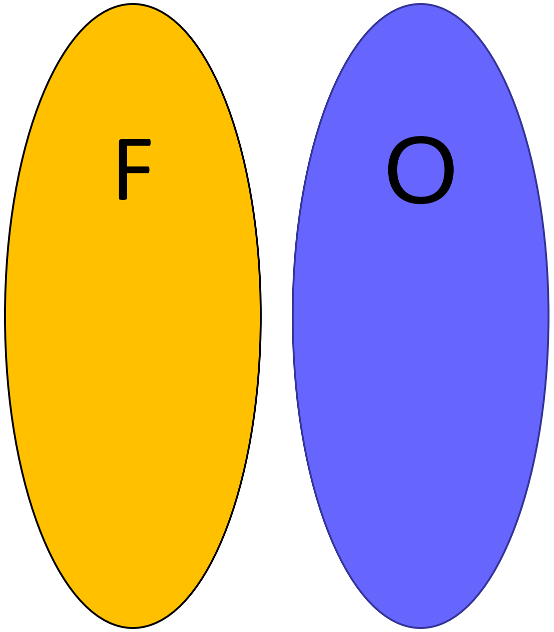
Impact forecast verification

- Translation of weather forecast information into potential weather impact can support informed *decisions* made by forecast users
  - Requires *understanding of operational decision processes based on multiple input parameters*
  - Requires *good (objective) observations* of the impact variable
- Note: Methods for evaluation/verification of impact forecasts are the same as methods for evaluation of the weather forecasts

Identifying verification goals

- What *questions* do we want to answer?
    - Examples:
        - In what locations/seasons/conditions does the forecast provide the most useful information?
        - Are there weather regimes in which the forecasts (predictability) are better or worse?
        - Is the forecast well calibrated (i.e., reliable)?
        - Do the forecasts correctly capture the full variability of the weather or impact even in severe/extreme cases?
        - How far out does the forecast provide useful information for decision making (usable lead time)
        - Can we predict the likely accuracy/reliability of individual forecasts?

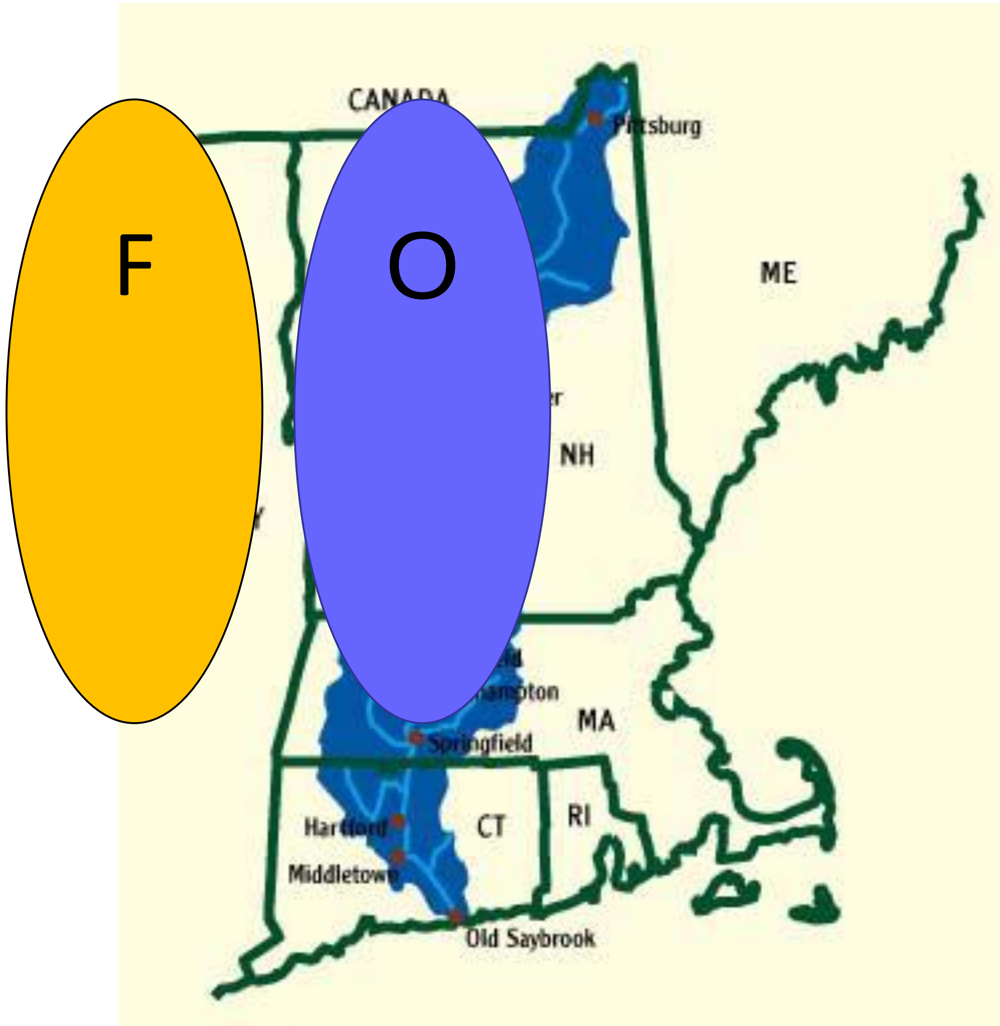
Good forecast or bad forecast?



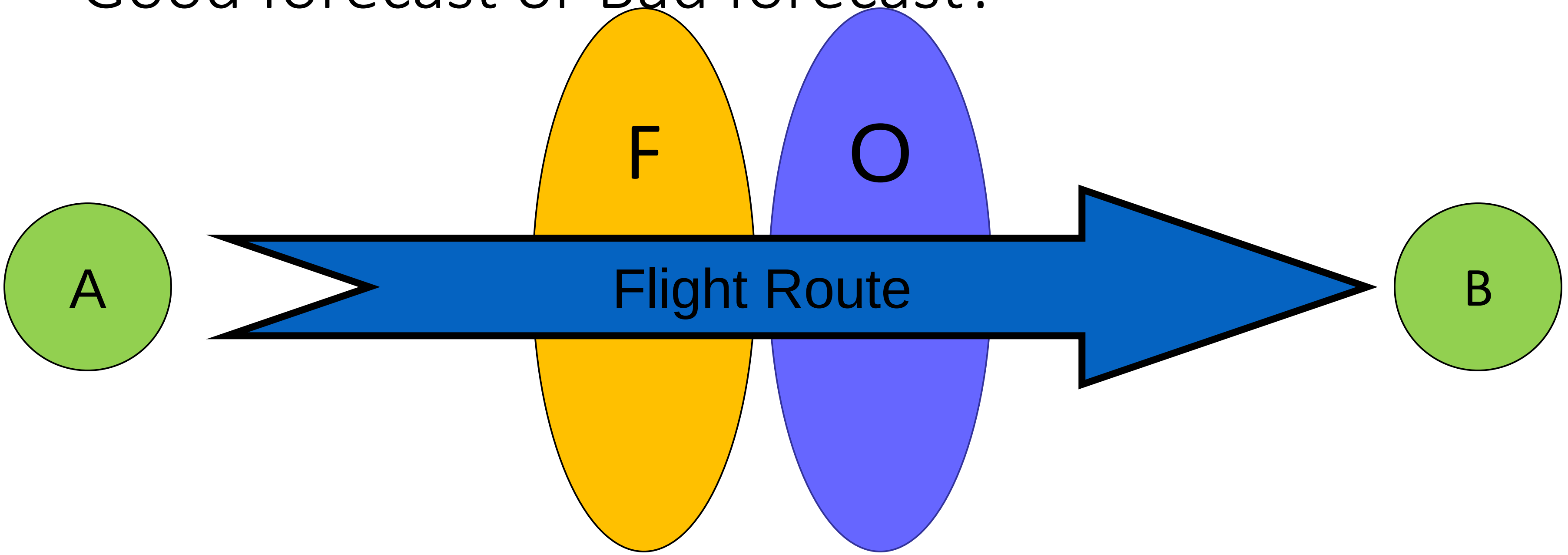
Many verification approaches would say that this weather forecast of a SIGWX area has NO skill and is very inaccurate.

Good forecast or Bad forecast?

For a water manager for this watershed, it’s a pretty bad forecast…



Good forecast or Bad forecast?



For a flow manager and the given route…

This will give a good estimate of capacity reduction

Different users have different requirements!

Different verification approaches can measure different types of "goodness"

Benefits of evaluating impact forecasts

- The example in the past few slides illustrates the benefit of using information from users to evaluate forecasts via *translation of a weather forecast into a potential impact forecast*
- For aviation, the example forecast correctly indicated there would be a *capacity reduction* along the proposed route, with a *location error*. Thus, the verification of the impact forecast is the following:
  - Correct forecast of an event occurrence
  - Correct forecast of the size of the impacted area on given route
  - Incorrect forecast of location/timing of the event
  - Early deviation of trajectory would result in increased delay

Selecting verification methods

- Selection of appropriate methods depends on
  - Type of forecast and observation
    - Continuous (e.g., wind speed, tropopause height, flight time)
    - Categorical (e.g., convective weather event, route blockage)
    - Probabilistic (e.g., probability of landing cross- winds exceeding a threshold)
    - Spatial (e.g., location and size of convective event, route blockage)
    - Questions of interest for decision makers
  - Verification <u>*attributes*</u> that can answer the questions
    - Attributes measure different aspects of forecast quality
    - <u>Examples</u>: Bias, correlation, accuracy, discrimination

Some key things to think about …

Who…

…needs to know?

What…

… do different stakeholder worry most about?

… kind of parameter are we evaluating? What are its characteristics (e.g., continuous, probabilistic)?

… thresholds (regulatory/operational) are important (if any)?

… forecast resolution is relevant (e.g., site-specific, area-average)?

… are the characteristics of the obs (e.g., quality, representativity, uncertainty)?

… are appropriate methods?

Why…

…do we need to verify it?

Some key things to think about…

How…

…do you need/want to present results (e.g., stratification/aggregation)?

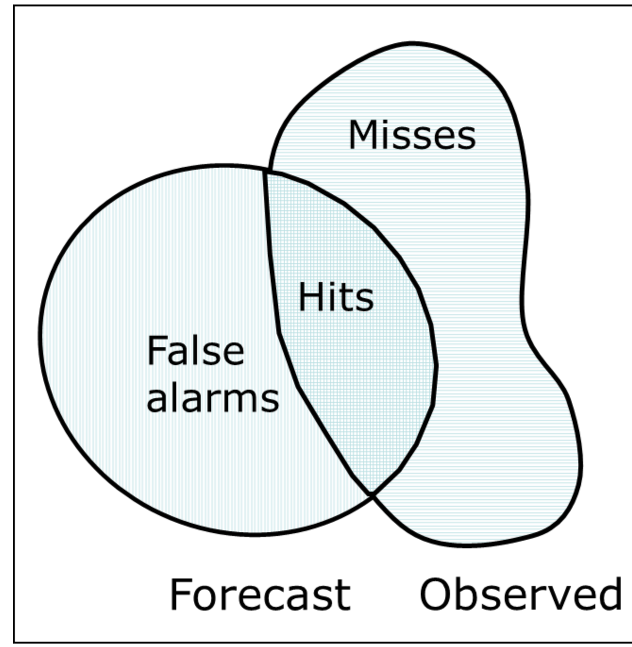
Which…

…methods and metrics are appropriate and understandable by users?

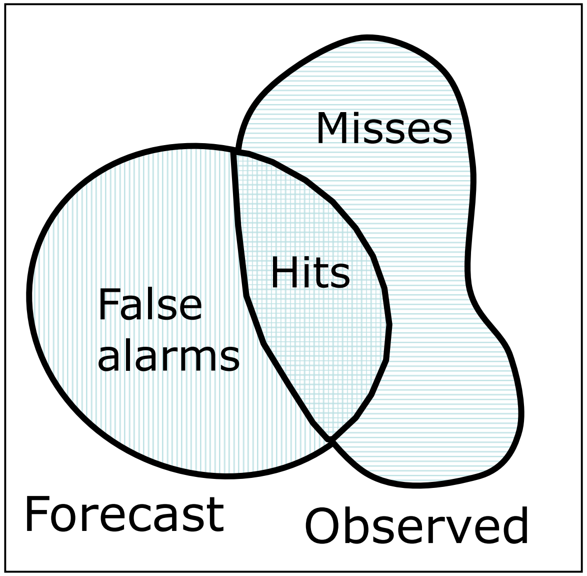
… methods are required (e.g., bias, event frequency, sample size, trending)

Categorical forecasts

- Categorical forecasts are generally Yes/No forecasts
  - "*Yes*" an electric storm will impact an airport from time *t0 to t1*
  - *"Yes/No"* a route will be blocked at time *t*
- Also may be related to an "exceedance"; for example:
  - "*Yes*" the storm will sit over a runway for 3 hours or more
  - "*Yes*" more than *X* flights will be affected



Categorical verification methods



Basic methods:

(1) Create contingency table by thresholding forecast and observed values for variable of interest, and counting forecast/observed pairs for each *cell* in the table

(2) Compute a variety of scores from the counts in the contingency table:

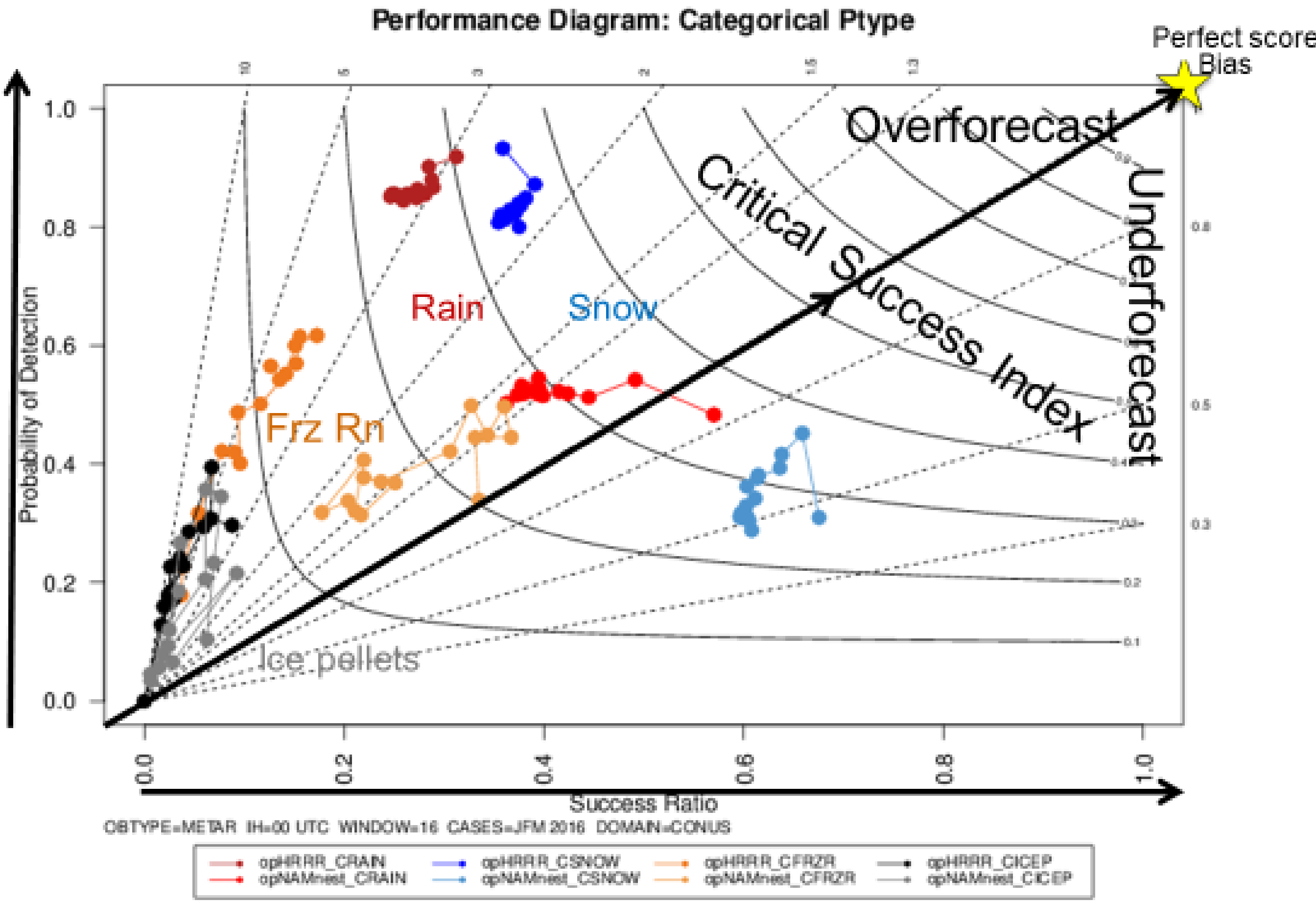
- *Probability of Detection* (POD) (measures ability to capture events)
- *False Alarm Ratio* (FAR) (measure of over-forecasting)
- *Threat score* (measure of overall accuracy taking into account POD and FAR)
- … And ***many*** other scores

Contingency Table

Forecast	Observed: yes	Observed: no
yes	*hits*	*false alarms*
no	*misses*	*correct negatives*

Could be applicable to precipitation, convection, route blockage, etc

Perfect forecast requires exact overlap!

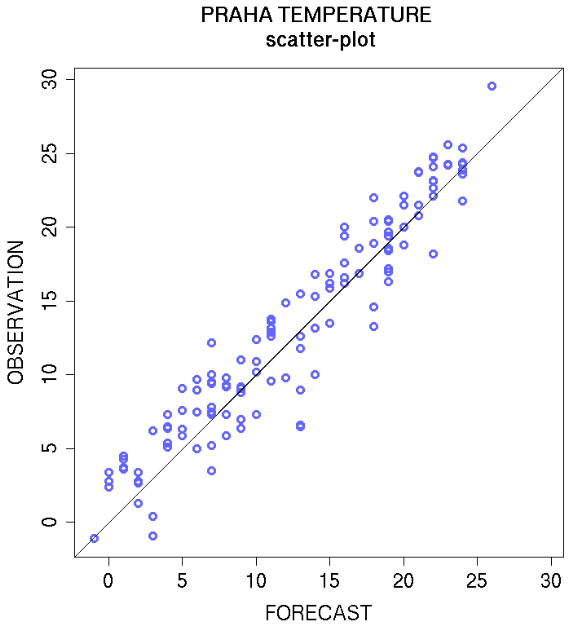


"Performance" diagrams allow display and comparison of several categorical verification measures simultaneously

Figure shows verification results for two models predicting precipitation type (inflight icing application)

Methods for continuous forecasts

- For continuous forecasts, forecast values at specific points are mathematically compared to observed values
  - Example: Flight level wind speed
- Many scores can be computed to measure a variety of attributes of interest



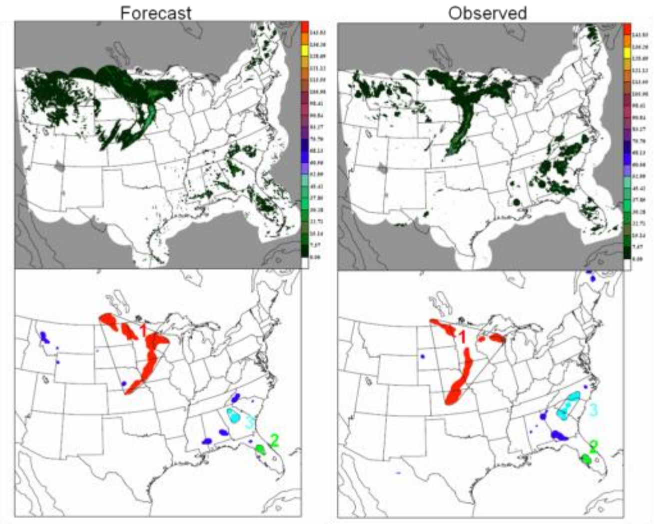
Continuous variable scores

- **Bias**: Average of errors (differences between forecast and observed values); also called Mean Error (ME)
  - Measures the "direction" of the error
  - Could be difference between two opposite errors in sample
- **Mean squared error (MSE):** Average of squared differences between forecast and observed values
  - Strongly penalizes large errors
  - Often presented as the square root of MSE (RMSE)
- **Mean absolute error (MAE):** Average of absolute values of differences between forecast and observed values
  - Less emphasis on large errors
- **Correlation**: Measures linear association
  - Ignores bias
  - Can be misleading
  - Penalizes higher resolution of forecasts

***Note**: Bias and MSE are not independent; an increased Bias leads to an increased MSE*

Methods for spatial forecasts

- Spatial verification methods have been developed to
  - Cope with the fact that ***Good forecasts may not require perfect overlap with the observed area (e.g., our route example)***
  - Provide diagnostic (user-relevant) information about forecast performance
- For example, spatial methods can answer questions such as:
  - *Was the warned region too big? Located in the correct place?*
  - *Was the route blockage in the latitude/longitude predicted?*
  - *Are there gaps in a CB area allowing flights to pass thru?*



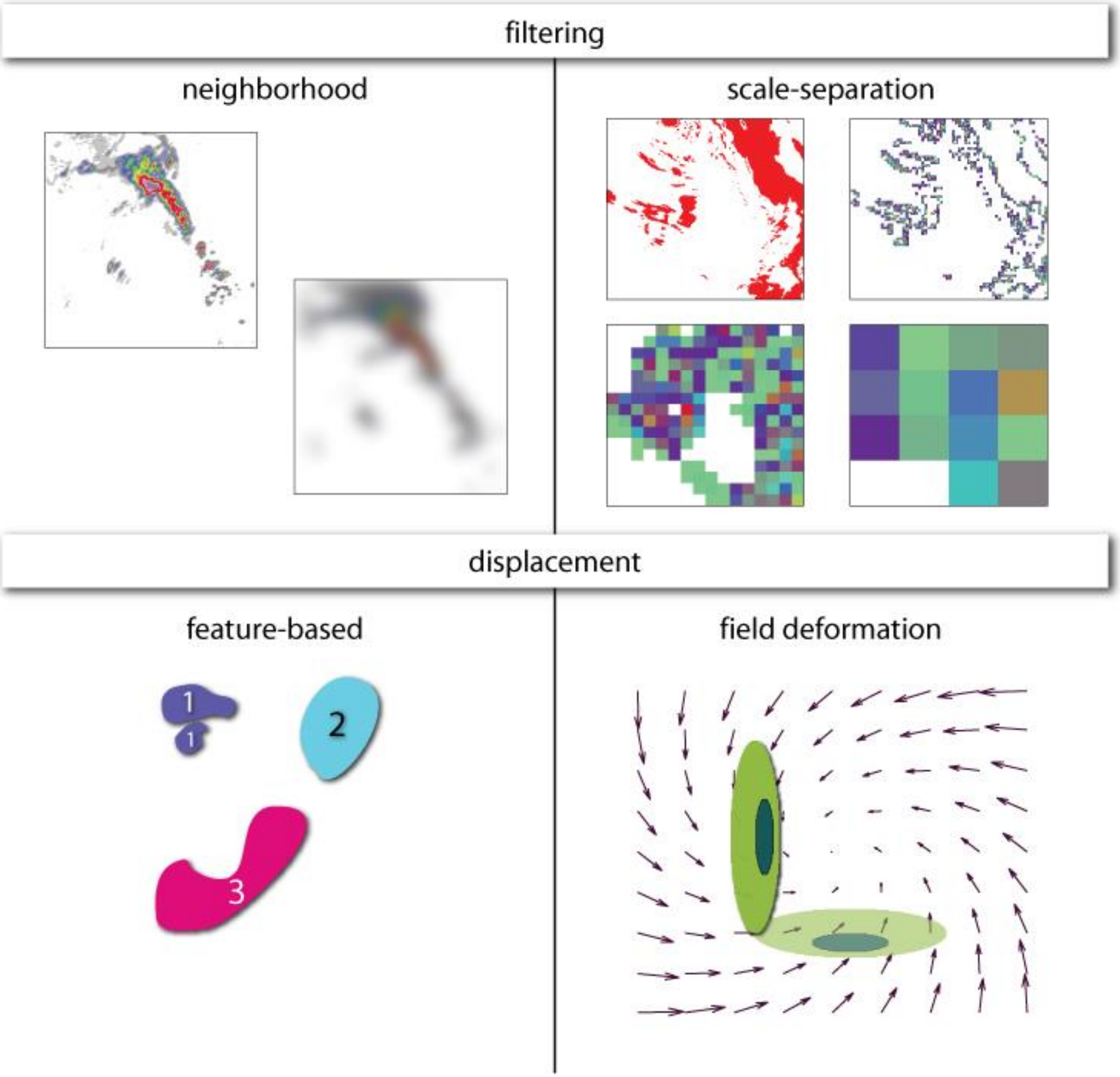
MODE example 2008

MODE identifies and compares characteristics of "objects" in the forecast and observed fields

Spatial Verification Approaches

To address limitations of traditional approaches, a new set of spatial verification methods have been developed

Goal is to provide more useful information about forecast performance



Methods for probabilistic forecasts

- Why probability forecasts?
  - Probabilistic forecasts provide useful information for decision-making, especially via automated decision-making systems (e.g., for routing decisions, fueling, etc.)
  - Reliable probabilistic forecasts can have greater economic value than non-probabilistic
  - Require adequate sample sizes
  - Need to “educate” users?
- Verification of probability forecasts involves measurement of
  - Accuracy
  - Reliability
  - Discrimination / Resolution

Measures for probabilistic forecasts

- Accuracy

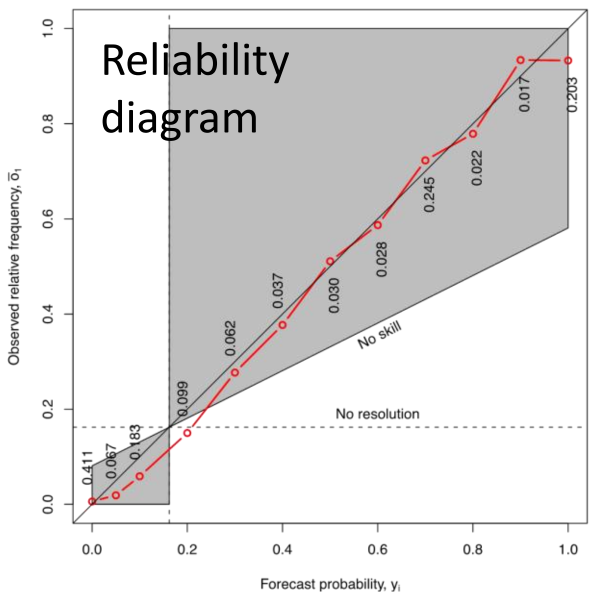
  Brier score: Average of squared differences between forecast probability and occurrence / non-occurrence of forecast event (like a MSE for probabilistic forecasts)

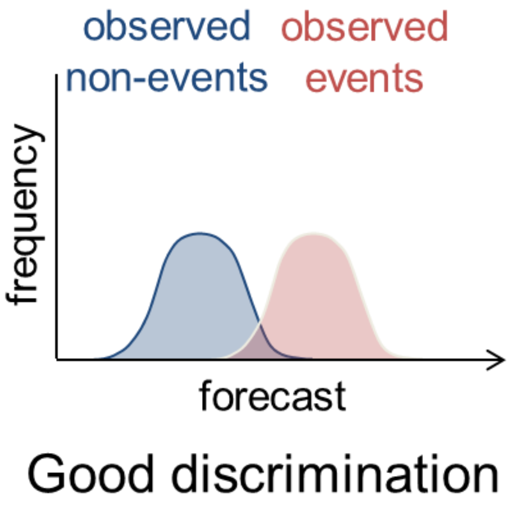
- Reliability

  Measures whether the frequency of an event occurring matches the probability forecast

- Discrimination

  Measures how different the forecasts are for occurrences and non-occurrences of the forecast event





Stratification of forecasts

- Meaningful verification depends on examining *homogeneous subsets* of forecasts
  - *Examples*: Categorization by season, event type (frontal passage, widespread convection, extreme vs. non-extreme)
- It is possible to arrive at meaningless results unless data are correctly stratified!
  - *Example*: Combining forecasts from winter and summer can lead to good results simply because a forecasting system is able to correctly forecast the climatology for winter and the climatology for summer (i.e., there may be no skill within winter or within summer)
- *However*: The need/desire for stratification must be balanced against the need to have an adequate sample size
  - Small samples can lead to erratic and inconsistent results (lack of robustness)

More stratification…

- Gives better estimate of expected accuracy in a given situation
- Maximizes achievable benefit when done properly
  - Example: IF situation of type „A“ is highly predictable, Situation of type „B“ fairly unpredictable,
  - then overall score would lead to missed benefits by <u>**under-use**</u> of forecast in cases A, and <u>**disappointment/ loss of confidence/ negative impact**</u> in cases B

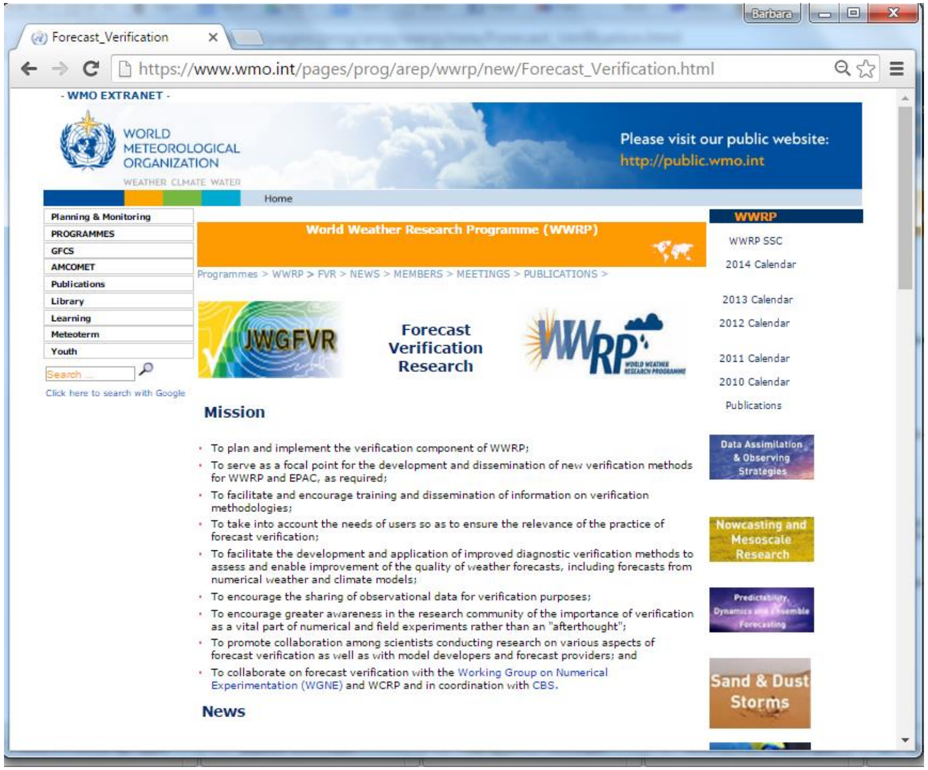
Summary

- Designing verification requires clear understanding of the *attributes* that are of interest and identification of the appropriate *methods* for measuring them
  - First step: Determine what questions need to be answered
- Verification is *multi-dimensional*: More than one measure is needed to provide a meaningful evaluation of a forecast!
- Careful *stratification* can provide the most useful information for decision-making
- Many *resources* are available as guidance for designing verification studies

Resources

Joint Working Group on Forecast Verification Research

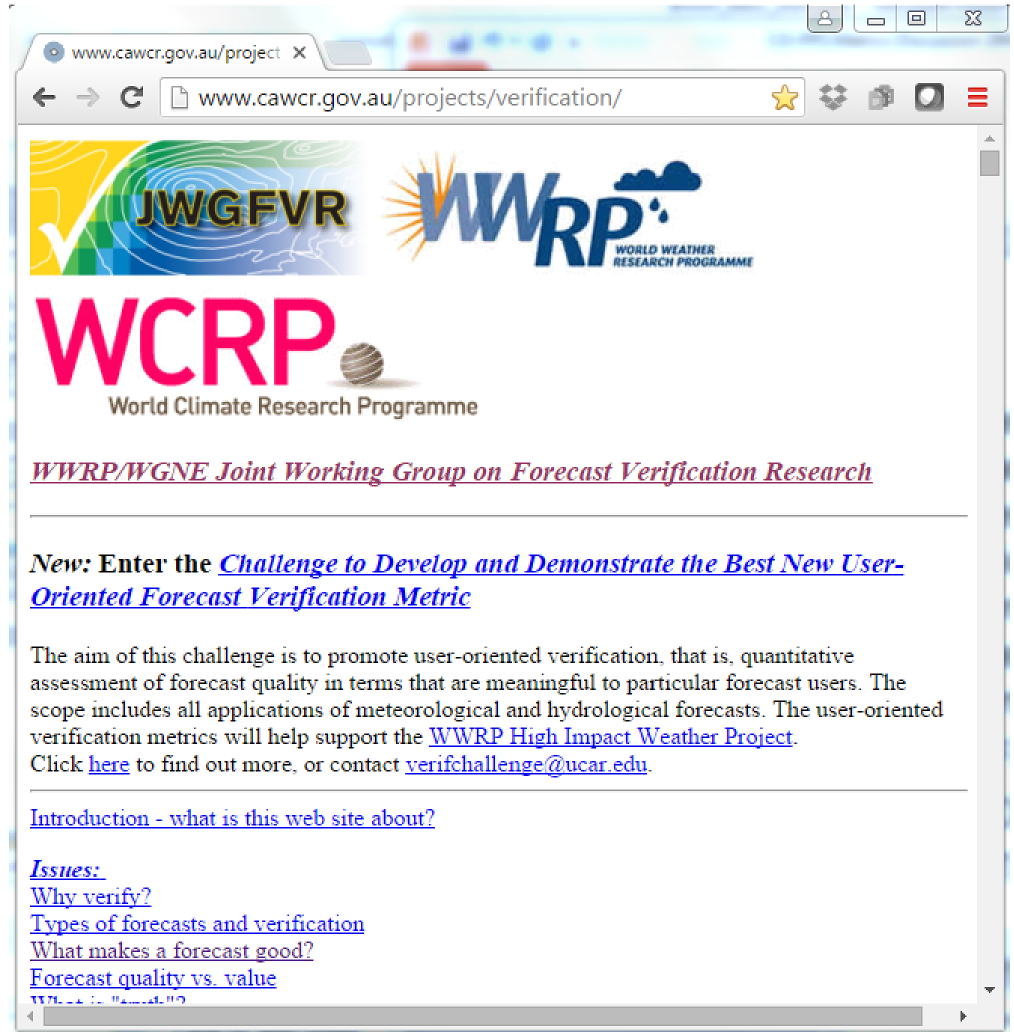
- Supports working groups and projects in WWRP and WGNE on verification topics
- Conducts and coordinates research on new verification methods (e.g., MesoVICT; https://www.ral.ucar.edu/projects/icp/)
- Workshops and tutorials



Resources

Web page with many links to presentations, articles, etc. from international community

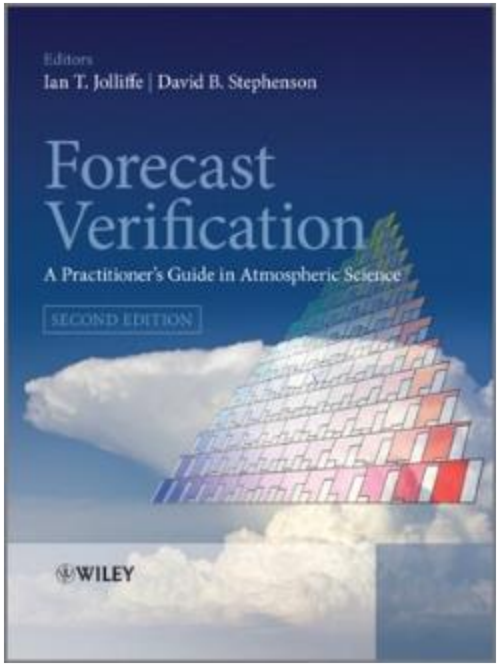
- FAQs
- Definitions
- Tools

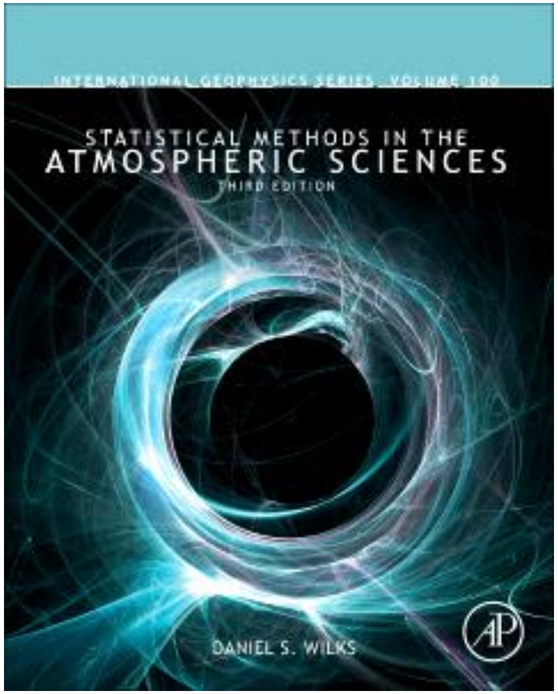


http://www.cawcr.gov.au/projects/verification/

Resources - Books

- Jolliffe and Stephenson (2012): *Forecast Verification: a practitioner's guide*, Wiley & Sons, 240 pp.
- Stanski, Burrows, Wilson (1989) *Survey of Common Verification Methods in Meteorology* (available at http://www.cawcr.gov.au/projects/verification/)
- Wilks (2011): *Statistical Methods in Atmospheric Science*, Academic press. (*Updated chapter on Forecast Verification*)





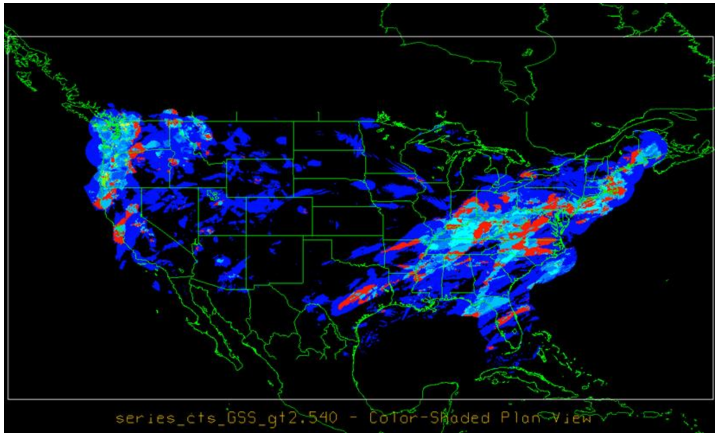
Resources

- Eric Gilleland's web page on spatial verification methods:
  http://www.ral.ucar.edu/projects/icp/
- Verification Issues, Methods and FAQ web page:
  http://www.cawcr.gov.au/projects/verification/
- EUMETCAL learning module on verification methods
  http://www.eumetcal.org/-learning-modules-

Tools for Forecast Evaluation

- Model Evaluation Tools (MET)
  - Includes Traditional approaches, Spatial methods (MODE, Scale, Neighborhood), Confidence Intervals Ensemble methods
  - Supported to the community (freely available)

http://www.dtcenter.org/met/users/



Spatial distribution of Gilbert Skill Score

- R libraries
  - Verification
  - Spatial-Vx
  - R is available at https://www.r-project.org/