# AvRDP verification considerations

Barbara Brown (NCAR, Boulder, USA; bgb@ucar.edu)

With significant contributions from Peter Li and Herbert Puempel

*AvrRDP final meeting: Pretoria, South Africa*

*21 August 2019*

# Outline

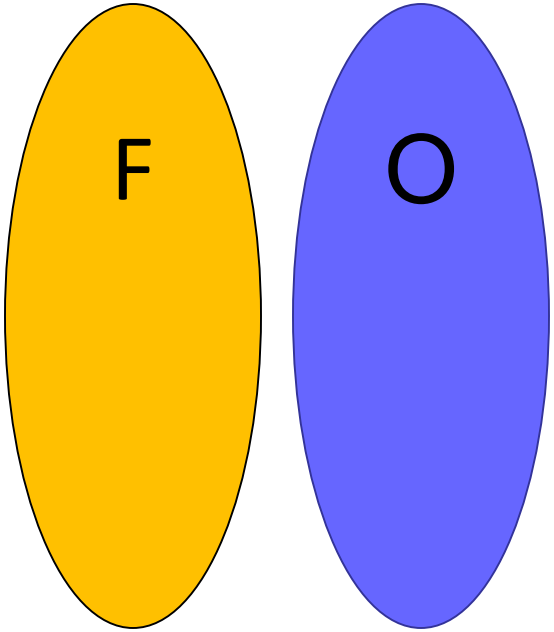
- Users of verification
- Good forecasts and bad forecasts…  User-relevant verification
- Identifying verification methods for different forecasts and purposes
- Simplifying verification results (Score cards)
- Resources

# Who are the users of verification information?

- Forecast developers
  - Calibrate and improve forecasts
- Air Traffic Managers
  - Determine the "usefulness" of products for decision-making
    - Based on information about forecast *quality* (calibration, skill)
  - Make decisions based on the forecast information, with known qualifications based on verification information
- Funders
  - Are the forecasts doing what they say they will do?
  - Should we invest more in their improvement?

**Note**: *"forecasts" and "verification information" can pertain to weather forecasts or to weather information translated into air traffic metrics*

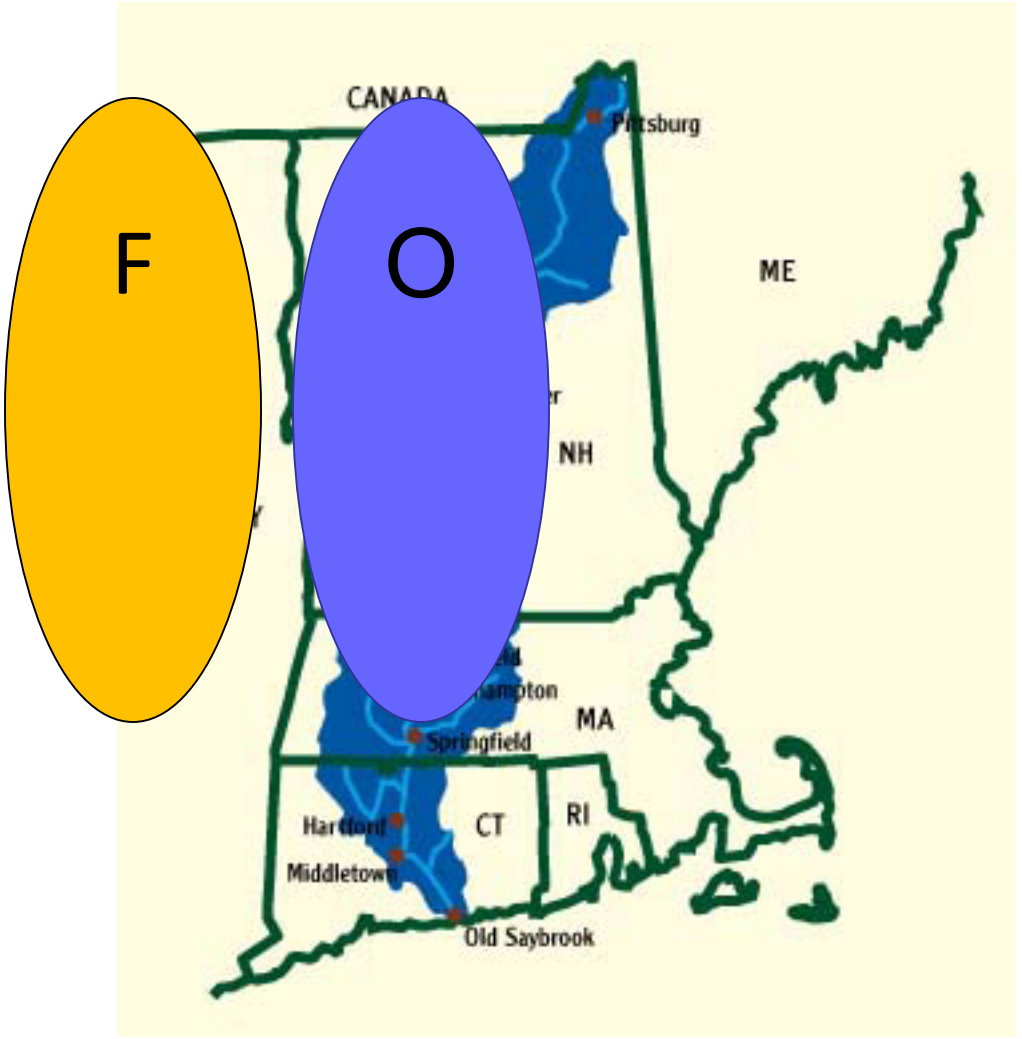
# Good forecast or bad forecast?



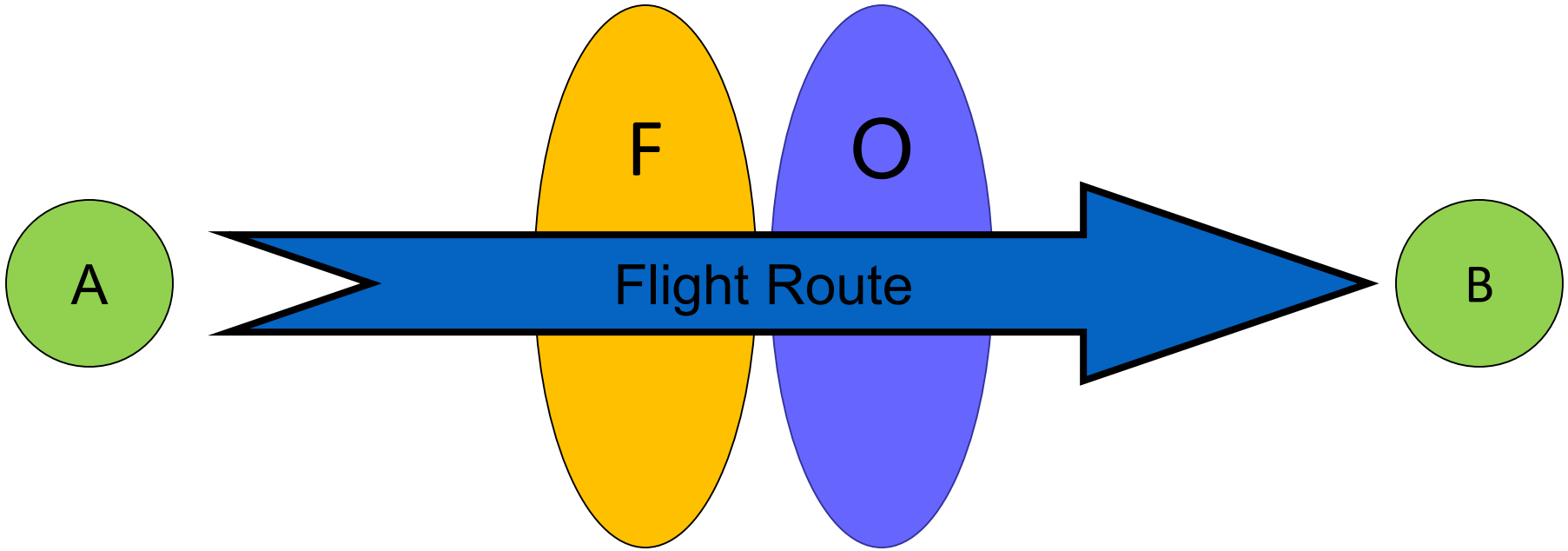

**Many verification approaches would say that this weather forecast of a significant weather area has NO skill and is very inaccurate.**

# Good forecast or Bad forecast?

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



# Good forecast or Bad forecast?



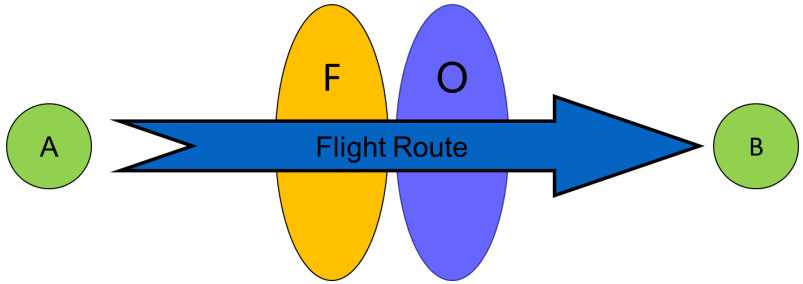

For a flow manager and the given route…

Different users have different requirements!

This will give a good estimate of capacity reduction

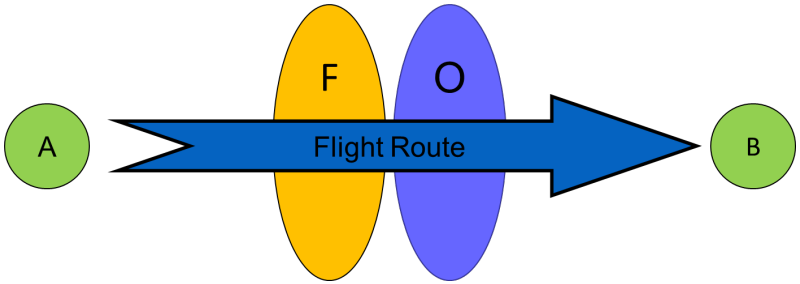
Different verification approaches can measure different types of "goodness"

# How can we evaluate this forecast in a meaningful way?




- As a weather forecast?
  - Traditional approach would indicate it has NO skill
    - ***No overlap*** between forecast and observed area
  - Alternative approaches: Spatial methods
    - *Neighborhood methods => Forecast has some skill because it is in the "neighborhood" of the observed region*
    - *Distance methods => Measure overall distance between the forecast and observed points*
    - *Object-based methods => Answer meaningful (physical) questions such as*
      - What is the *distance* between the forecast and observed areas (e.g., centroid difference)
      - Is the *area covered* by the forecast the same size as the area covered by the observed storm?
      - Is the *orientation* of the forecast correct?
      - …

# How can we evaluate this forecast in a meaningful way?



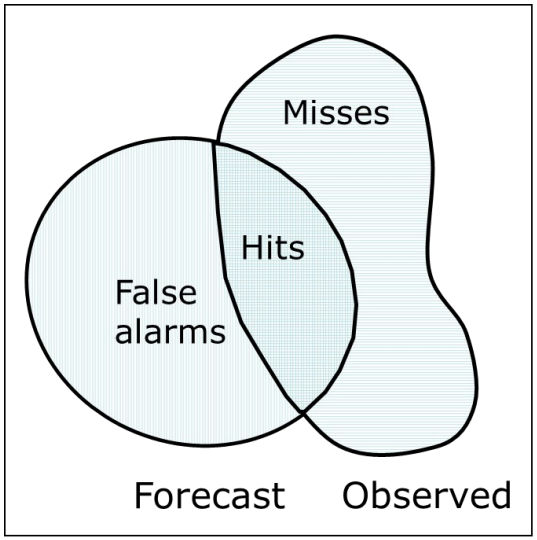

- As an ATM forecast (*translated* from the weather forecast)?
  - Estimate expected traffic flow/capacity (or delays etc.) if forecast is correct => *Translated forecast*
    - How many flights would be able to get through if the forecast is correct?
    - How many delays are expected if the forecast is correct?
  - Measure observed traffic flow/delays
    - Compare to expected flow
    - How well did the forecasted flow match the observed flow?
    - *Note*: Need to be able to take into account (and separate out) other factors that might lead to diversions, reduced flow, delays

# Forecast verification methods

- Pertain to *any kind of forecast*
  - Weather/climate
  - Medicine
  - Economics
  - **ATM impacts – enroute/terminal etc.**
- To evaluate weather forecasts or forecasts of ATM impacts requires good ***observations*** of the weather/impacts
  - Reliable measurements
  - Understanding of uncertainties
- Specific verification approaches are required …
  - For different types of forecasts/observations
  - To answer different types of questions (*Are the forecasts reliable? Do they have skill over other methods?*)
- Identifying the availability of ***observations*** and the ***questions*** to be answered are ***critical first steps***! (for both weather and impact forecast verification)

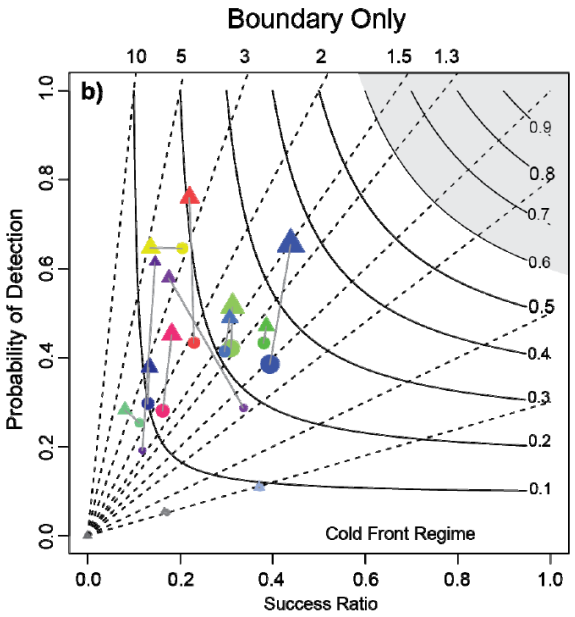
# Categorical forecasts and observations

- Typically these are 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



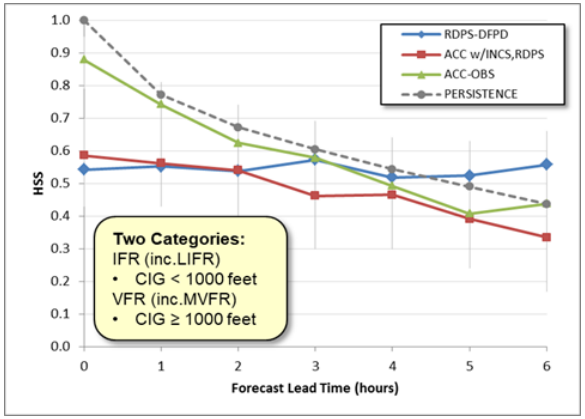
# Categorical forecast examples

- Categorical statistics include
  - POD (Probability of Detection)
  - FAR (False Alarm Ratio)
  - CSI (Critical Success Index)
  - ETS (Equitable Threat Score)
  - HSS (Heidke Skill Score)




*Example Performance Diagram for convective initiation forecasts (from Roberts et al. 2012; Weather and Forecasting)*

- Can be applied to yes/no decision making
  - Ex: Closing approach route due to convection




*CYYZ (Toronto) verification; (Nov 2015 – Mar 2016). Credit: Janti Reid*

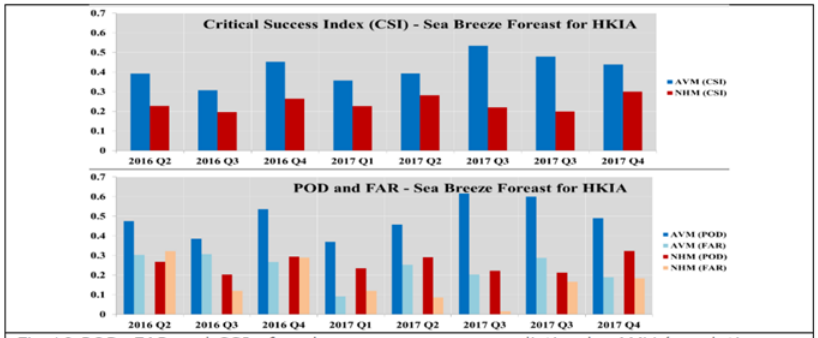
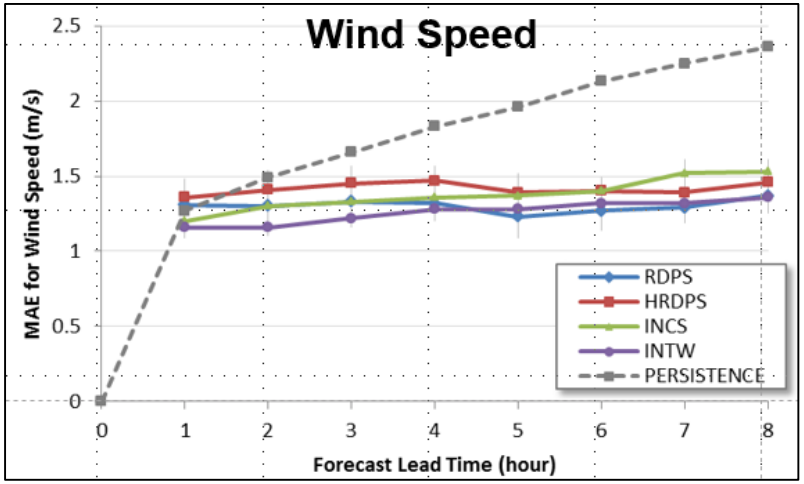



Fig.10 POD, FAR and CSI of seabreeze occurrence prediction by AMV (resolution 200m) and NHM (resolution 2km)

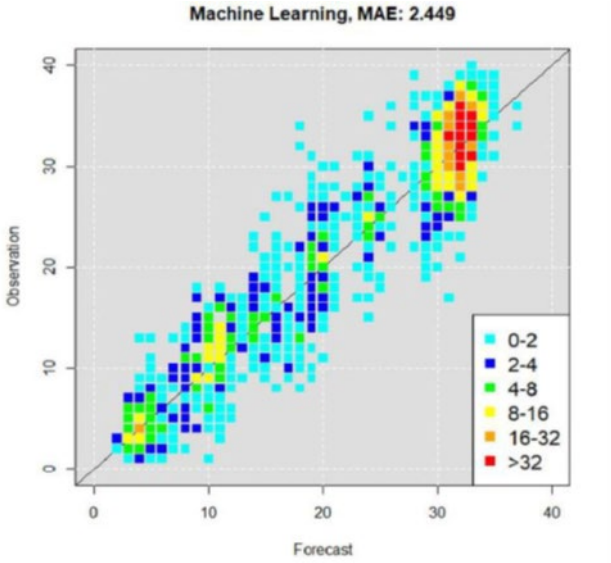
*From HKO Final report for AVRDP (2019)*

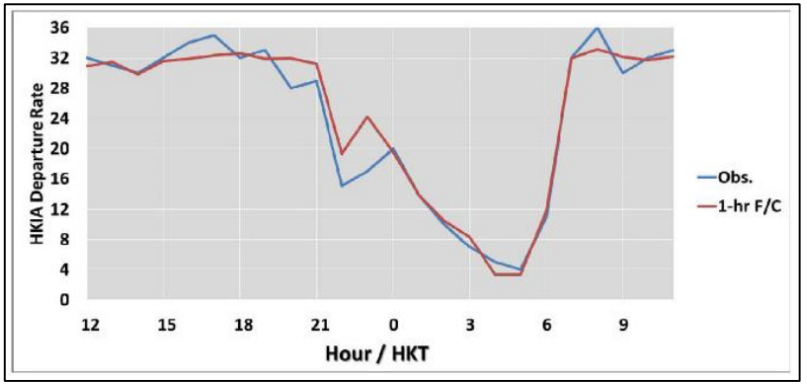
# Continuous forecasts

- Comparison of continuous forecast and observed values (e.g., wind speed)
- Examples: Root Mean Squared Error (MSE), Mean Absolute Error (MAE), ME (Mean Error or Arithmetic Bias)
  - Note that scores are inter-related




*CYYZ (Toronto) verification; (Nov 2015 – Mar 2016).*
*Credit: Janti Reid*




***Departure Rate predictions***
***From HKO Final report for AVRDP (2019)***

# 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)

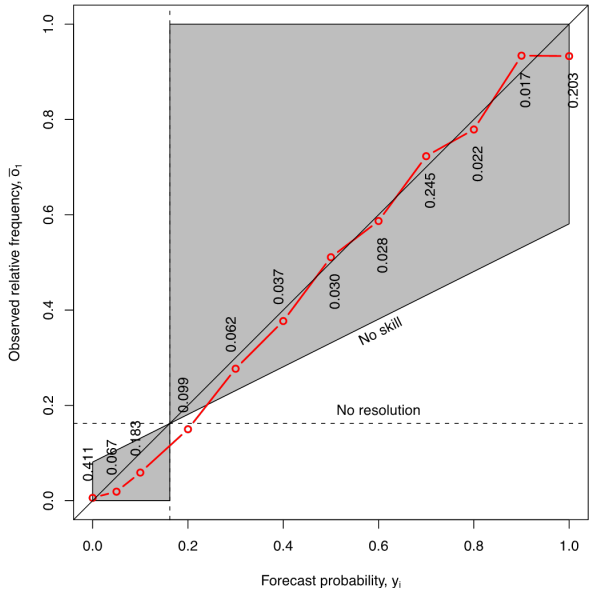
- Reliability

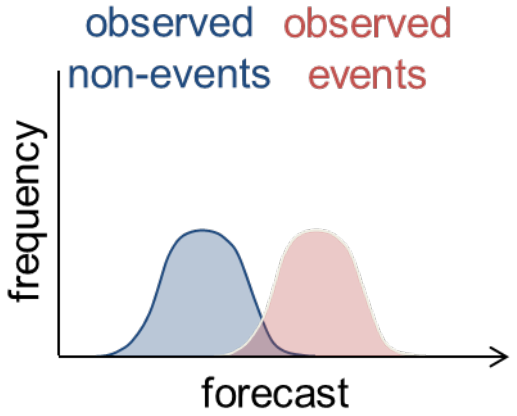
  Measures whether the frequency of an event occurring matches the probability forecast

- Discrimination

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

Reliability diagram




Good discrimination

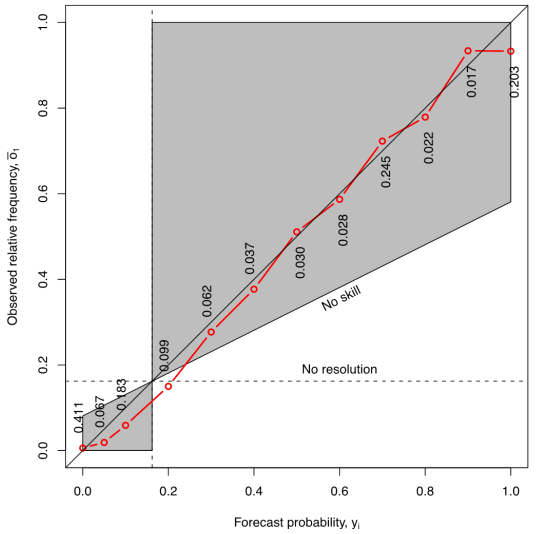
# Recommended approaches for probability forecasts

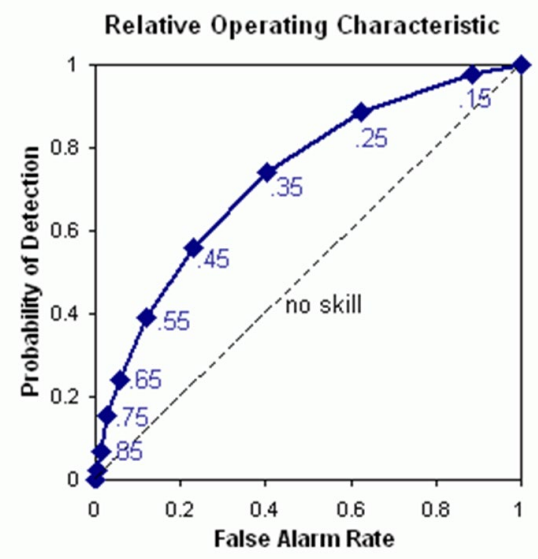
## A good combination of measures:

- *Reliability*: Does the "event" occur approximately as often as predicted?
- *Relative Operating Characteristic (ROC):* How well does the forecast discriminate between events and non-events?
  - Can be translated into a ***potential cost-loss measure***
  - Ignores calibration/reliability

## These two measures provide a "complete" evaluation of probabilistic (2-category) forecasts

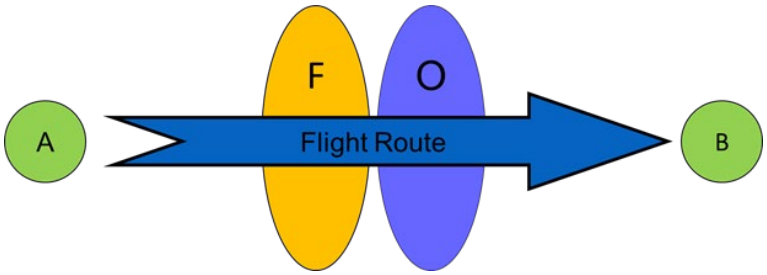
Reliability diagram

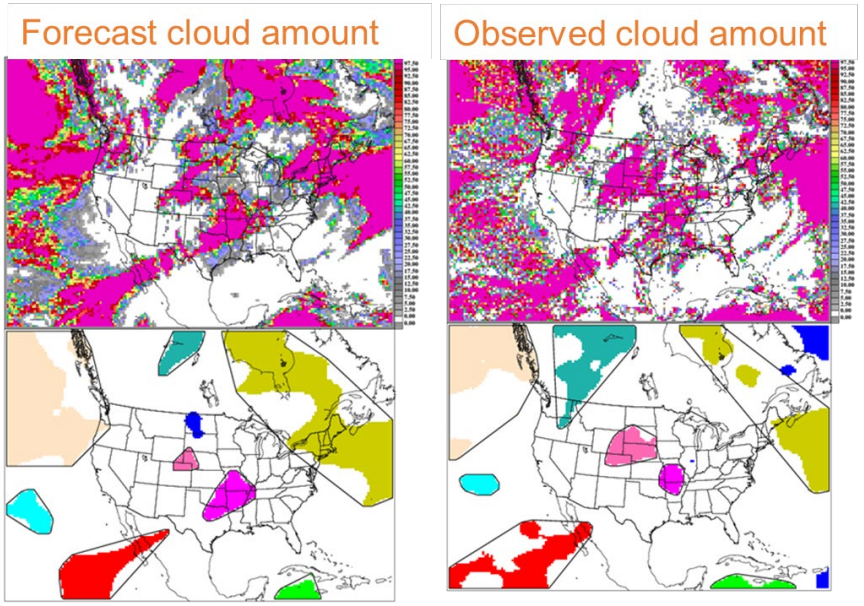


# Spatial approaches

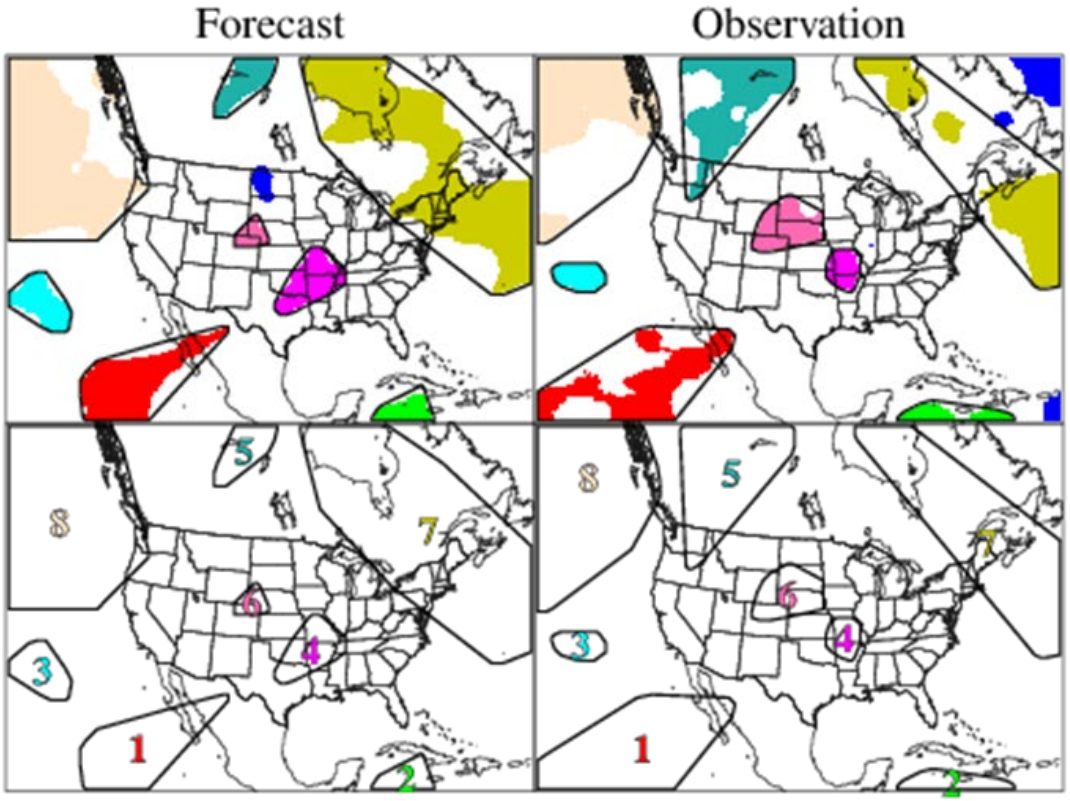
- Provide the opportunity to evaluate characteristics of forecasts that are *directly relevant to users*
  - Mis-placement of convective weather
  - Areal coverage of hazardous weather
  - Intensity of storms
- Several categories of approaches
  - Object-based
  - Neighborhood
  - Scale separation
  - Distance
  - Field deformation
- These approaches have been applied in a variety of studies involving clouds, convection, etc. that are relevant for aviation




MODE object-based approach applied to forecast and observed cloud amount

# Cluster Object Information




- Some displacement of all clusters
- Large area differences, for some objects

… Etc.

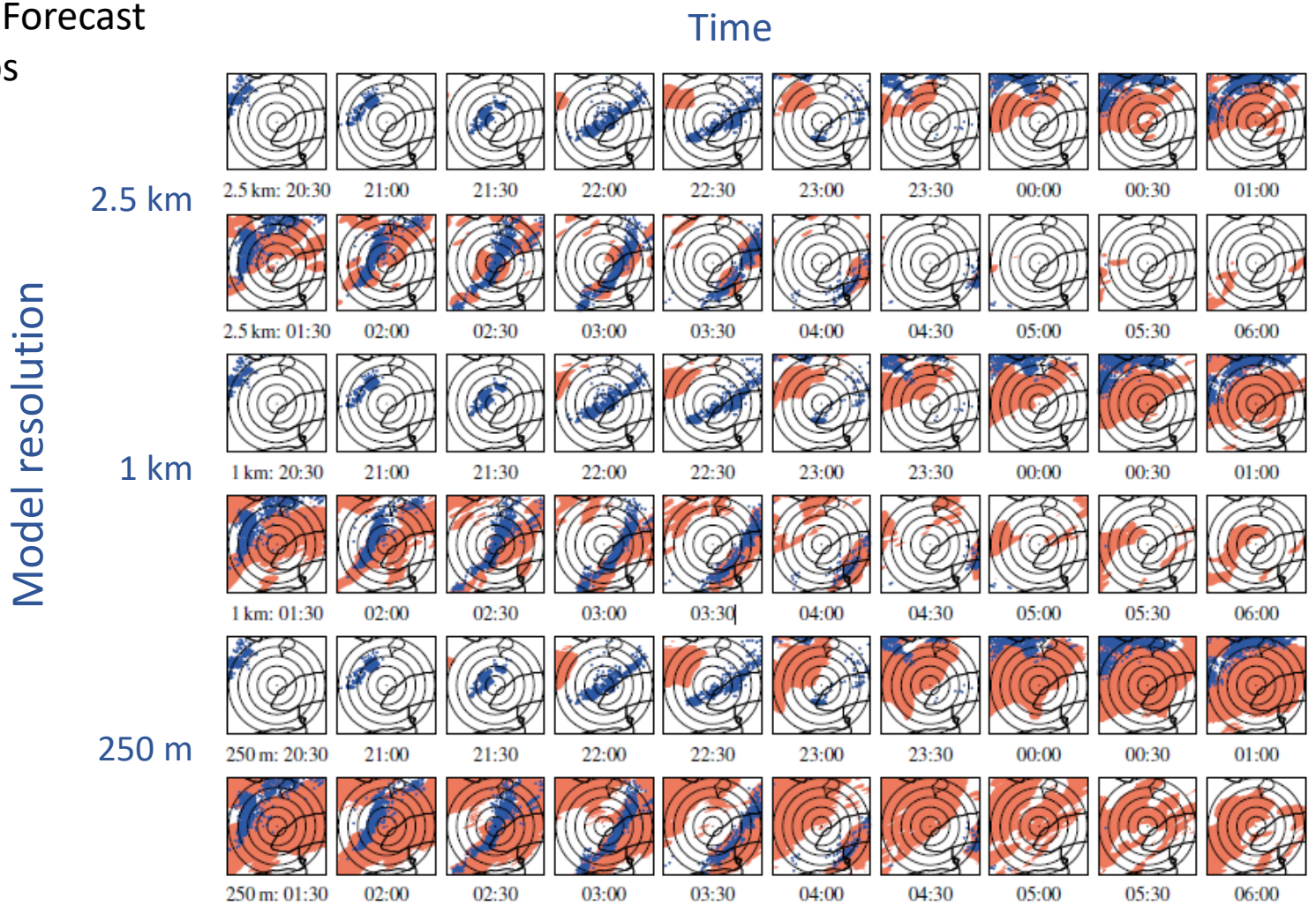
| CLUS PAIR | CEN DIST | ANG DIFF | FCST AREA | OBS AREA | INTER AREA | UNION AREA | SYMM DIFF | FCST INT 50 | OBS INT 50 | FCST INT 90 | OBS INT 90 | TOT INTR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 8.53 | 10.08 | 689 | 816 | 504 | 1001 | 497 | 100.00 | 100.00 | 100.00 | 100.00 | 1.0000 |
| 2 | 6.18 | 10.69 | 131 | 138 | 87 | 182 | 95 | 100.00 | 100.00 | 100.00 | 100.00 | 1.0000 |
| 3 | 9.80 | 35.64 | 247 | 145 | 33 | 359 | 326 | 89.00 | 100.00 | 100.00 | 100.00 | 0.9411 |
| 4 | 4.69 | 51.94 | 299 | 130 | 121 | 308 | 187 | 100.00 | 100.00 | 100.00 | 100.00 | 0.9158 |
| 5 | 16.56 | 13.02 | 229 | 829 | 196 | 862 | 666 | 100.00 | 100.00 | 100.00 | 100.00 | 0.9018 |
| 6 | 3.47 | 19.33 | 81 | 305 | 81 | 305 | 224 | 100.00 | 100.00 | 100.00 | 100.00 | 0.8958 |
| 7 | 11.74 | 2.27 | 2366 | 1049 | 1001 | 2414 | 1413 | 100.00 | 100.00 | 100.00 | 100.00 | 0.9407 |
| 8 | 15.77 | 38.71 | 1921 | 1157 | 773 | 2305 | 1532 | 100.00 | 100.00 | 100.00 | 100.00 | 0.9607 |

# **Example**: Spatio-temporal User-centric Distance for Forecast Verification (Brunet et al. 2018; MetZ); Lightning forecasts

## Measures errors in predicted distances to observed events

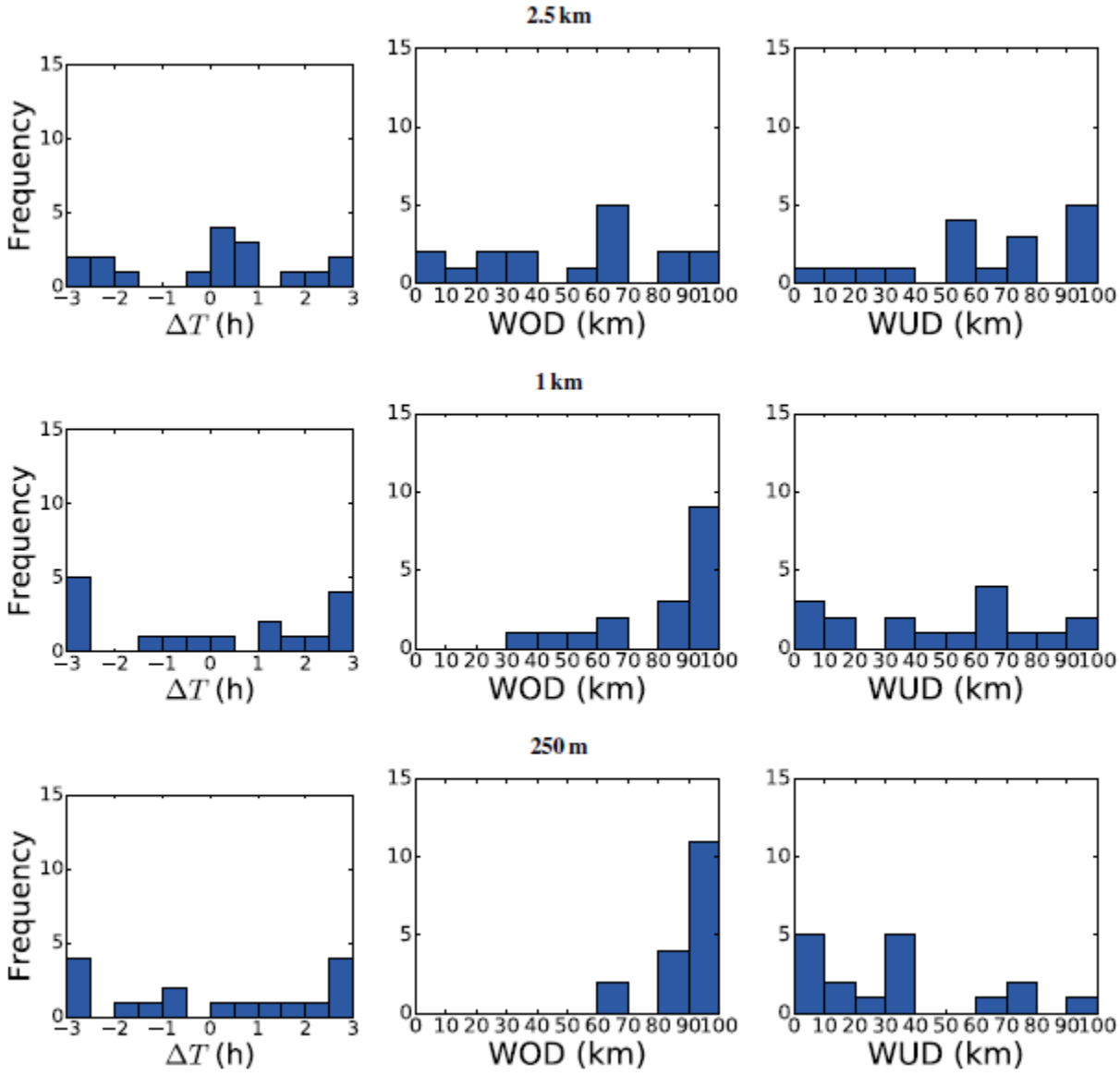
Orange = Forecast

Blue = Obs



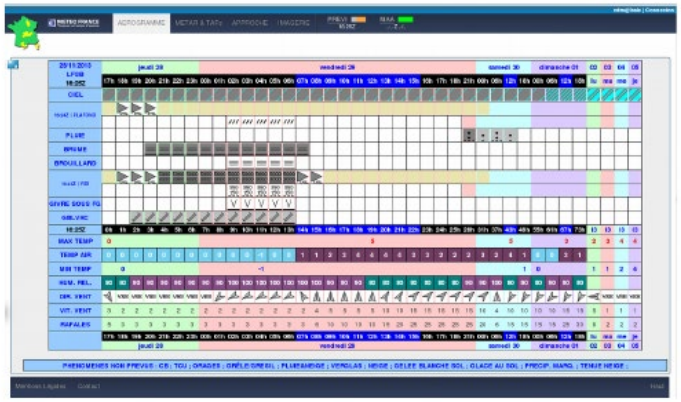
# Summaries of example results (Brunet et al. spatial approach)

- WOD = Worst Overforecast Distance
- WUD = Worst Underforecast Distance



# Simplifying verification information for decision making

- As noted frequently, the number of scores used for verification can be over-whelming and hard to understand
- Hence there is a need for
  - Simpler methods?
  - Combined scores (e.g., NWP or GO index)?
  - Score cards to summarize and clearly define important results



Example forecast display (CDG; credit S. Desbios)

Example TC verification scoring table (NCAR)

| Forecast Hour | | 12 | 24 | 36 | 48 | 60 | 72 | 84 | 96 | 108 | 120 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Atlantic Basin | A | 699 | 649 | 582 | 516 | 455 | 407 | 369 | 327 | 290 | 266 |
| | | -1.3 | -3.2 | -5.8 | -11.3 | -14.7 | -20.1 | -25.0 | -27.1 | -41.4 | -64.6 |
| | | -5% | -7% | -9% | -15% | -16% | -18% | -18% | -16% | -20% | -27% |
| | | 0.903 | 0.959 | 0.968 | 0.990 | 0.987 | 0.971 | 0.935 | 0.879 | 0.878 | 0.932 |
| | B | 671 | 619 | 556 | 494 | 438 | 392 | 355 | 313 | 276 | 249 |
| | | -1.1 | -4.3 | -8.5 | -11.6 | -12.9 | -16.1 | -21.5 | -27.6 | -52.1 | -84.2 |
| | | -4% | -10% | -14% | -15% | -14% | -14% | -16% | -17% | -28% | -40% |
| | | 0.774 | 0.984 | 0.997 | 0.988 | 0.957 | 0.918 | 0.908 | 0.832 | 0.893 | 0.957 |
| | C | 686 | 638 | 569 | 504 | 445 | 400 | 359 | 317 | 282 | 260 |
| | | 3.5 | 6.8 | 8.5 | 7.2 | 6.9 | 4.9 | 4.0 | 7.2 | -8.1 | -29.6 |
| | | 11% | 12% | 11% | 8% | 6% | 4% | 2% | 4% | -3% | -11% |
| | | 0.999 | 0.999 | 0.995 | 0.833 | 0.748 | 0.426 | 0.236 | 0.337 | 0.270 | 0.683 |
| | D | 688 | 640 | 571 | 506 | 449 | 404 | 363 | 319 | 285 | 259 |
| | | 3.6 | 5.2 | 9.7 | 11.6 | 15.0 | 19.5 | 21.4 | 21.8 | 18.1 | 14.7 |
| | | 11% | 10% | 13% | 12% | 12% | 13% | 12% | 10% | 7% | 5% |
| | | 0.999 | 0.999 | 0.999 | 0.999 | 0.999 | 0.997 | 0.993 | 0.904 | 0.713 | 0.518 |
| Eastern North Pacific Basin | A | 421 | 392 | 339 | 287 | 244 | 205 | 170 | 134 | 109 | 87 |
| | | -1.1 | 1.7 | 5.9 | 12.0 | 10.9 | 6.2 | -8.8 | -28.5 | -49.9 | -87.0 |
| | | -4% | 4% | 9% | 13% | 10% | 5% | -6% | -16% | -25% | -40% |
| | | 0.708 | 0.520 | 0.798 | 0.881 | 0.657 | 0.312 | 0.316 | 0.559 | 0.581 | 0.593 |
| | B | 409 | 382 | 329 | 282 | 239 | 202 | 169 | 140 | 114 | 89 |
| | | -1.4 | -2.8 | -6.0 | -16.2 | -30.2 | -49.0 | -72.6 | -91.1 | -113.2 | -158.9 |
| | | -5% | -7% | -11% | -24% | -40% | -58% | -76% | -81% | -86% | -109% |
| | | 0.927 | 0.745 | 0.840 | 0.975 | 0.999 | 0.999 | 0.999 | 0.999 | 0.984 | 0.955 |
| | C | 419 | 388 | 337 | 289 | 244 | 205 | 169 | 137 | 110 | 86 |
| | | 2.3 | 6.2 | 9.6 | 11 | 11.2 | 6.5 | 1.9 | 3.4 | -3 | -24.8 |
| | | 8% | 12% | 14% | 12% | 10% | 5% | 1% | 2% | -1% | -9% |
| | | 0.946 | 0.99 | 0.99 | 0.933 | 0.769 | 0.377 | 0.086 | 0.106 | 0.069 | 0.394 |
| | D | 419 | 390 | 338 | 289 | 246 | 209 | 175 | 141 | 114 | 87 |
| | | 2.3 | 4.8 | 6.9 | 7.9 | 7.7 | 3.9 | 0.2 | -4.5 | -9.9 | -26.7 |
| | | 7% | 10% | 10% | 9% | 7% | 3% | 0% | -2% | -4% | -10% |
| | | 0.999 | 0.998 | 0.993 | 0.942 | 0.836 | 0.421 | 0.018 | 0.4 | 0.415 | 0.625 |

# METViewer CAM Scorecard

for GFDLFV3 and HRRR

2018-04-30 00:00:00 - 2018-06-01 00:00:00

| | | | Daily Domain | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | 12 hr | 14 hr | 16 hr | 18 hr | 20 hr | 22 hr | 24 hr | 26 hr | 28 hr | 30 hr | 32 hr | 34 hr | 36 hr |
| Fraction Skill Score | Composite Reflectivity | >=25.0 | ▼ | ▼ | | | ▾ | | ▾ | ▼ | ▼ | ▾ | | | |
| | | >=30.0 | ▼ | ▼ | ▾ | | ▼ | | ▼ | ▼ | ▼ | ▼ | | | |
| | | >=35.0 | ▼ | ▼ | ▼ | | | | ▼ | ▼ | ▼ | ▾ | | | |
| | | >=40.0 | ▼ | ▾ | ▼ | ▾ | | | ▼ | ▾ | ▾ | | | | |
| | | >=45.0 | ▼ | | ▾ | | | | ▼ | ▼ | ▾ | | | | |
| | | >=50.0 | ▾ | ▾ | | | ▾ | | ▼ | ▼ | | | | | |
| CSI | Composite Reflectivity | >=25.0 | ▾ | ▾ | | | | | | | | | | | |
| | | >=30.0 | ▼ | ▼ | | | | | | | | | | | |
| | | >=35.0 | ▾ | ▾ | | | | | | | | | | | |
| | | >=40.0 | ▼ | | | | | | | | | | | | |
| | | >=45.0 | | | | | | | | | | | | | |
| | | >=50.0 | | | ▲ | | | | | | | | | | |

| | |
|---|---|
| ▲ | GFDLFV3 is better than HRRR at the 99.9% significance level |
| ▴ | GFDLFV3 is better than HRRR at the 99% significance level |
| | GFDLFV3 is better than HRRR at the 95% significance level |
| | No statistically significant difference between GFDLFV3 and HRRR |
| | GFDLFV3 is worse than HRRR at the 95% significance level |
| ▾ | GFDLFV3 is worse than HRRR at the 99% significance level |
| ▼ | GFDLFV3 is worse than HRRR at the 99.9% significance level |
| | Not statistically relevant |

From T. Jensen, 2018

# Other factors to remember…

- Stratification
  - "Difficulty" of forecast (e.g., days with highly forced convection vs. other days)
  - Location, Topography, Season, etc. (the standard stuff…)
- Observation uncertainty
  - A big issue for weather…
  - An even bigger issue for ATM impacts
- Reporting/presentation of results
  - Must be meaningful to users of the results - will be different for different users (e.g., Air Traffic Managers vs. administrators)

# Some comments from Herbert

- It's most important to address verification of *high-impact events*
- Verification information may need to be simplified to be readily understood by operators
- Complex scores (e.g., Brier, ROC) should only be used by service provider/developers – generally not practical for operators
  - This may be addressed via specialized displays or summaries (e.g., via score cards or other displays)
  - These scores also may be translated to be useful for operators – e.g., red/amber/green
- It is valuable and important to find clear connections between weather verification and operational impacts
- To be useful, ensembles must be calibrated
- It's important to understand and communicate predictability

# Summary

- Choice of verification method is key to obtaining useful information about quality of
  - Weather forecasts
  - Weather forecasts translated to ATM impacts
- Verification methods should be selected to match the *type of forecast/observation* as well as the *questions* that are relevant for users of the forecasts
- *Intuitive displays and summaries* can make all the difference in usefulness of verification information!

# Where to go from here…

- Identify some specific weather and ATM events to evaluate based on data collection so far or in the future
  - Convective weather impacts on terminal aircraft acceptance rates
  - Convective impacts on traffic flow/capacity
- Identify questions to be answered and test
- Consider further how to translate product evaluations into user-relevant terms
  - Cost savings? Delay reductions? Etc.

# Resources

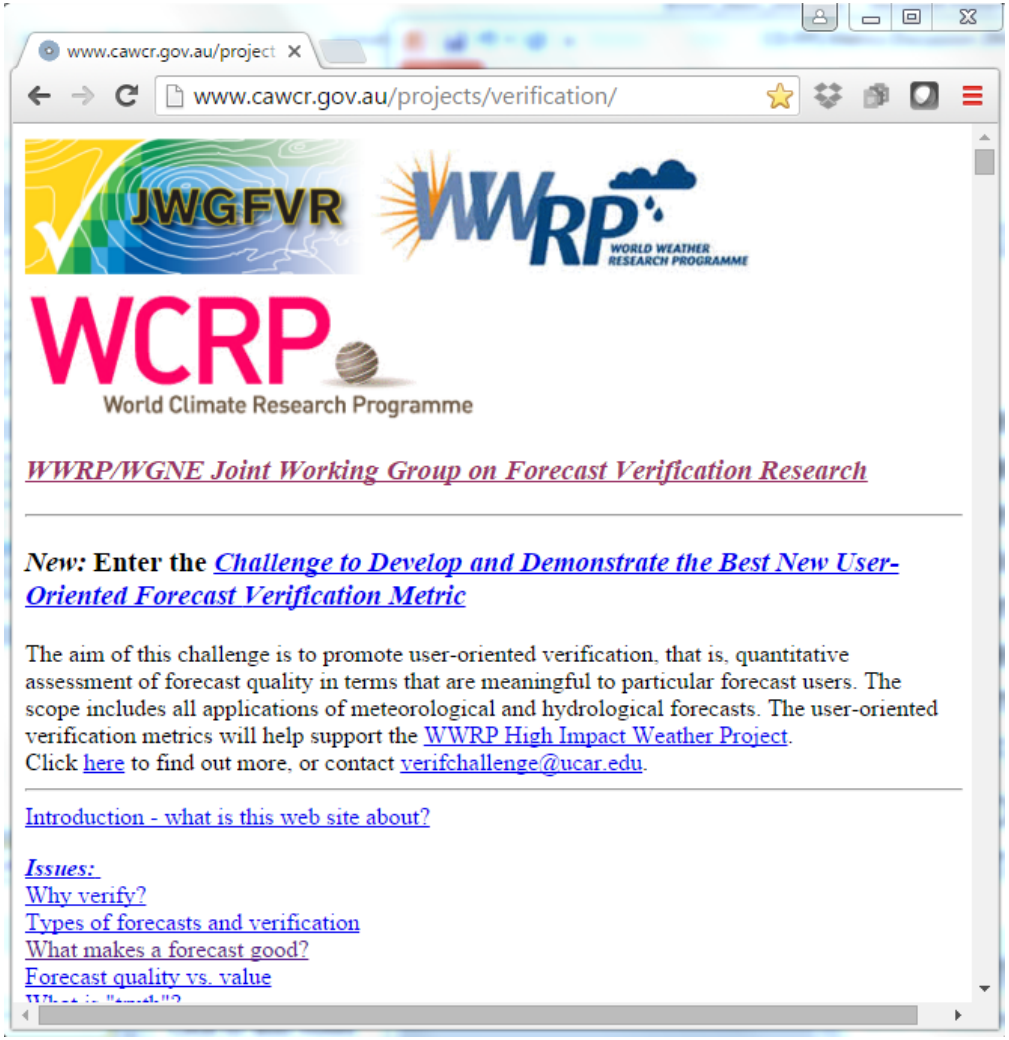
# Joint Working Group on Forecast Verification Research

- 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

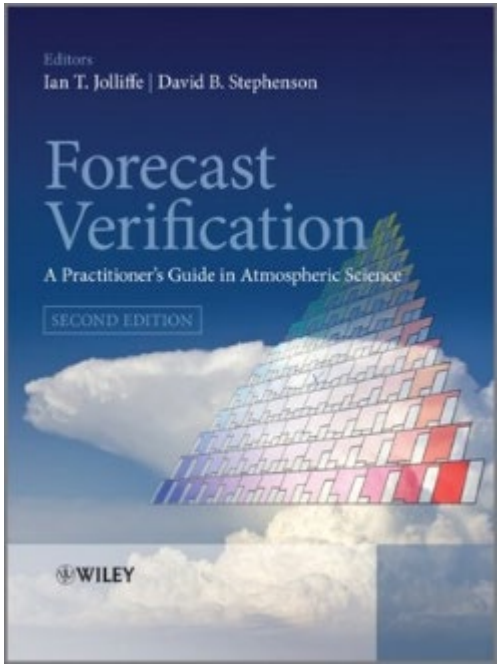
- FAQs
- Definitions
- Tools

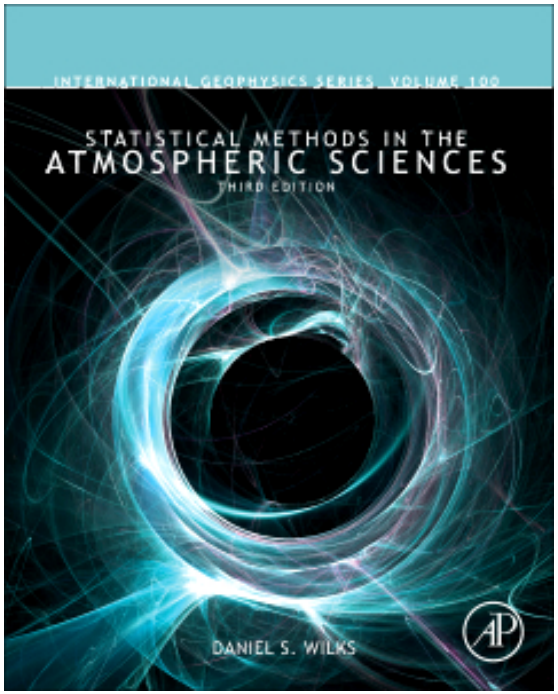



http://www.cawcr.gov.au/projects/verification/

# 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*)



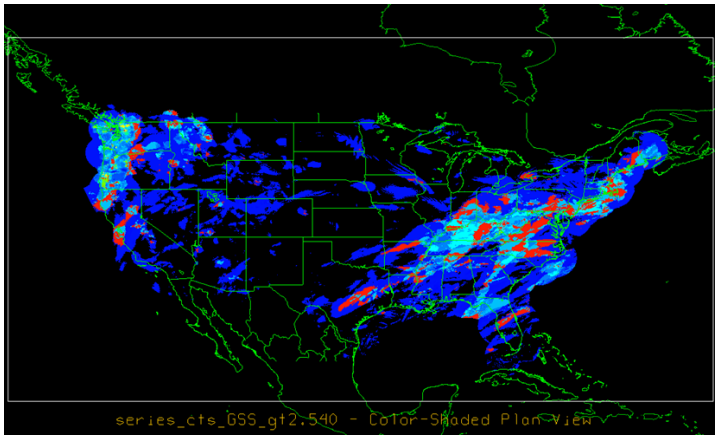



# Resources

- 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)




Spatial distribution of Gilbert Skill Score

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

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