An overview of the current status and recent advances in forecast evaluation methods

Barbara Brown

NCAR, Boulder, Colorado, USA

bgb@ucar.edu

Seminar 1

Central Weather Bureau; Taipei, Taiwan

13 March 2018

NCAR

NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Goals

To understand where we are and where we are going, it's helpful to understand where we have been and what we have learned…

- Evolution of verification of forecasts
  - ❖Including some ideas for S2S and Climate
- Challenges
  - ❖Observations and Uncertainty
  - ❖User-relevant approaches
  - ❖Methods for S2S and climate forecasts

Early verification

- Finley period… 1880's (Murphy paper: "*The Finley Affair*"; *WAF*, **11**, 1996)
- Focused on contingency table (categorical) statistics

		Observed	
		Yes	No
Forecast	Yes	*Hits*	*false alarms*
	No	*Misses*	*correct negatives*

Computing categorical verification measures

Yes/No contingency table

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

Use ***contingency table counts*** to compute a variety of measures

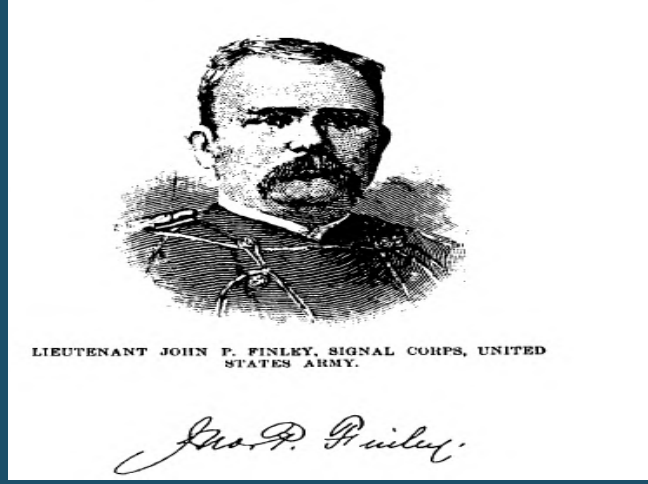
- POD, FAR, Freq. Bias, CSI, Gilbert Skill Score (= ETS), etc.

Important issues:

Choice of scores is critical

The traditional measures are not independent of each other

Finley Tornado Data (1884)



Forecast focused on the question:

Will there be a tornado?

YES

NO

Observation answered the question:

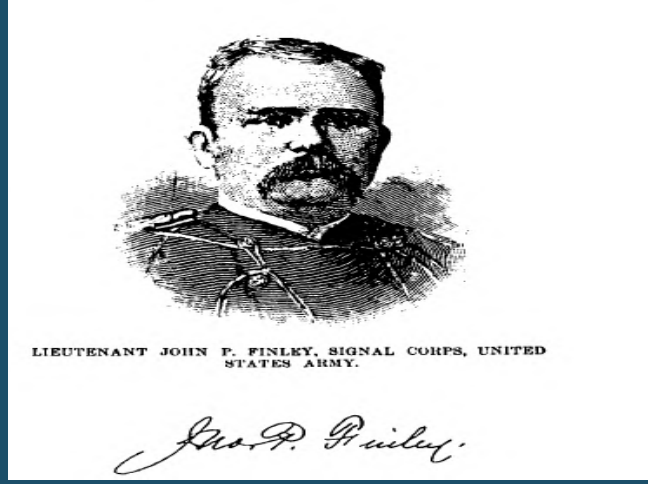
Did a tornado occur?

YES

NO

Answers fall into 1 of 2 categories ⁂ Forecasts and Obs are Binary

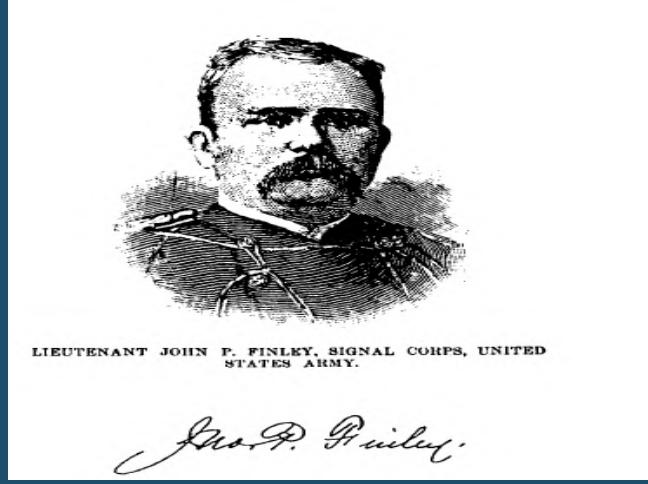
A Success?



Forecast	Observed		
	Yes	No	Total
Yes	28	72	100
No	23	2680	2703
Total	51	2752	2803

Percent Correct = (28+2680)/2803 = 96.6% !

What if forecaster never forecasted a tornado?



LIEUTENANT JOHN P. FINLEY, SIGNAL CORPS, UNITED STATES ARMY.

Forecast	Observed		
	Yes	No	Total
Yes	0	0	0
No	52	2752	2803
Total	51	2752	2803

Percent Correct = (0+2752)/2803 = 98.2%

See Murphy 1996 (*Weather and Forecasting*)

Lessons from Finley

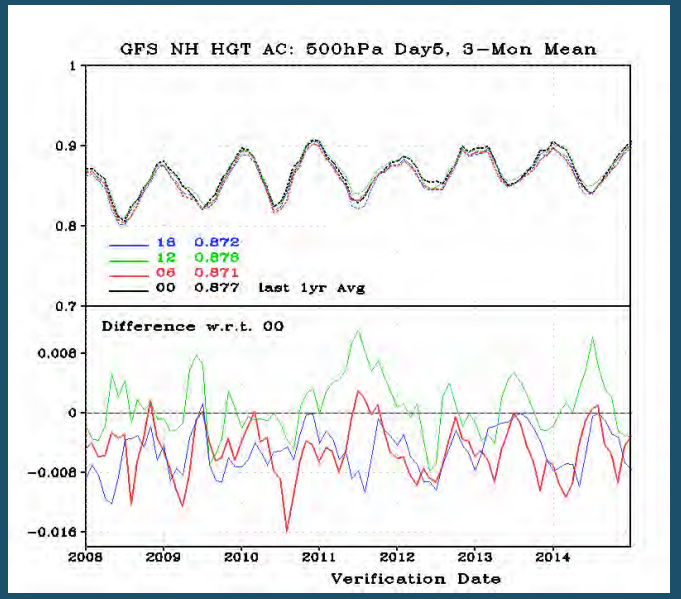
- Not all verification measures are (always) meaningful!
- Different measures needed for different purposes
- Many new measures developed (Gilbert, Peirce, Heidke, etc.)
- Discussion of "goodness" and "value" of forecasts

		Observed: Yes	Observed: No
Forecast	Yes	*Hits*	*false alarms*
Forecast	No	*Misses*	*correct negatives*

- Contingency table ("*Categorical*") methods are still the backbone of many verification efforts (e.g., warnings, seasonal)
- **Important note:** *Scores are not independent!*

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!



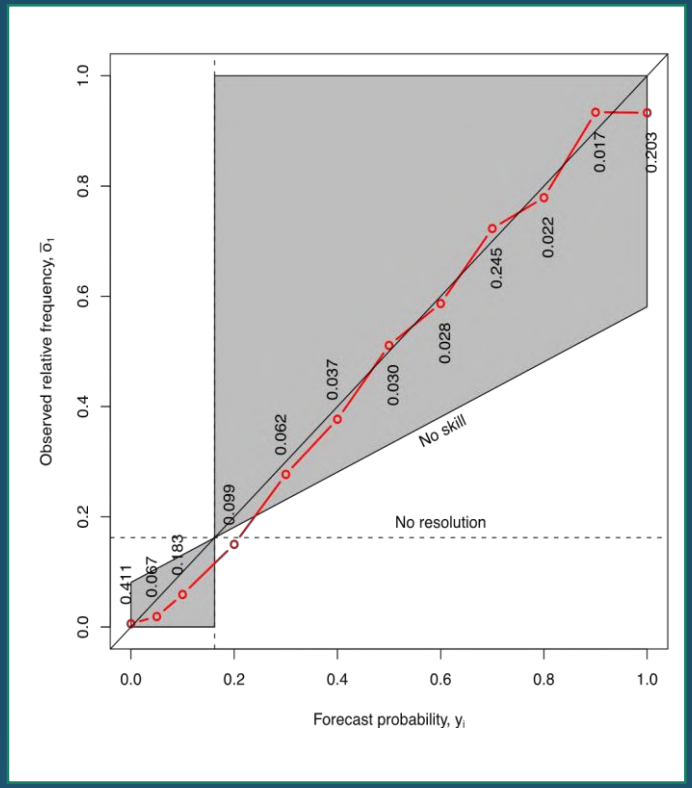
Development of "NWP" measures

- S1 score
- Anomaly correlation
- Still relied on for monitoring and comparing performance of NWP systems

Note: *Reliance on squared error statistics means we are optimizing toward the average – not toward extremes!*

The "Renaissance": The Allan Murphy era

- Expanded methods for probabilistic forecasts
  - Decompositions of scores => *meaningful interpretations of verification results*
  - Brier Score = Reliability – Resolution + Uncertainty
  - Attribute diagram
- Statistical framework for forecast verification
  - Joint distribution of forecasts and observations and their factorizations
  - Placed verification in a statistical context



Dimensionality of the forecast problem:

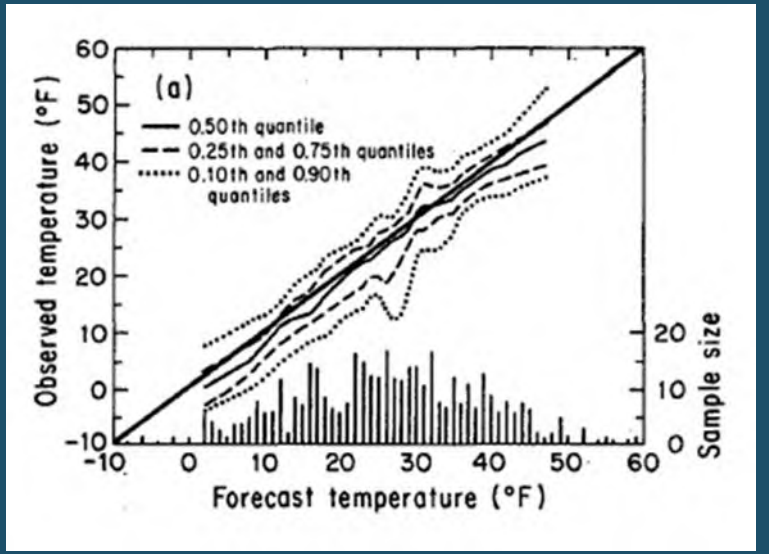
d= nf*nx - 1

For 2x2 contingency table, d=3!

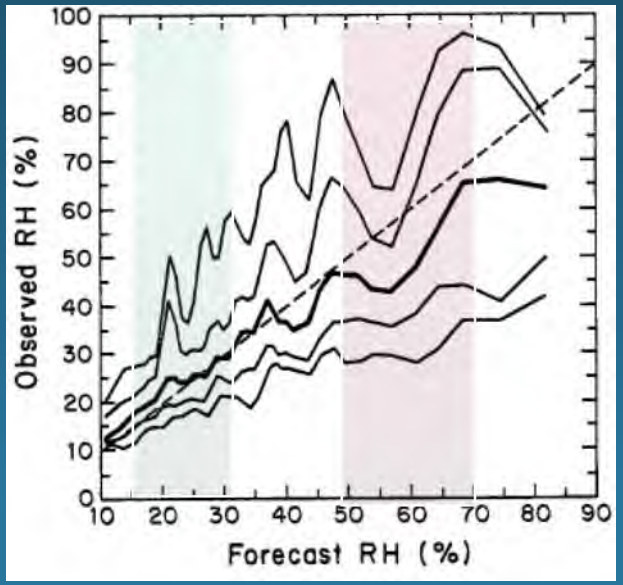
Murphy era cont.

"Diagnostic" verification

- Focus on measuring *attributes* of performance rather than *summary measures*
- *A revolutionary idea: Instead of relying on a single measure of "overall" performance*
  - *ask questions about performance*
  - *measure attributes that can answer those questions*



Example: Use of conditional quantile plots to examine conditional biases in forecacsts



Murphy: Forecast quality depends on a wide variety of forecast attributes

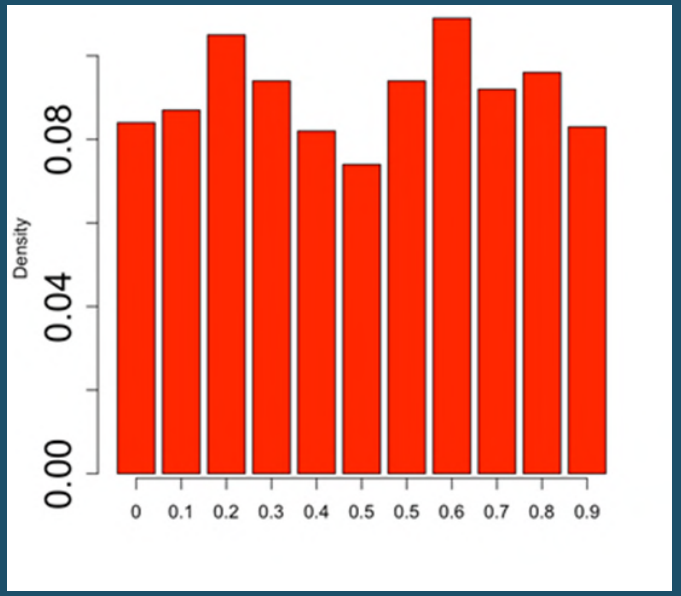
- Defined using the joint distribution of forecasts and observations
- Different combinations of attributes represent different characteristics relevant for different users

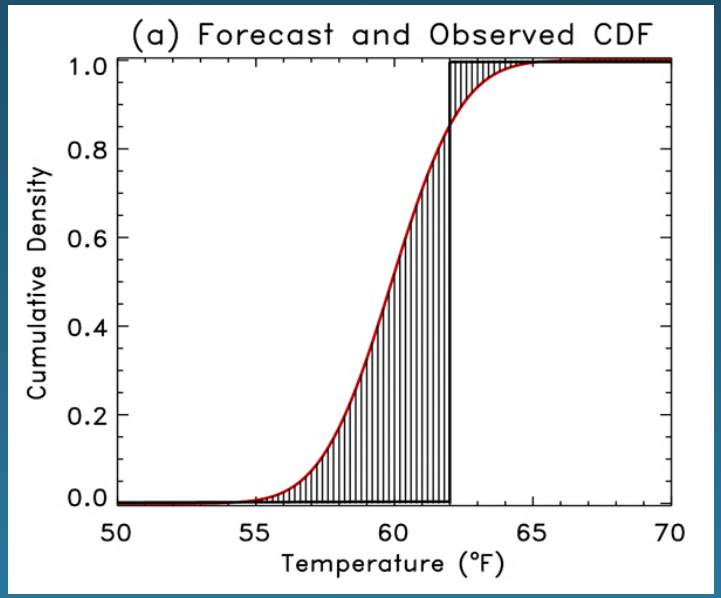
Aspect	Definition	Relevant distribution(s)
Bias	Correspondence between mean forecast and mean observation	$p(f)$ and $p(x)$
Association	Overall strength of linear relationship between individual pairs of forecasts and observations	$p(f, x)$
Accuracy	Average correspondence between individual pairs of forecasts and observations	$p(f, x)$
Skill	Accuracy of forecasts of interest relative to accuracy of forecasts produced by standard of reference	$p(f, x)$
Reliability	Correspondence between conditional mean observation and conditioning forecast, averaged over all forecasts	$p(x\|f)$ and $p(f)$
Resolution	Difference between conditional mean observation and unconditional mean observation, averaged over all forecasts	$p(x\|f)$ and $p(f)$
Sharpness	Variability of forecasts as described by distribution of forecasts	$p(f)$
Discrimination 1	Correspondence between conditional mean forecast and conditioning observation, averaged over all observations	$p(f\|x)$ and $p(x)$
Discrimination 2	Difference between conditional mean forecast and unconditional mean forecast, averaged over all observations	$p(f\|x)$ and $p(x)$
Uncertainty	Variability of observations as described by distribution of observations	$p(x)$

Murphy, 1993 (WAF)

The "Modern" era

- New focus on evaluation of probability and ensemble forecasts
  - Development of new methods specific to ensembles (rank histogram, CRPS)
- Greater understanding of limitations of methods
  - "Meta" verification
  - Examples:
    - Propriety: Don't encourage hedging
    - Equitability: "Bad" forecasts are represented consistently

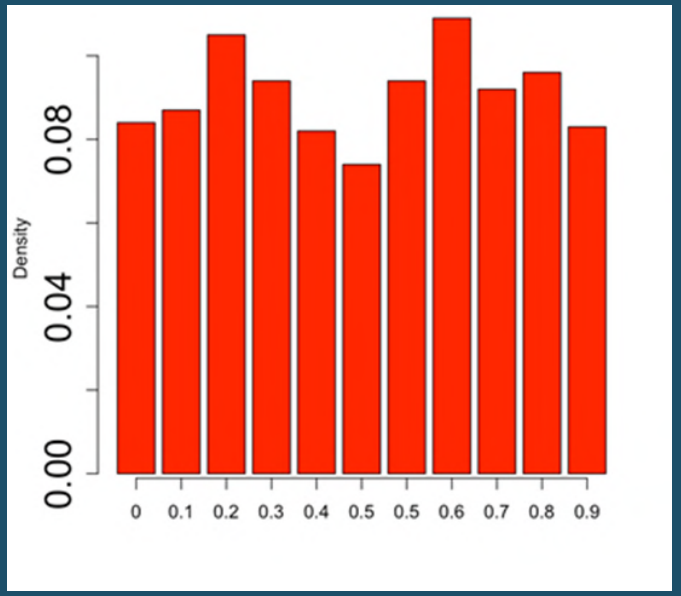


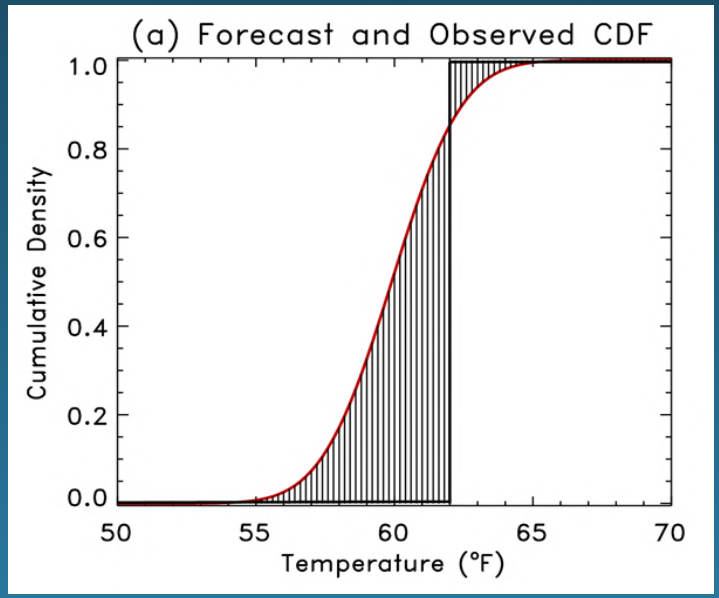


Measure	Attribute evaluated	Comments
Probability forecasts		
Brier score	Accuracy	Based on squared error
Resolution	Resolution (resolving different categories)	Compares forecast category climatologies to overall climatology
Reliability	Calibration	
Skill score	Skill	Skill involves *comparison* of forecasts
Sharpness measure	Sharpness	Only considers distribution of forecasts
ROC	Discrimination	Ignores calibration
C/L Value	Value	Ignores calibration
Ensemble distribution		
Rank histogram	Calibration	Can be misleading
Spread-skill	Calibration	Difficult to achieve
CRPS	Accuracy	Squared difference between forecast and observed distributions Analogous to MAE in limit
log p score	Accuracy (IGN = -log2 p_C)	Local score, rewards for correct category; infinite if observed category has 0 density

The "Modern" era

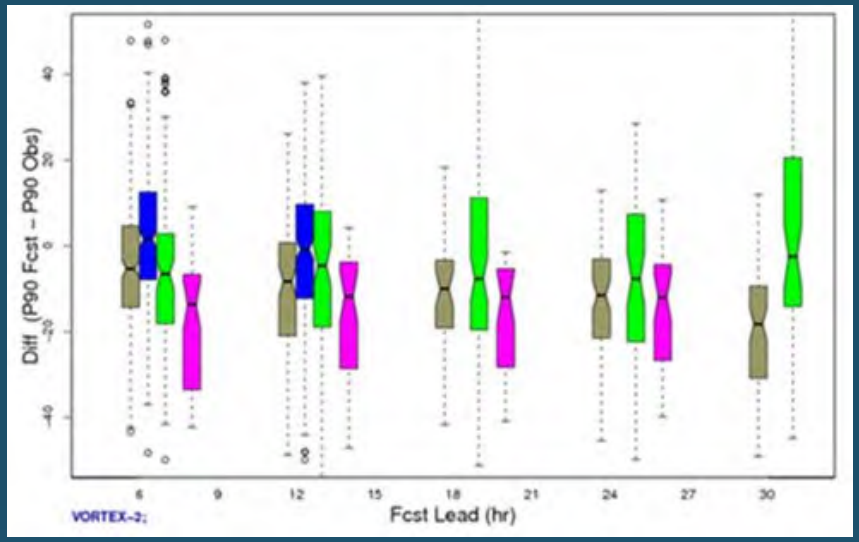
- New focus on evaluation of probability and ensemble forecasts
  - Development of new methods specific to ensembles (rank histogram, CRPS)
- Greater understanding of limitations of methods
  - "Meta" verification
  - Examples:
    - Propriety: Don't encourage hedging
    - Equitability: "Bad" forecasts are represented consistently

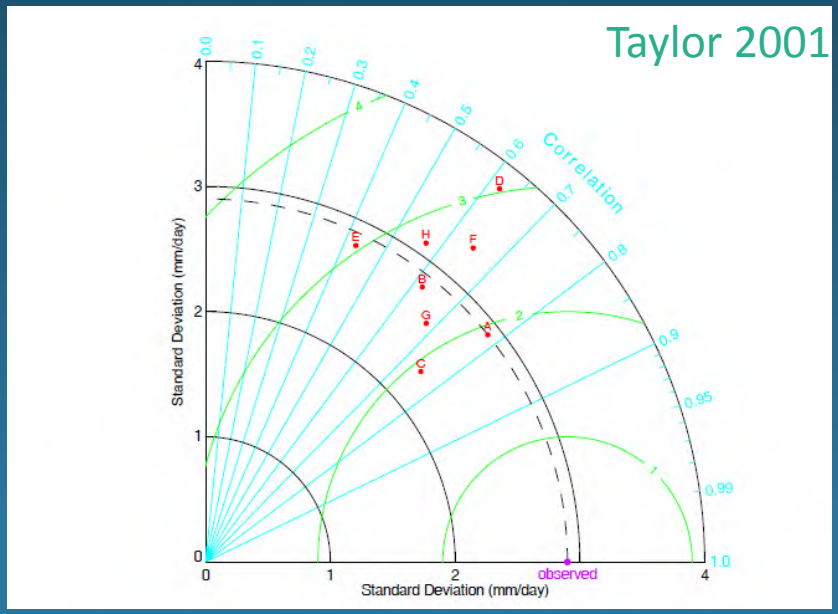


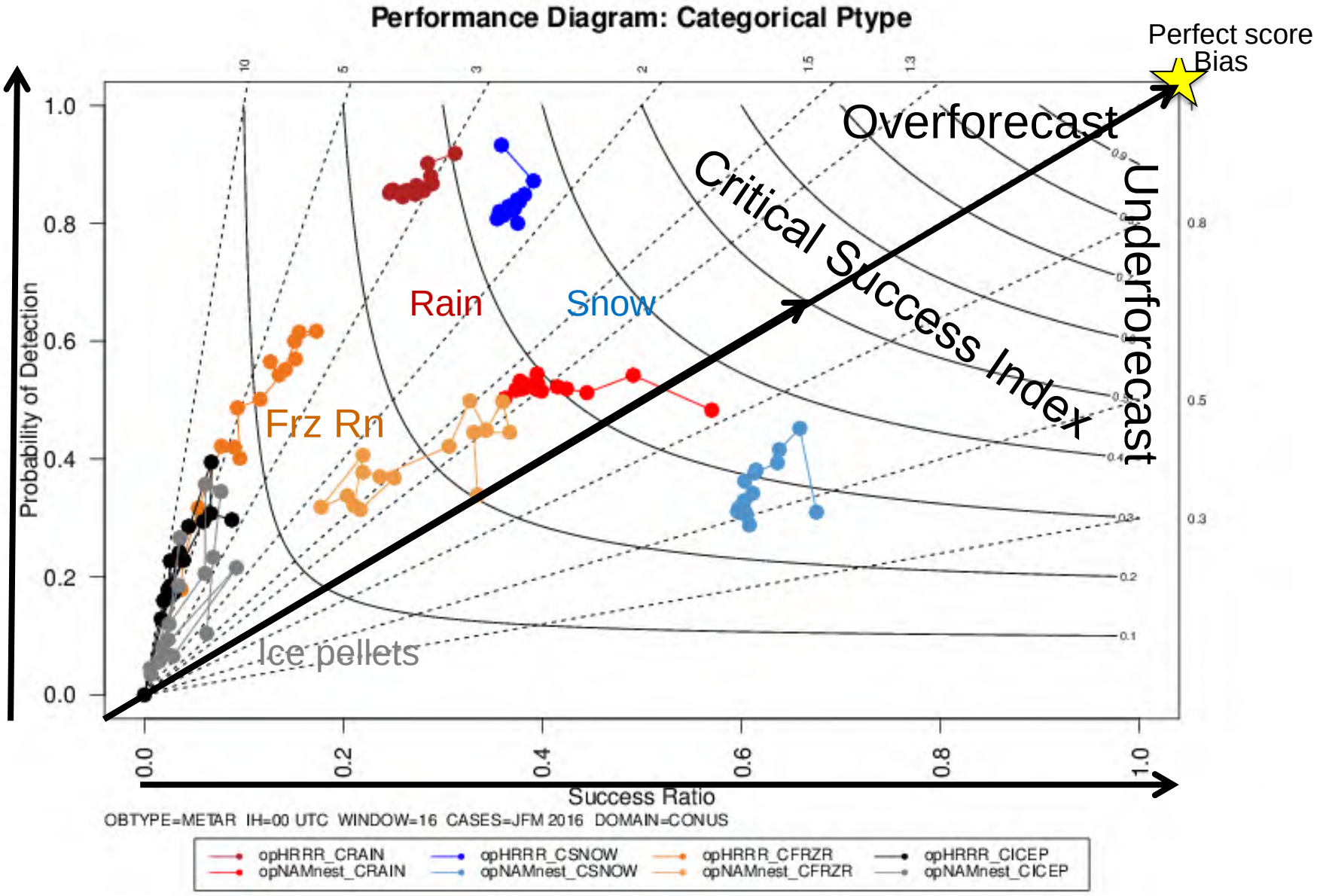


The Modern era cont'.

- Evaluation of sampling uncertainty in verification measures
  - Confidence intervals (parametric or bootstrapped)
  - Distributions of errors
- Approaches to evaluate multiple attributes simultaneously
  - Note: this idea is an extension of Murphy's attribute diagram idea to other types of measures
  - Ex: Performance diagrams, Taylor diagrams





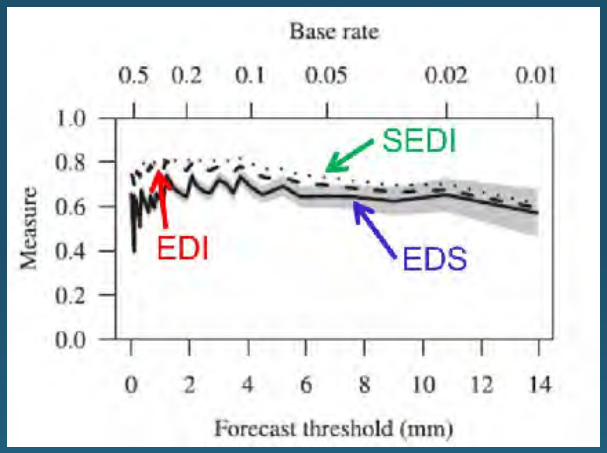


Credit: J. Wolff, NCAR

The "Modern" era cont.

- International Verification Community
  - Workshops, textbooks…
- 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

WMO Joint Working Group on Forecast Verification Research

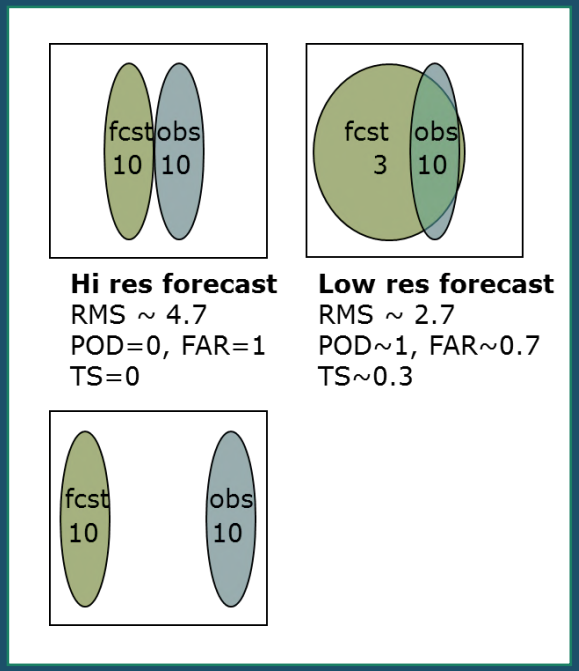


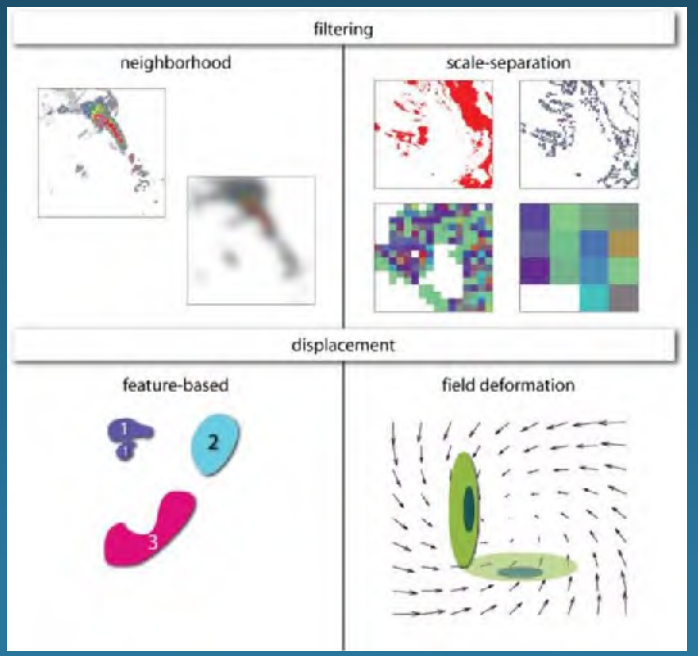
From Ferro and Stephenson 2011 (*Wx and Forecasting*)

Spatial verification methods

Inspired by lack of *diagnostic* information from traditional approaches

- Difficult to distinguish differences between forecasts
- Double penalty problem
  - Forecasts fail the "eye" test
  - Smoother forecasts often "win"
- Want score to say what went wrong or was good about a forecast





New Spatial Verification Approaches

Neighborhood

Successive smoothing of forecasts/obs

Gives credit to "close" forecasts

Scale separation

Measure scale-dependent error

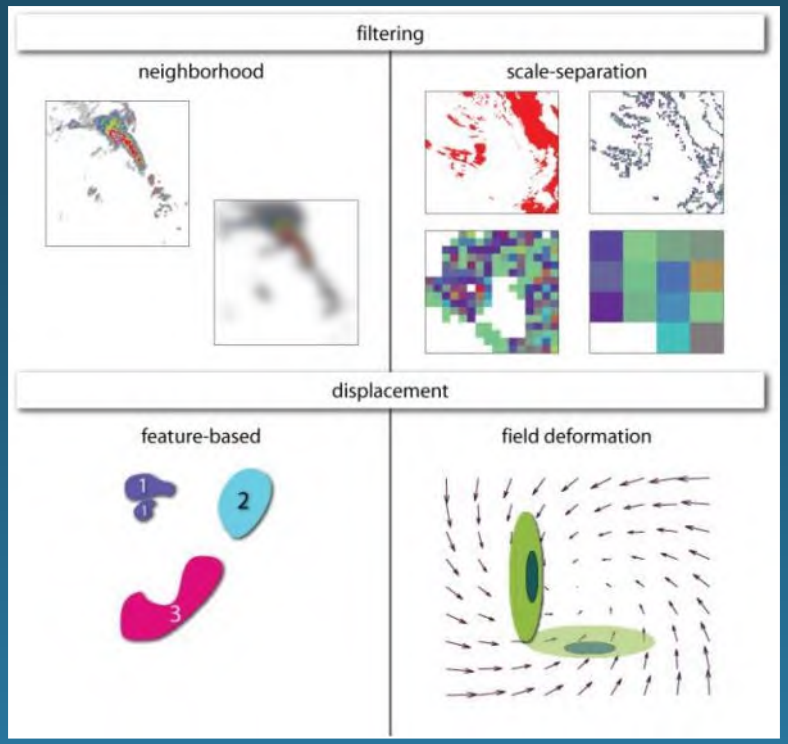
Field deformation

Measure distortion and displacement (phase error) for whole field

How should the forecast be adjusted to make the best match with the observed field?

Object- and feature-based

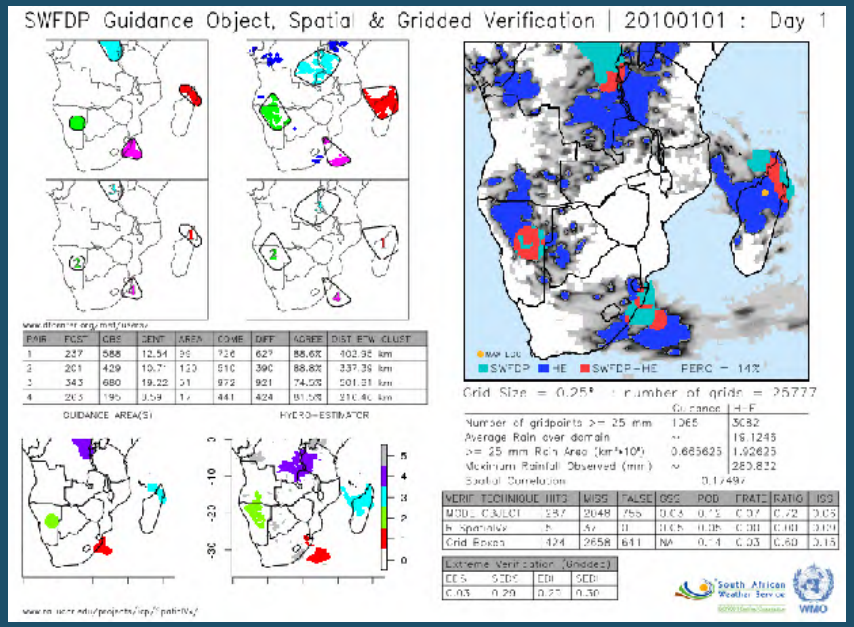
Evaluate attributes of identifiable features



http://www.ral.ucar.edu/projects/icp/

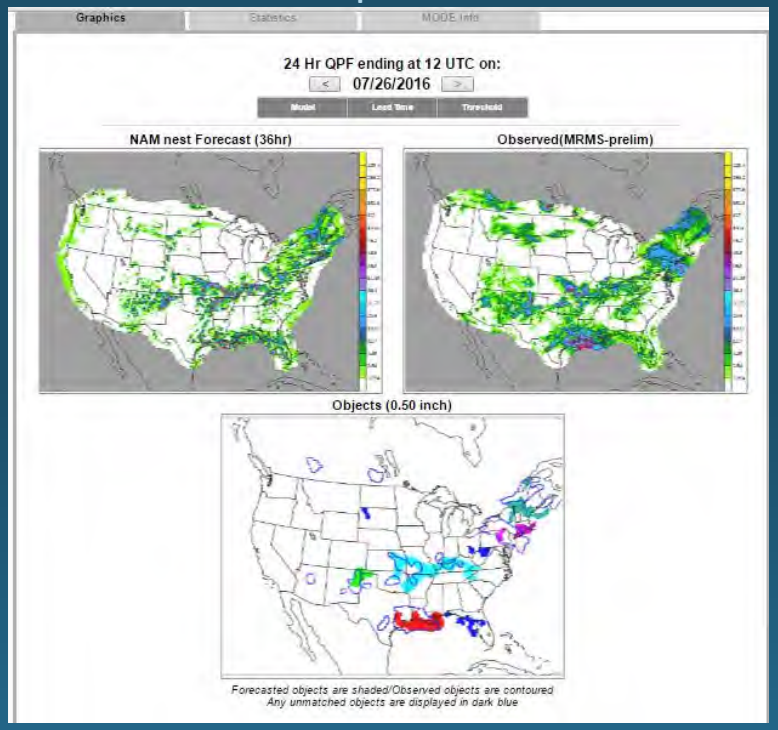
Example Applications

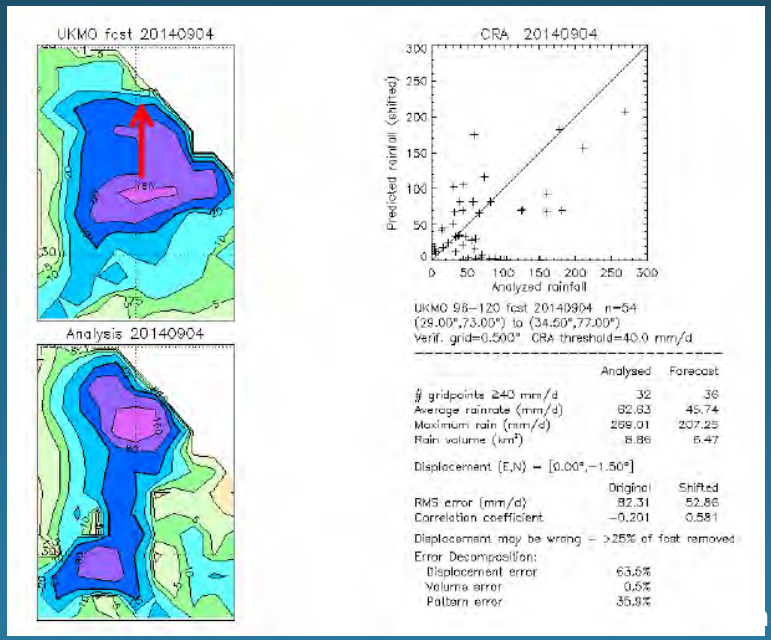
SWFDP, South Africa



From Landman and Marx 2015 presentation

US Weather prediction Center





Ebert and Ashrit (2015): CRA

Object-based extreme rainfall evaluation:

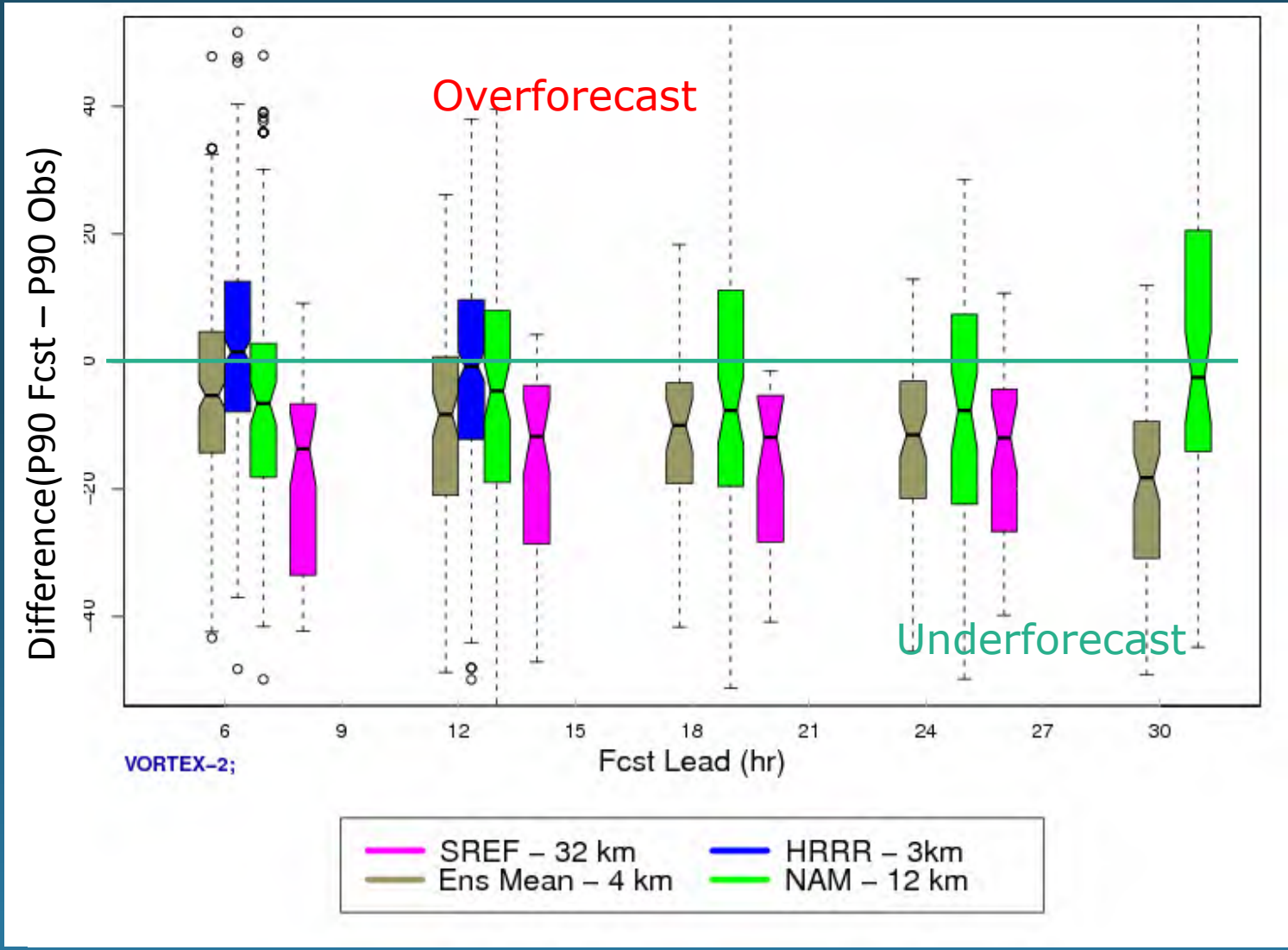
6hr Accumulated Precipitation Near Peak (90th%) Intensity Difference (Fcst – Obs)

High Resolution Deterministic Does Fairly Well

High Resolution Ensemble Mean Underpredicts

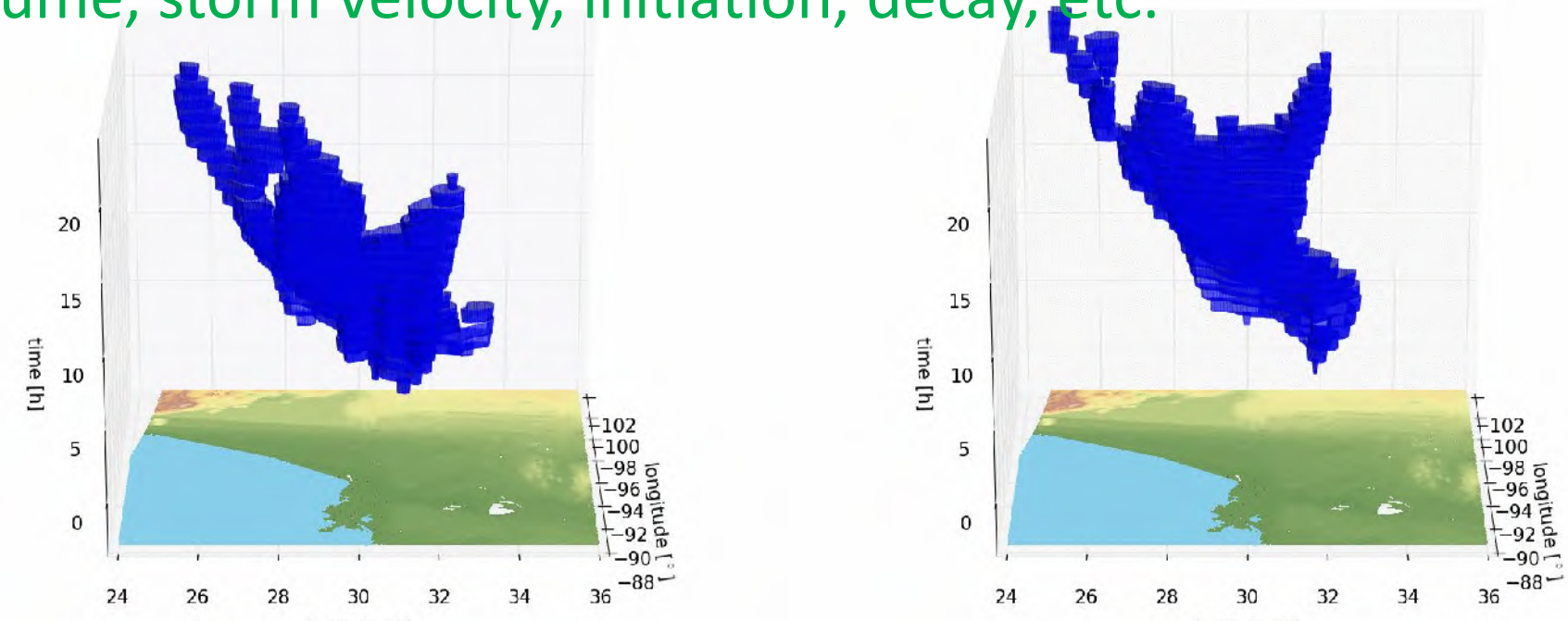
Mesoscale Deterministic Underpredicts

Mesoscale Ensemble Underpredicts the most



MODE Time Domain: Adding the time Dimension

MODE-TD allows evaluation of timing errors, storm volume, storm velocity, initiation, decay, etc.



Application of MODE-TD to WRF prediction of an MCS in 2007 (Credit: A. Prein, NCAR)

MODE and MODE-TD are available through the Model Evaluation Tools (http://www.dtcenter.org/met/users/)

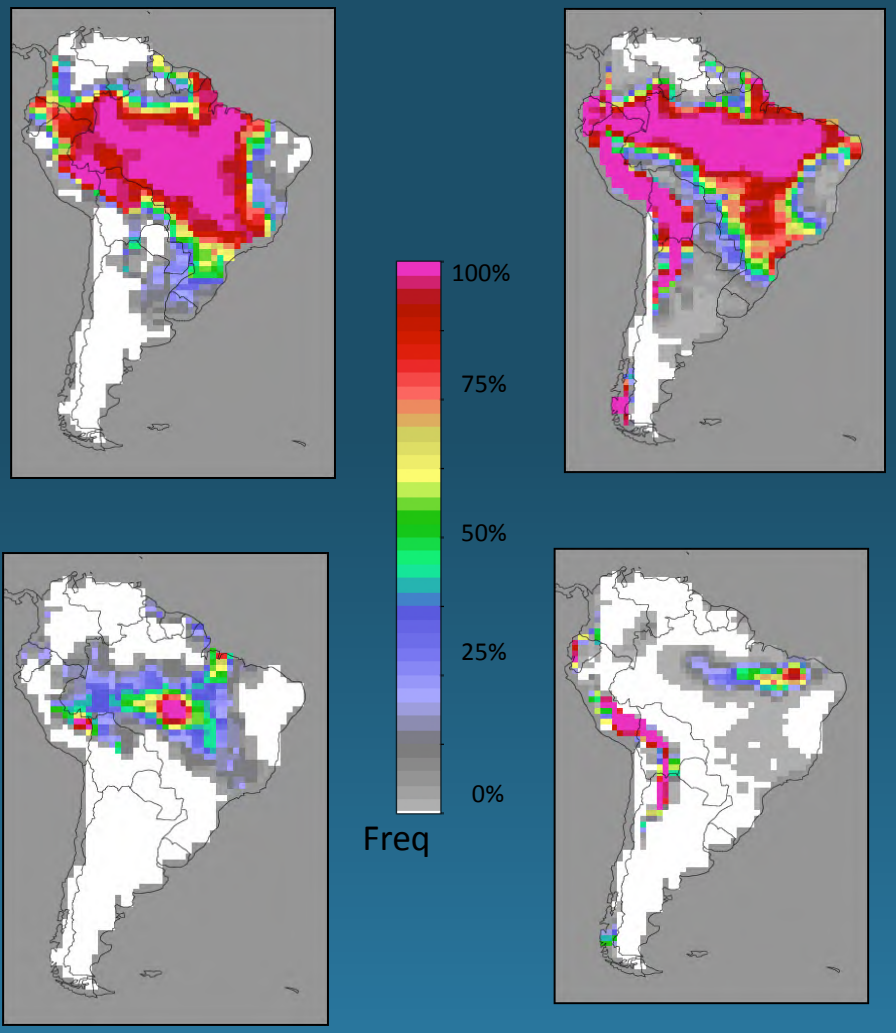
Climate application of MODE

CRU TS3.21

CESM-LE

Object frequency >500mm

Object frequency >900mm



WMO/WWRP/WCRP: S2S Verification recommendations

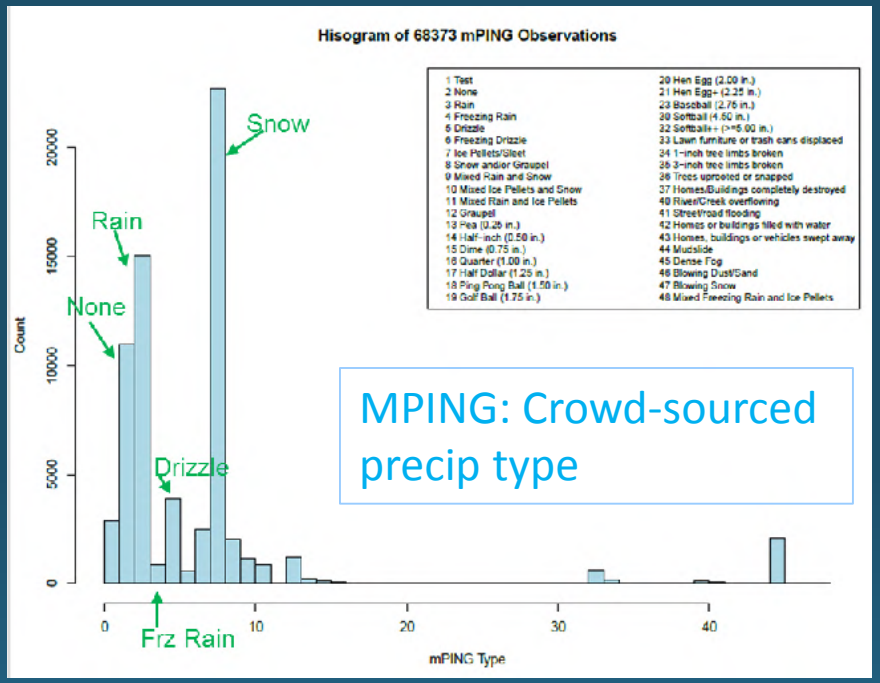
- Development of user-relevant metrics, thresholds, etc.
  - Identify relevant variables (e.g., rainfall phases) as well as procedures – beyond standard "average" events
  - Phase space methods (e.g., for MJO)
- Implement S2S framework for evaluating real-time and retrospective forecast skill
- Conditional verification (e.g., by ENSO, MJO)
- Appropriate measures for extremes and discrimination
- Spatial methods
- Account for sampling uncertainty

From book in preparation: *The Gap between Weather and Climate Forecasting: Sub-seasonal to Seasonal Prediction*; Chapter on "Forecast Verification for S2S Time Scales" (Coelho, Brown, Wilson, Mittermaier, and Casati)

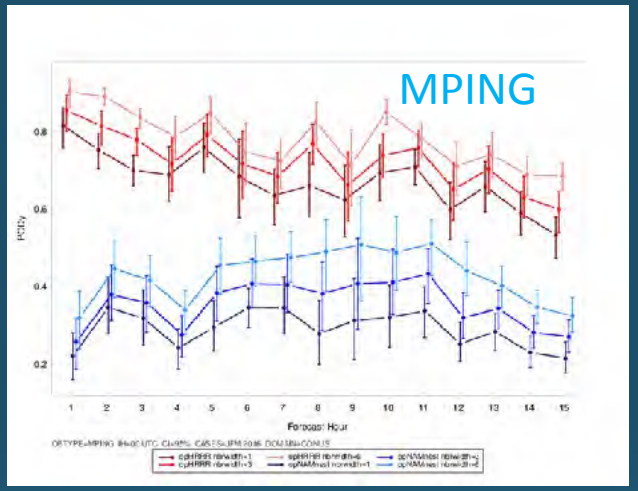
Challenges

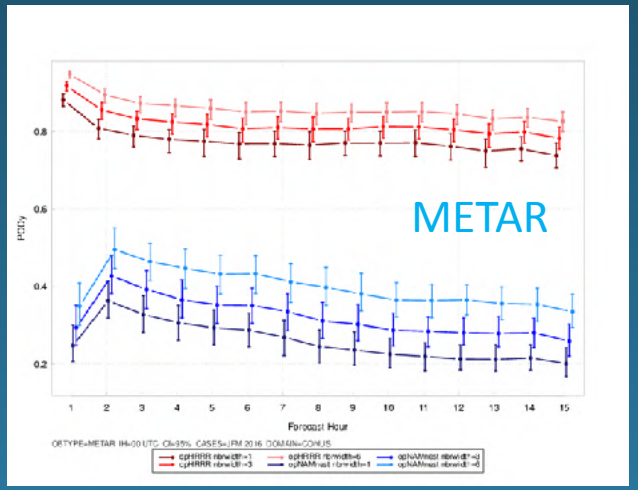
- Observation limitations
  - Representativeness
  - Biases
- Measuring and incorporating uncertainty information
  - **Sampling**: Methods are available but not typically applied
  - **Observation:** Few methods available; not clear how to do this in general
- User-relevant verification
  - Evaluating forecasts in the context of user applications and decision making

Example: Precipitation Type



Snow precip type (2 models): POD vs lead time





Human-generated observations have biases (e.g., in types observed)

Type of observation impacts the verification results

Credit: J. Wolff (NCAR)

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….

Summary

- Much progress has been made in the last few decades
  Advancing capabilities and 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
  Thinking beyond contingency tables
- Thoughtfulness in selecting and implementing verification approaches will pay off in more meaningful results
  Optimize forecasts for what we care about

But still more challenges ahead…

Xièxiè　谢谢

resources

WMO Working Group on Forecast Verification Research

- Working Group under the World Weather Research Program (WWRP) and Working Group on Numerical Experimentation (WGNE)
- International representation
- Activities:
  - Verification research
  - Training
  - Workshops
  - Publications on “best practices”: Precipitation, Clouds, Tropical Cyclones

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

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

Or email vx-discuss@mail.rap.ucar.edu

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

Spatial Method Intercomparison Project

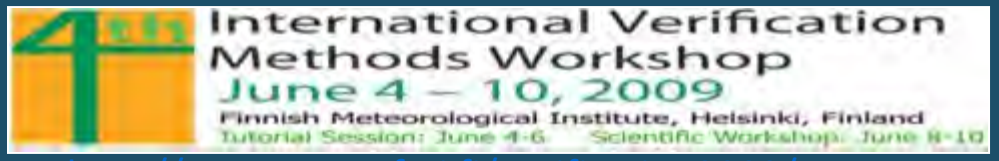
- International effort to evaluate and compare new verification methods
- Second intercomparison in progress
  - Focus on complex terrain, ensembles, precipitation and wind
- Many references on spatial methods

http://www.ral.ucar.edu/projects/icp/

Workshops and tutorials

- WMO Tutorials (3rd, 4th, 5th, 6th workshops)
  - Presentations available
- EUMETCAL tutorial
  - Hands-on tutorial

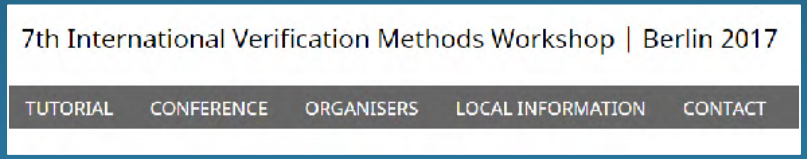
http://www.eumetcal.org/-Eumetcal-modules-



http://www.space.fmi.fi/Verification2009/

http://cawcr.gov.au/events/verif2011/

http://www.ncmrwf.gov.in/verif2014/Tutorial-Program.php



https://www.7thverificationworkshop.de/

Resources: Overview papers

- Casati et al. 2008: *Forecast verification: current status and future directions*.

  Meteorological Applications, **15**, 3-18.

- Ebert et al. 2013: *Progress and challenges in forecast verification*

  Meteorological Applications, **20**, 130-139.

Papers summarizing outcomes and discussions from 3rd and 5th International Workshop on Verification Methods

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*)