Verification of nowcasts and short-range forecasts, including aviation weather

Barbara Brown

NCAR, Boulder, Colorado, USA

WMO WWRP 4th International Symposium on Nowcasting and Very-short-range Forecast 2016 (WSN16)

Hong Kong; July 2016

NCAR
NATIONAL CENTER FOR ATMOSPHERIC RESEARCH

Goals

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

- Evolution of verification of short-range forecasts
- Challenges
  - Observations and Uncertainty
  - User-relevant approaches

Early verification

- Finley period… 1880's (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…

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

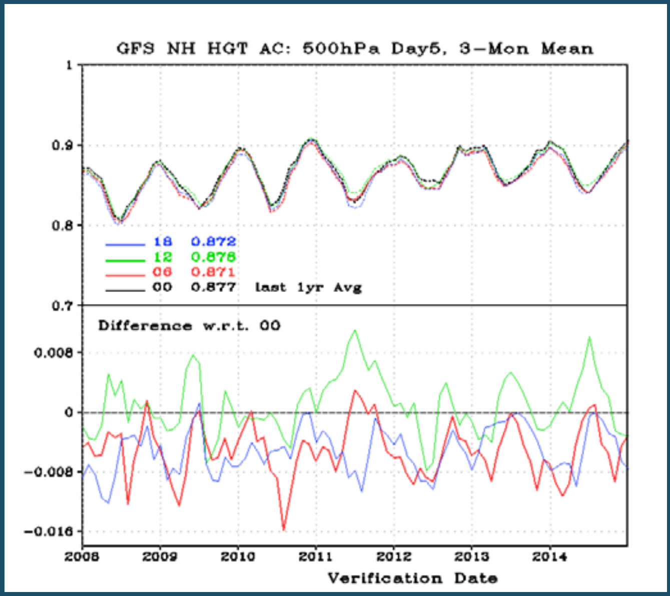
These methods are still the ***backbone*** of many verification efforts (e.g., warnings)

Important notes:

- Many categorical scores are not independent!
- At least 3 metrics are needed to fully characterize the bivariate distribution of forecasts and observations

Early years continued: Continuous measures

- Focus on squared error statistics
  - Mean-squared error
  - Correlation
  - Bias
  - <u>Note</u>: Little recognition before Murphy of the non-independence of these measures
- Extension to probabilistic forecasts
  - Brier Score (1950) – well before prevalence of probability forecasts!

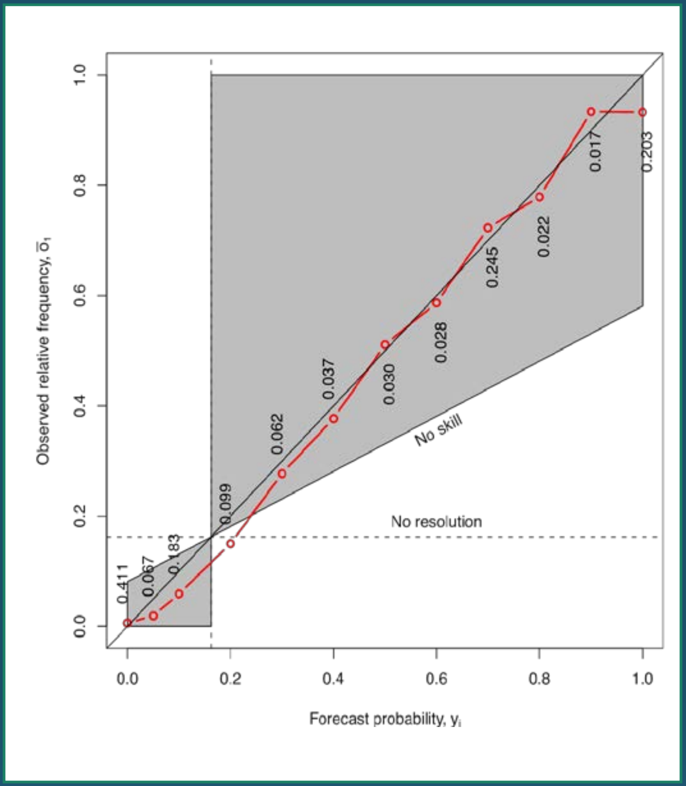


Development of "NWP" measures

- S1 score
- Anomaly correlation
- Still relied on for monitoring and comparing performance of NWP systems (Are these still the best measures for this purpose?)

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

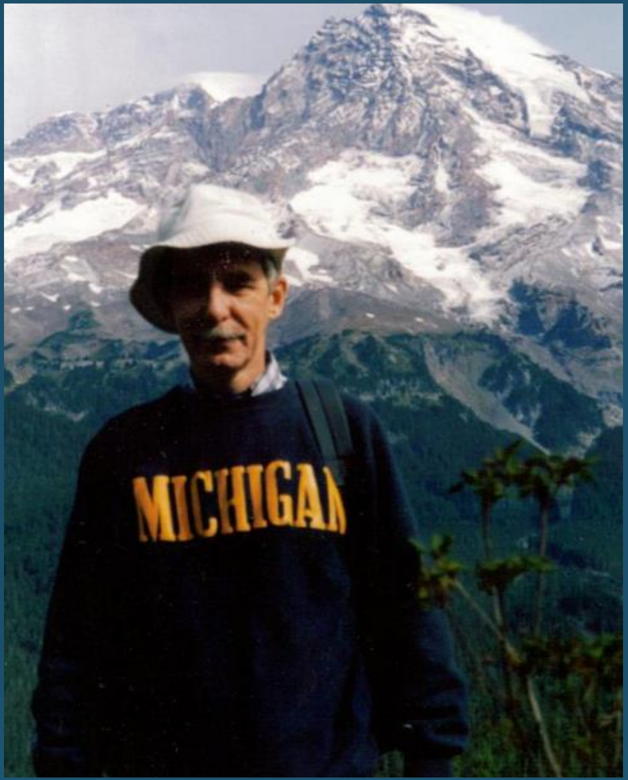
The “Renaissance”: The Allan Murphy era



- Expanded methods for probabilistic forecasts
  - Decompositions of scores led to more meaningful interpretations of verification results
  - Attribute diagram
- Initiation of ideas of meta verification: Equitabiltiy, Propriety
- 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:

$$d = n_f * n_x - 1$$

"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"

The Murphy era cont.

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

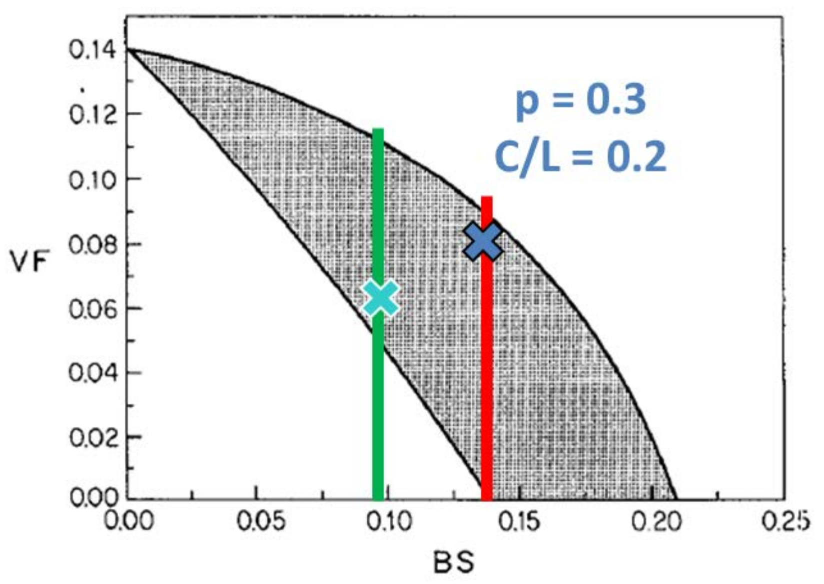


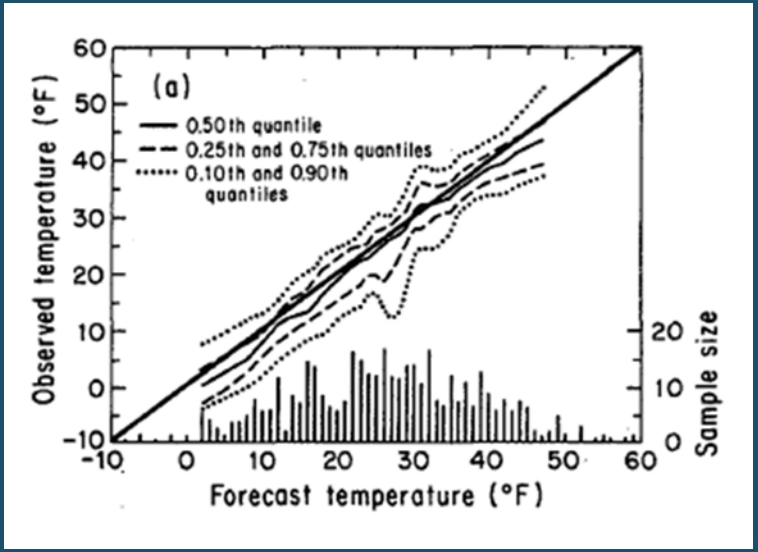
FIG. 4. Relationship between forecast accuracy and forecast value in the cost–loss ratio situation, with climatological probability $\pi = 0.3$ and cost–loss ratio $C/L = 0.2$ (taken from Murphy and Ehrendorfer 1987).

From Murphy, 1993 (*Weather and Forecasting*)

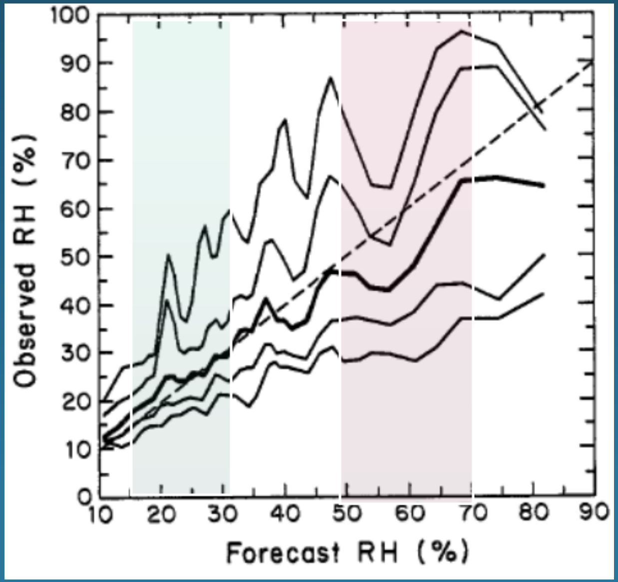
Murphy era cont.

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*

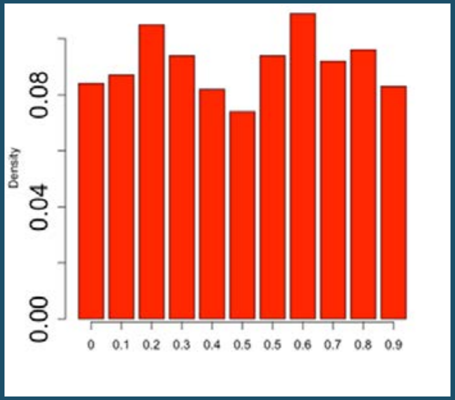


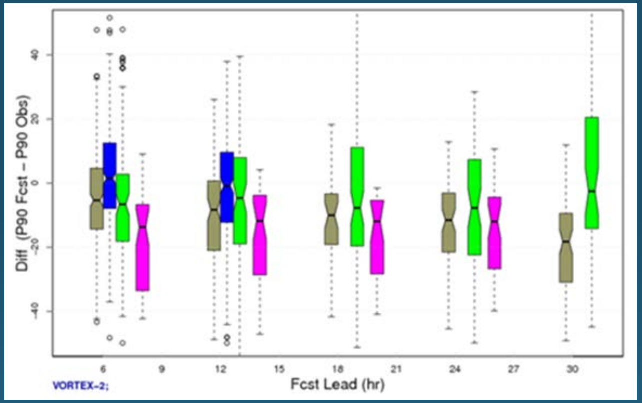
Example: Use of conditional quantile plots to examine conditional biases in forecacsts

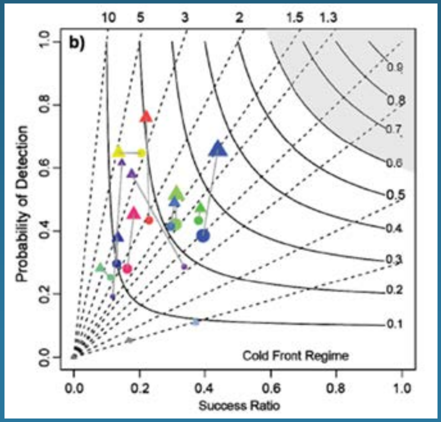


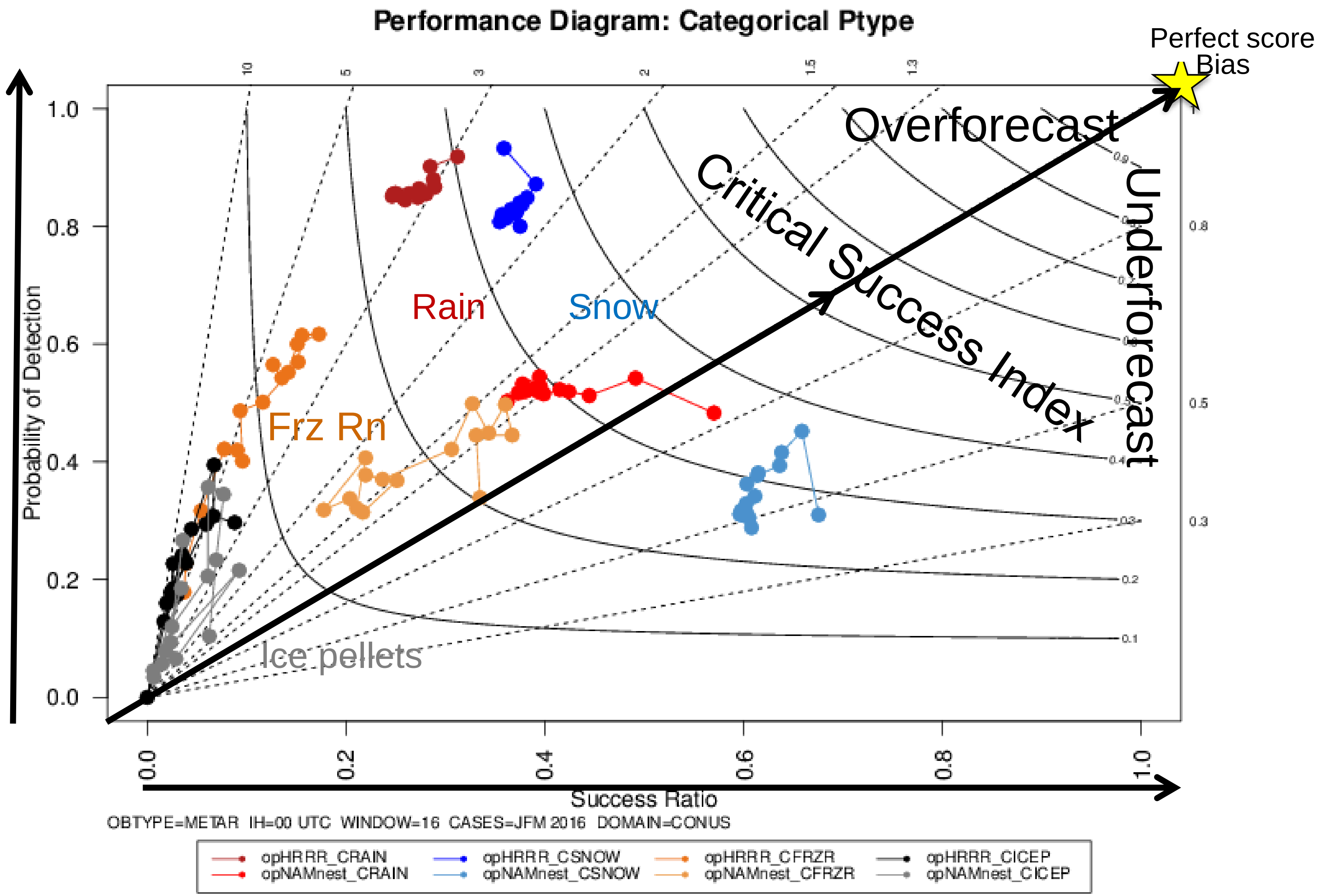
The "Modern" era

- 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







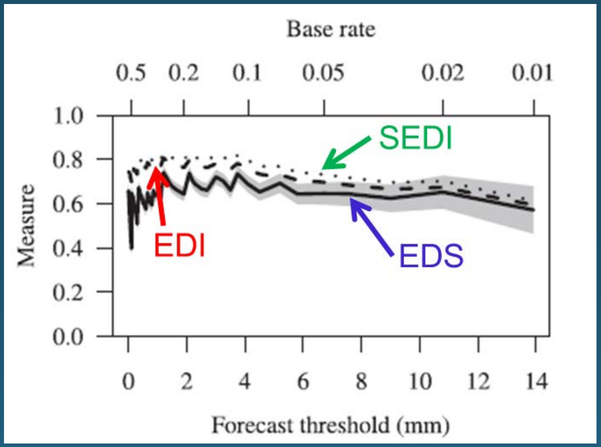


Credit: J. Wolff, NCAR

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

WMO Joint Working Group on Forecast Verification Research

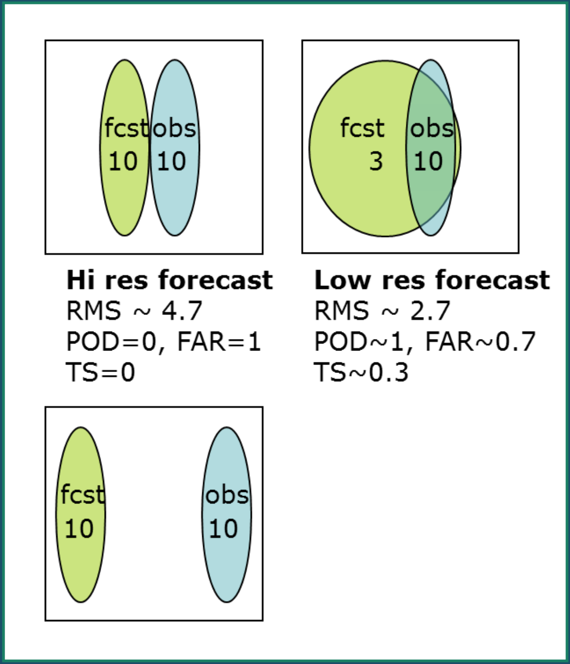


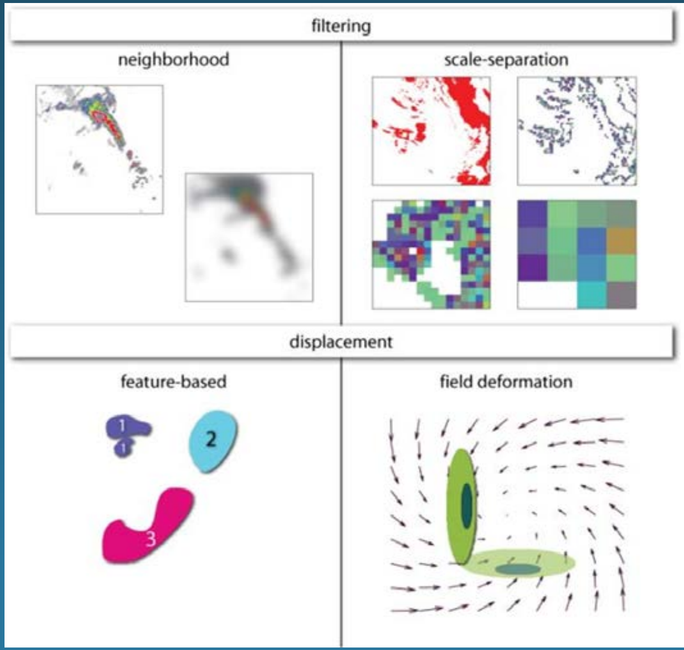
From Ferro and Stephenson 2011 (*Wx and Forecasting*)

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

Object- and feature-based

Evaluate attributes of identifiable features

Neighborhood

Successive smoothing of forecasts/obs

Gives credit to "close" forecasts

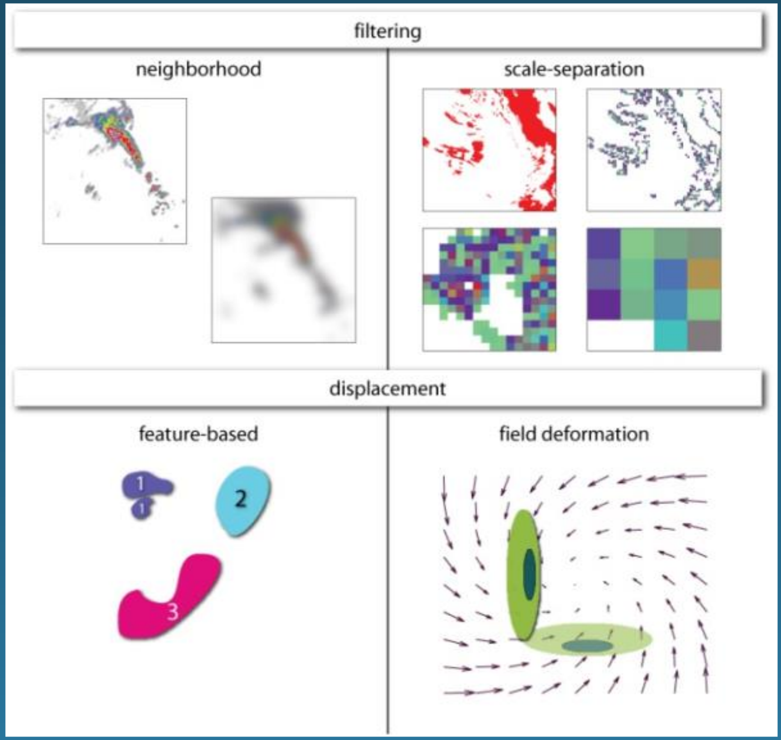
Scale separation

Measure scale-dependent error

Field deformation

Measure distortion and displacement (phase error) for whole field

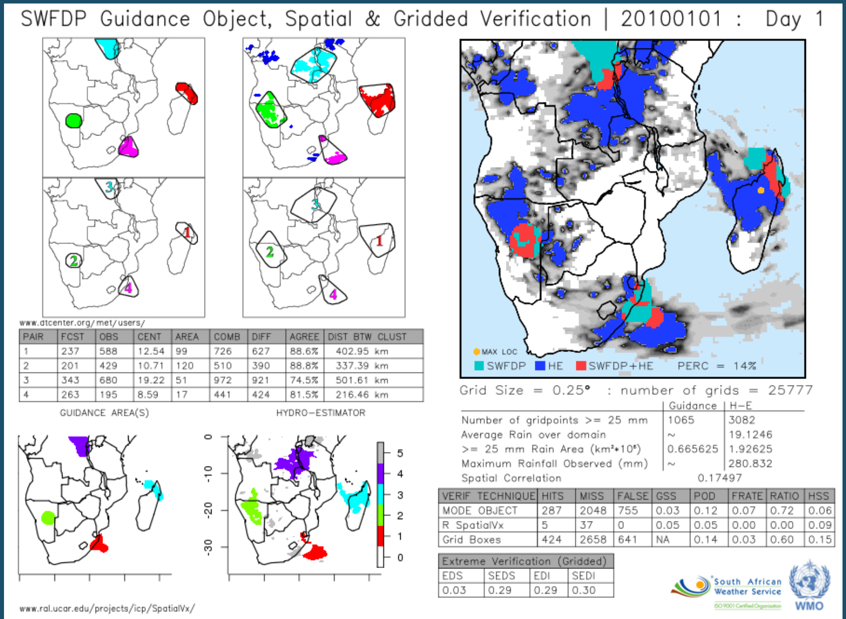
How should the forecast be adjusted to make the best match with the observed field?



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

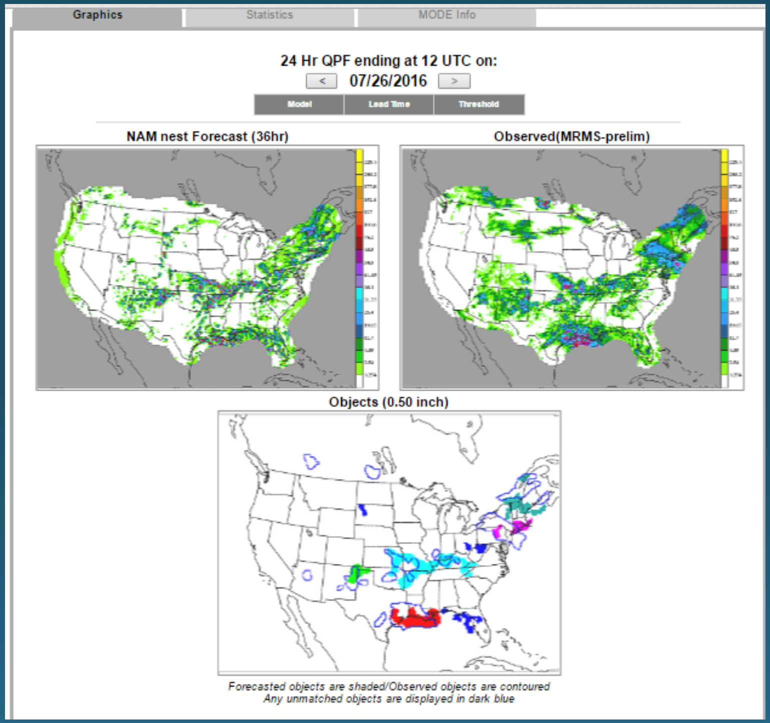
Example Applications

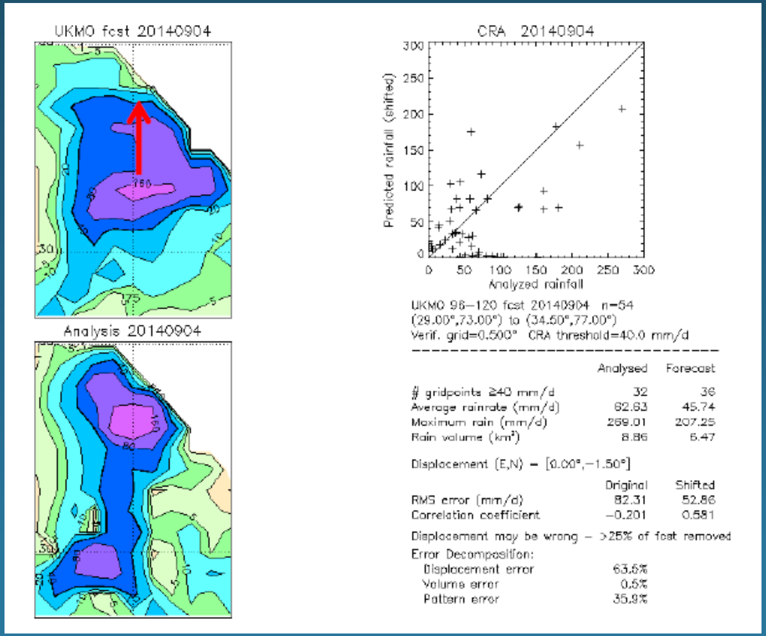
SWFDP, South Africa



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 (90^{th}%)

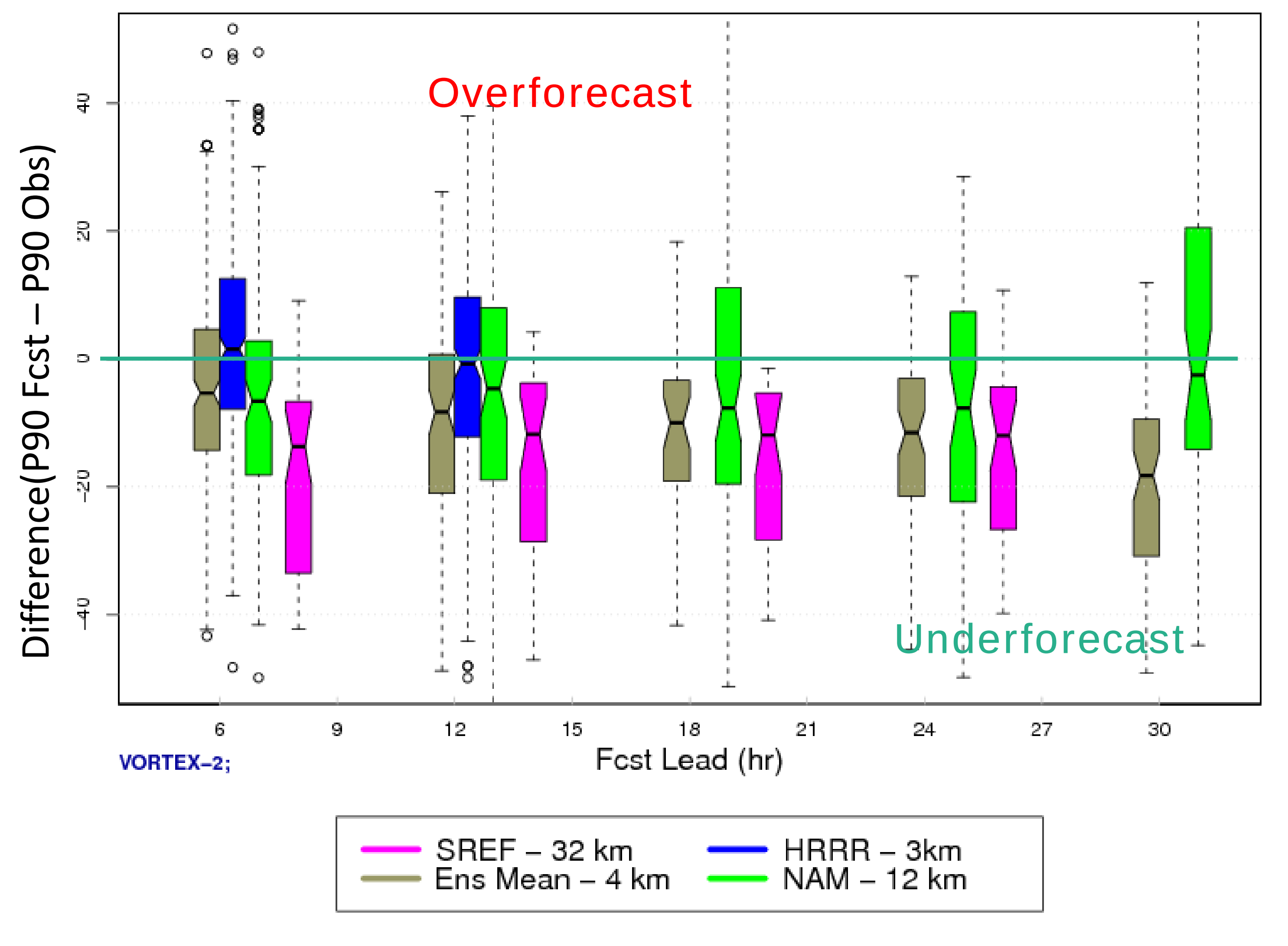
Intensity Difference (Fcst – Obs)

High Resolution Deterministic Does Fairly Well

High Resolution Ensemble Mean Underpredicts

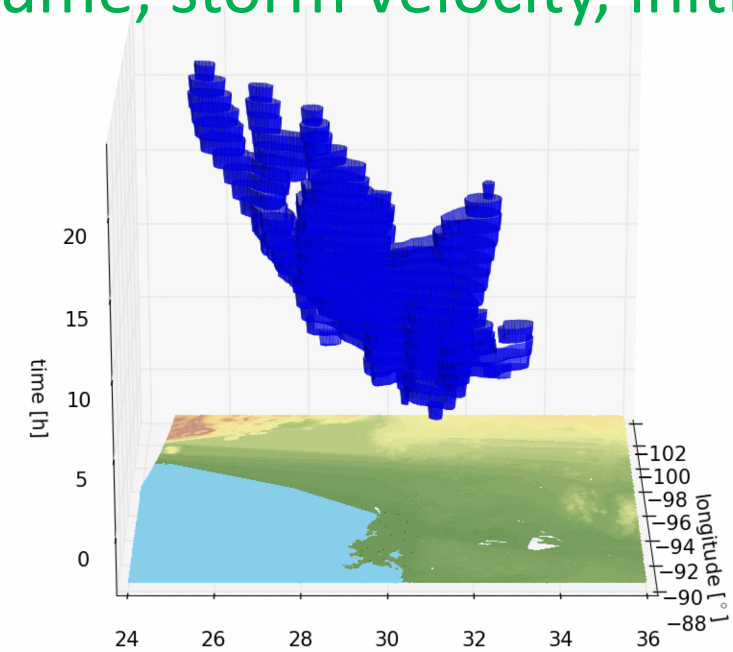
Mesoscale Deterministic Underpredicts

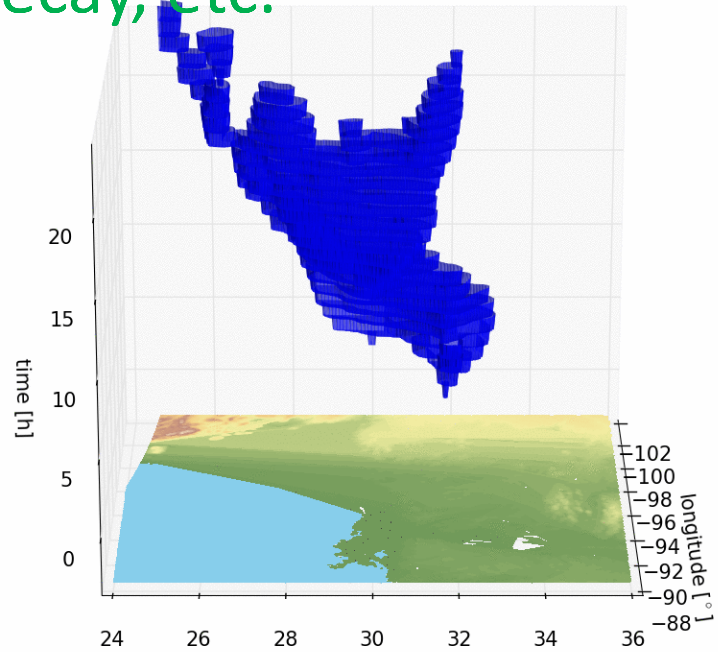
Mesoscale Ensemble Underpredicts the most



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/)

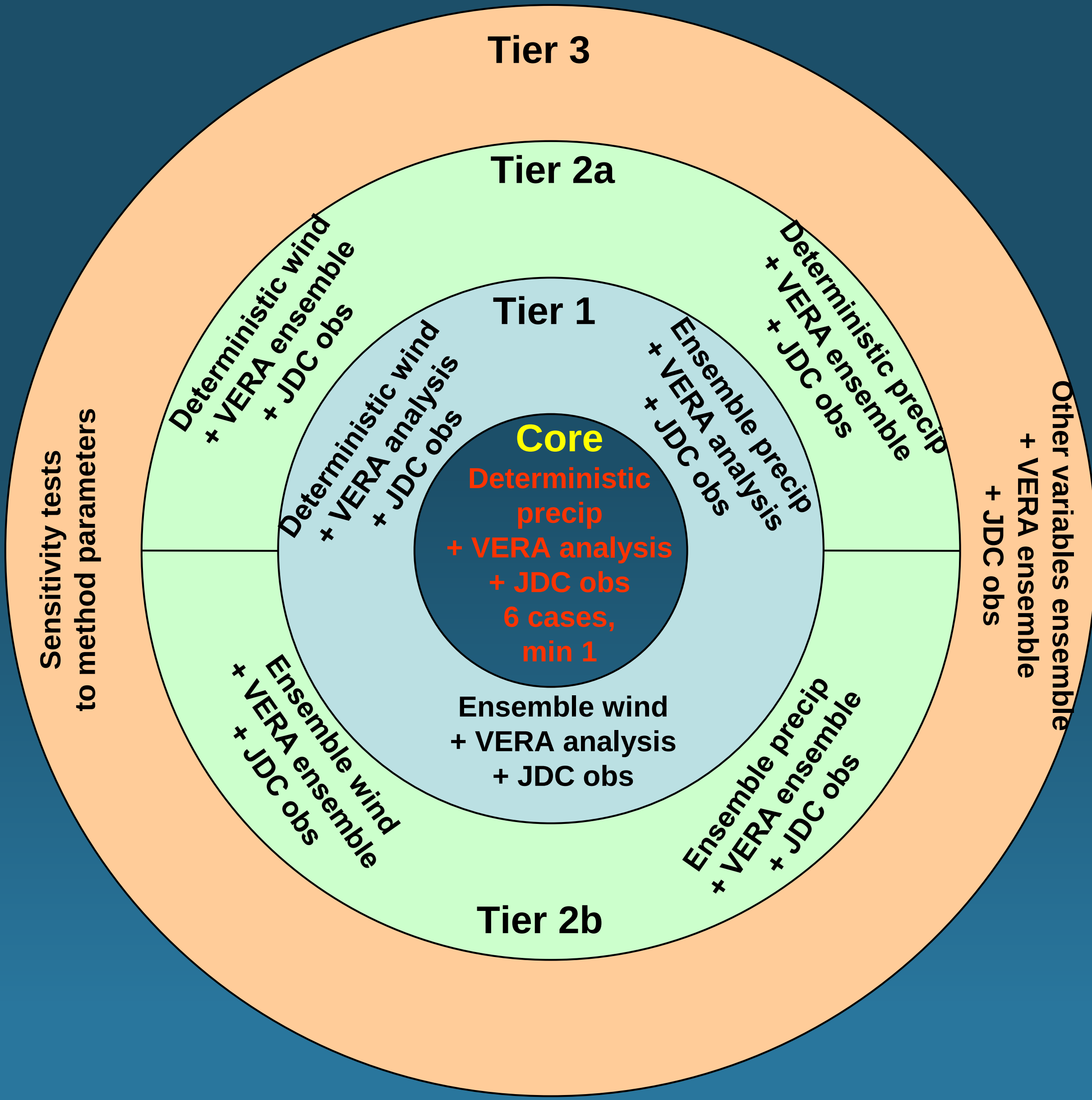
Meta-evaluation of spatial methods: *What are the capabilities of the new methods?*

- Initial intercomparison (2005-2011): Considered method capabilities for precipitation in High Plains of the US (https://www.ral.ucar.edu/projects/icp/)
- MesoVICT (Mesoscale Verification in Complex Terrain); 2013-??? considers

How do/can spatial methods:

- Transfer to other regions with complex terrain (Alpine region), and other parameters: e.g., wind (speed and direction) ?
- Work with forecast ensembles?
- Incorporate observations uncertainty (analysis ensemble)?

MesoVICT



- 3 tiers
- Complex terrain
- Mesoscale model forecasts from MAP-Dphase
- Precipitation and wind
- Deterministic and Ensemble
- Verification with VERA

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 genera;
- User-relevant verification
  - Evaluating forecasts in the context of user applications and decision making

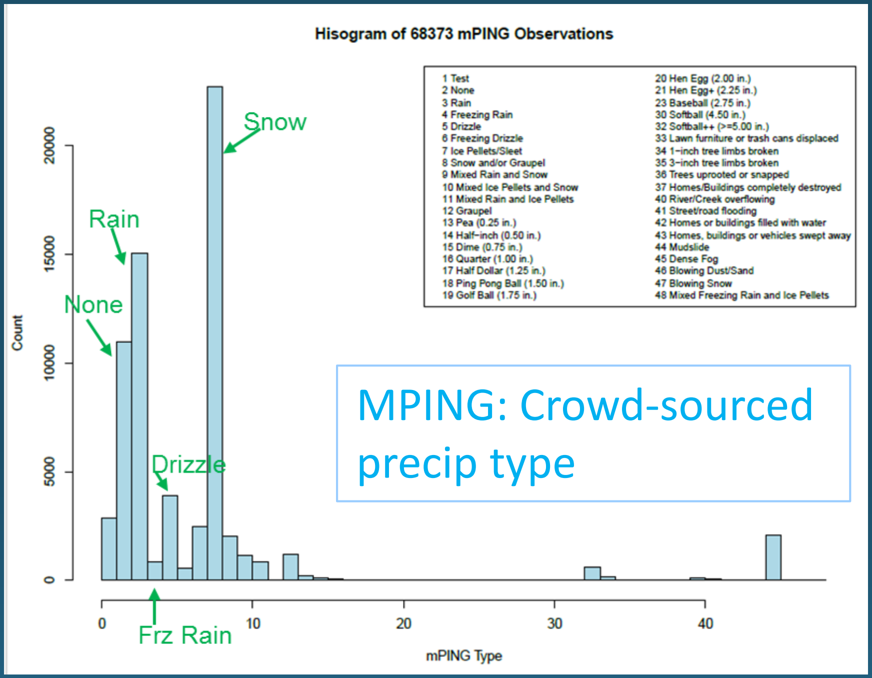
Observation limitations

Observations are still often the limiting factor in verification

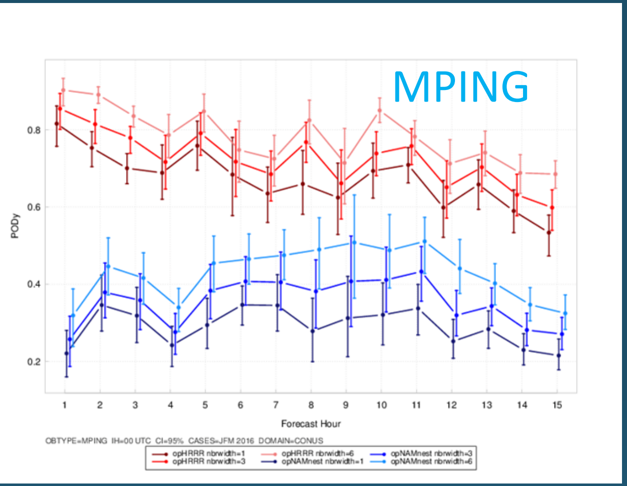
Example: Aviation weather

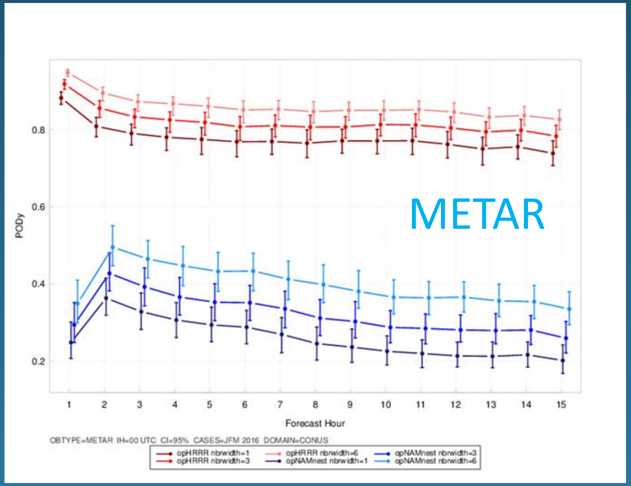
- Observations can be characterized by
  - **Sparseness**: Difficult, especially for many aviation variables (e.g., icing turbulence, precipitation type)
  - **Representativeness**: How to evaluate "analysis" products that provide nowcasts at locations with no observations?
  - **Biases**: Observations of extreme conditions (e.g., icing, turbulence) biased against where the event occurs! (pilot avoidance)
- Verification methods must take these attributes into account (e.g., choice of verification measures)

Example: Precipitation Type



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



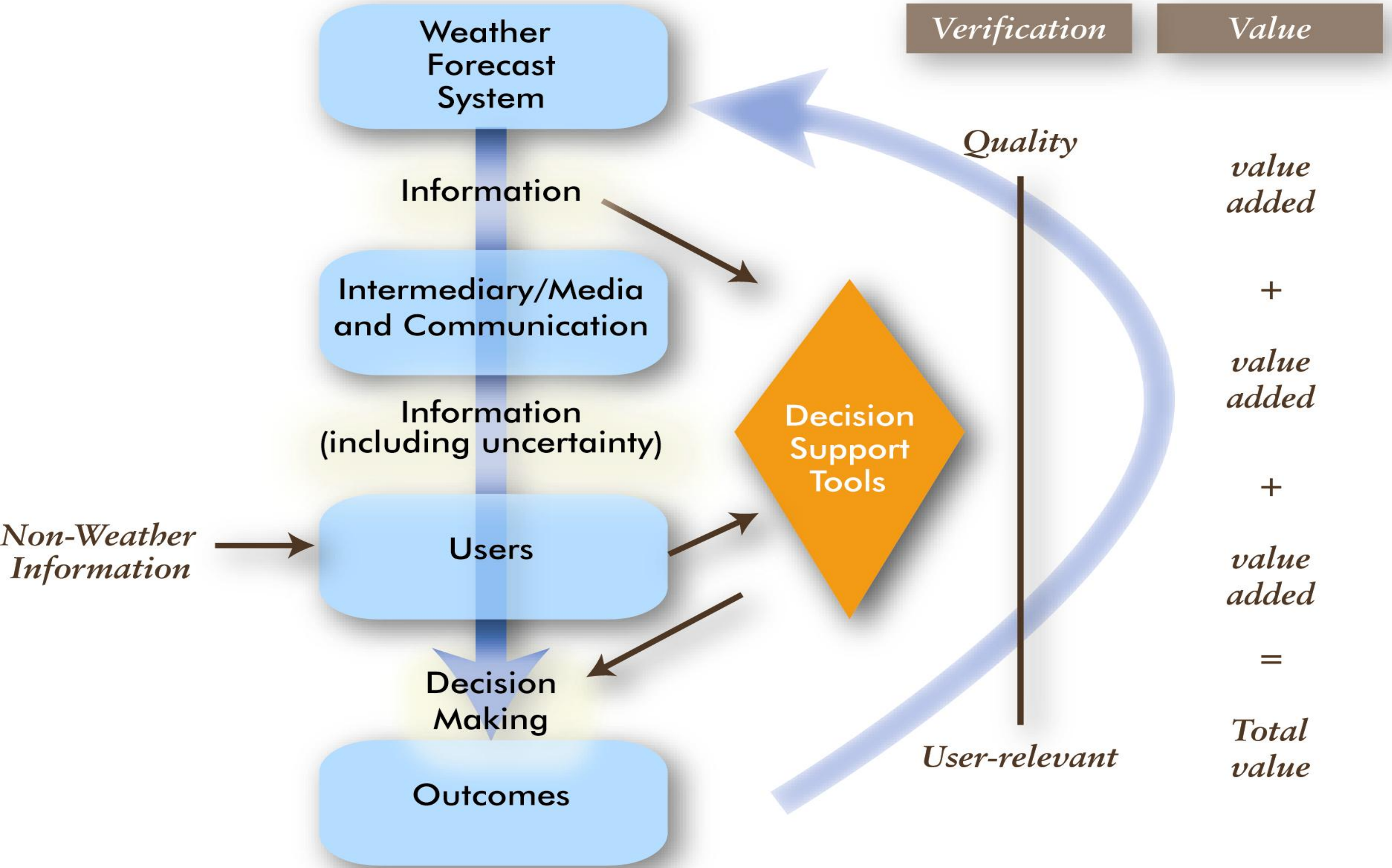


Human-generated observations have biases (e.g., in types observed)

Type of observation impacts the verification results

Credit: J. Wolff (NCAR)

Conceptual Model: Forecast Quality and Value



Morss et al. 2008 (BAMS)

User-relevant verification

Levels of user-relevance

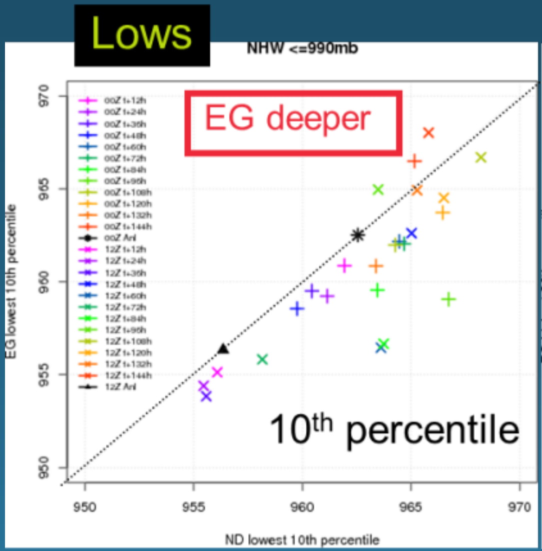
1. Making traditional verification methods useful for a range of users (e.g., variety of thresholds)
2. Developing and applying specific methods for particular users [Ex: Particular statistics; user-relevant variables]
3. Applying meaningful diagnostic (e.g., spatial) methods that are relevant for a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

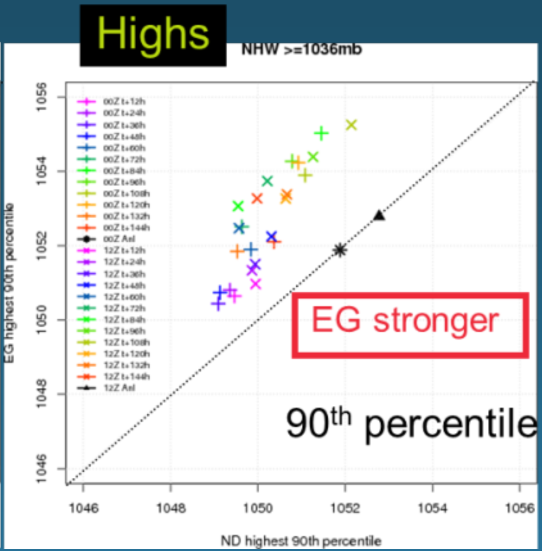
Most verification studies are at Levels 1 and 2, with some approaching 3, and very few actually at Level 4

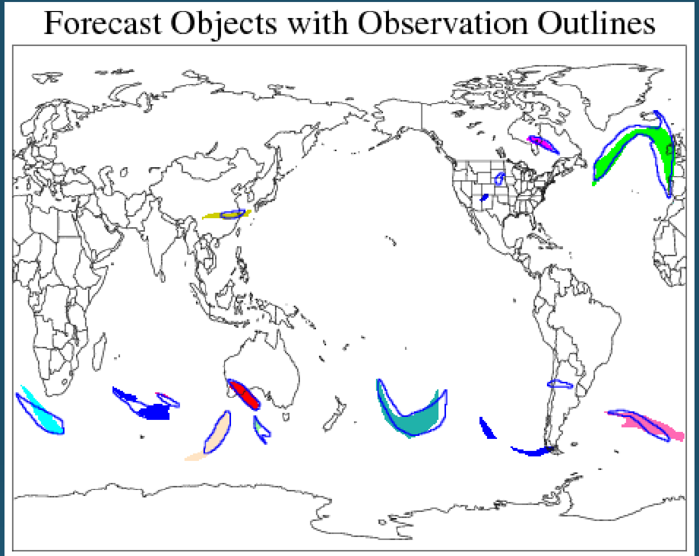
Some examples….

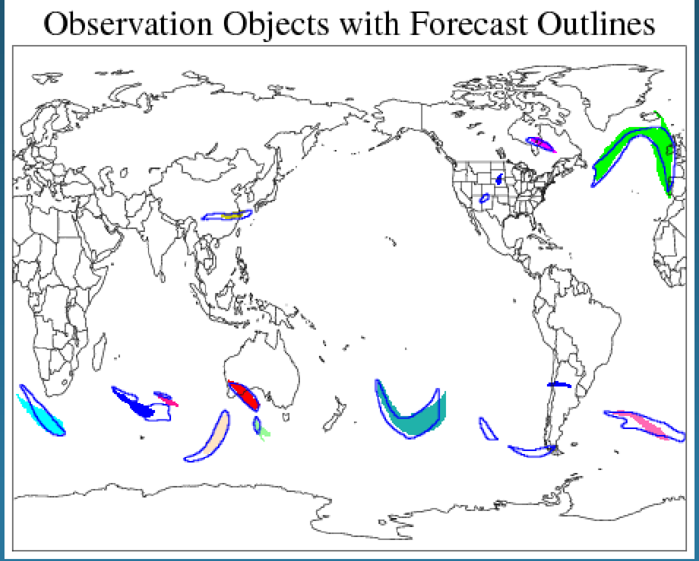
Applications of object –based approaches

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





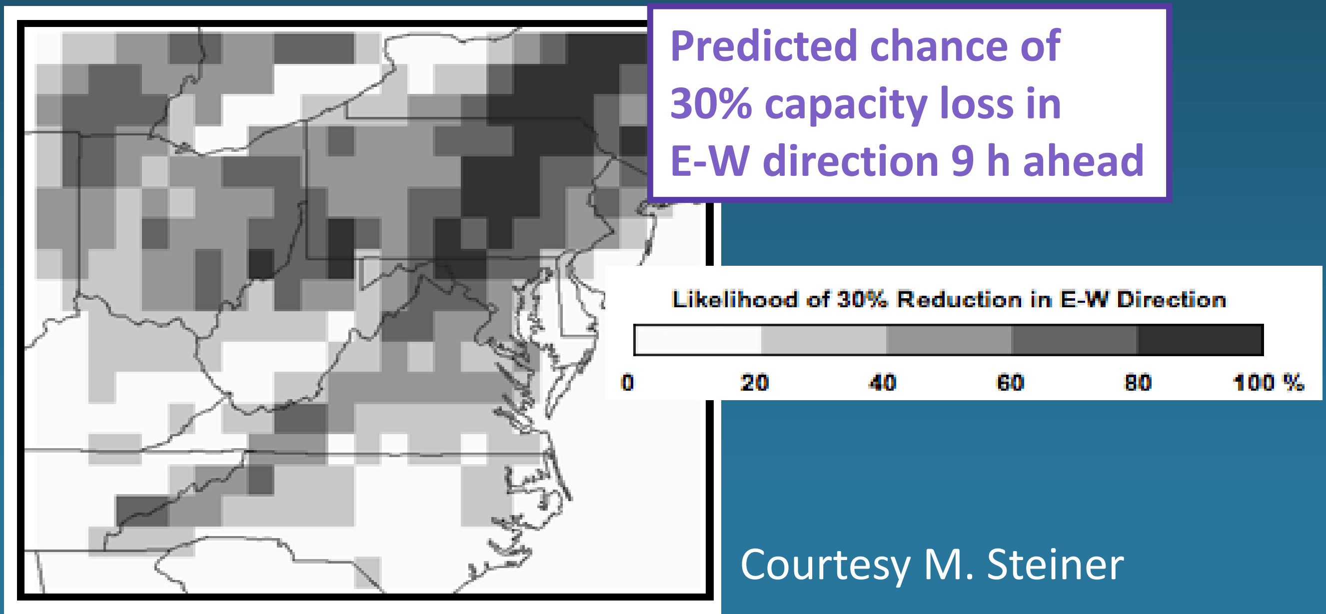




Courtesy Marion Mittermaier, UK Met Office

"User" approach to ensemble evaluation…

- Translate ensemble info into "user-relevant" information
- Evaluate on the basis of the "impact" variable
- <u>*Ideal*</u>: User-*specific* info for many users; more general, user-*relevant* info for others…



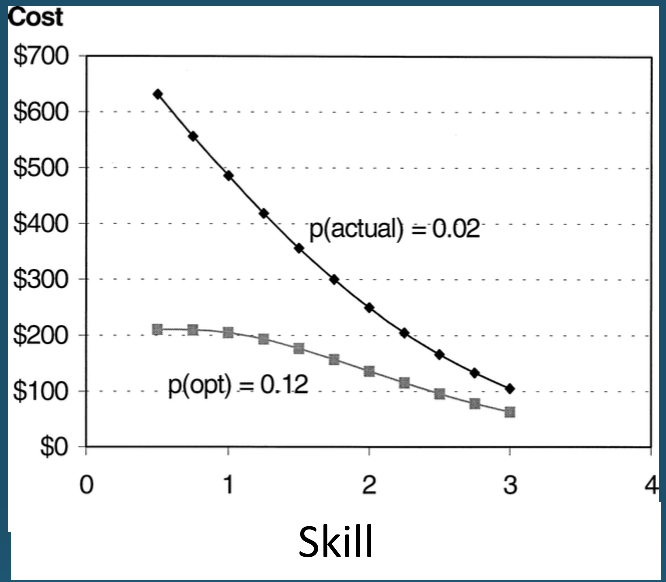
Courtesy M. Steiner

<u>Steiner:</u>
Translate convective ensembles into probability maps of aircraft "capacity"

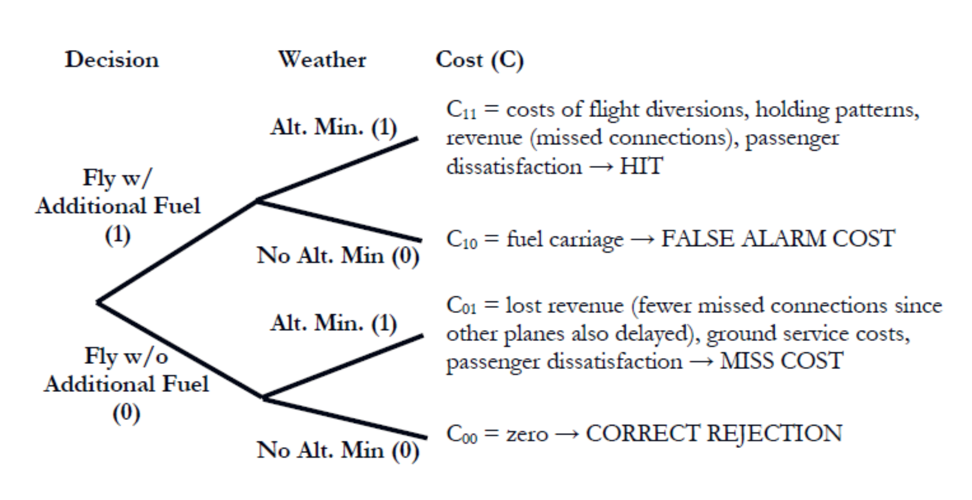
Examples of user-based forecast verification and value studies: Looking at the relationship between quality and value

Keith (2003; *Weather and Forecasting*) – Value of ceiling forecasts for fuel savings:

Cost/loss evaluation of alternate airport fuel loading needs



Keith (2005; unpublished): an average of $23K is saved per flight using probabilistic forecasts
=> Savings of approximately $50M per year in operating costs due to more optimal balance between false alarms and misses



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!

Challenge: Develop best new user-relevant verification method

- Sponsored by WMO/WWRP
  - JWGFVR (Verification Working Group)
  - High Impact Weather, Sub-seasonal to seasonal, and Polar Prediction projects
- Focus
  - All applications of weather/climate/hydro forecasts
  - Metrics can be quantitative scores or diagnostics
- Criteria for being selected as "best"
  - Originality, user relevance, simplicity, robustness, resistance to hedging.
  - Desirable characteristics:
    - (i) Clear statistical foundation;
    - (ii) Applicability to a broader set of problems

Challenge: Develop best new user-relevant verification method

- Deadline for submission: 31 Oct 2016
- Prize: Invited keynote talk at the 7th International Verification Methods Workshop in May, 2017 (Berlin)
- 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 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…

Remaining challenges (some examples)

- Expansion of user-relevant metrics

  Providing a breadth of information to users
- Sorting out how to incorporate uncertainty appropriately
  - *Spatial / Temporal*
  - *Measurement / Observation*
  - *Sampling*
- Improving communication

  Developing ways to communicate forecast quality information to the general public, specific users