FORECAST VERIFICATION

T-NOTE – Buenos Aires, 5-16 August 2013 – Celeste Saulo and Juan Ruiz

ACKNOWLEDGMENTS

This material has been prepared following the lectures:

- Introduction to Verification of Forecasts and Nowcasts (WSN12) by Rita Roberts[1], Barbara Brown[1], Beth Ebert[2], and Tressa Fowler[1]
- ECMWF training lectures on Ensamble Verification by Renate Hagedorn
- Lectures from the 4th International Verification Methods Workshop, Helsinki, 2009. WWRP 2010 -1, in particular those by: B. Ebert, B. Brown, L. Wilson.
- "the" link for continuous update on new methods, examples, explanations, etc. etc.:
- http://www.cawcr.gov.au/projects/verification/
- WMO Working Group on Forecast Verification Research

OUTLINE

- General ideas about verification: why verify?, what verify?
- Finding “your score”. Steps in verification
- Reference data set
- Scores for Spatial verification
- Probabilistic verification

WHAT DO WE MEAN BY FORECAST *VERIFICATION*?

→ Measuring the quality of a forecast by comparison with observations

A forecast is like an experiment...

You make a hypothesis about what will happen.

You would not consider an experiment to be complete until you found out what happened.

→ VERIFICATION

→ *Verification* is a critical part of the *forecasting process*

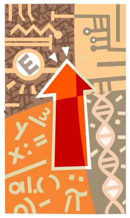
WHAT IS VERIFICATION?

- 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 to a user***)
- For many purposes a more appropriate term is "***evaluation***"

¿WHY VERIFY?

- Scientific purposes (understand sources of model errors)
- Monitor forecast quality (administrative-economic issues)
- Quantify model errors so that we can
  - Help operational forecasters understand model biases and select models for use in different conditions
  - Help "users" interpret forecasts (e.g., "What does a temperature forecast of 0 degrees really mean?")
  - Identify forecast weaknesses, strengths, differences

WHY VERIFY FORECASTS?



To show that your forecasts have a positive ***impact*** → ***Skill scores, value scores***

To ***monitor*** whether your forecasts are improving over time → ***Summary scores***

To ***evaluate*** and ***compare*** forecasting systems → ***Continuous and categorical scores***

To ***understand*** the errors, so that you can improve the forecasts → ***Diagnostic methods***

OUTLINE

- General ideas about verification: why verify?, what verify?
- **Finding "your score". Steps in verification**
- Reference data set
- Scores for Spatial verification
- Probabilistic verification

The verification method(s) you should use depends on the purpose of the verification

You will be measuring a specific *__attribute__*

VERIFICATION STEPS

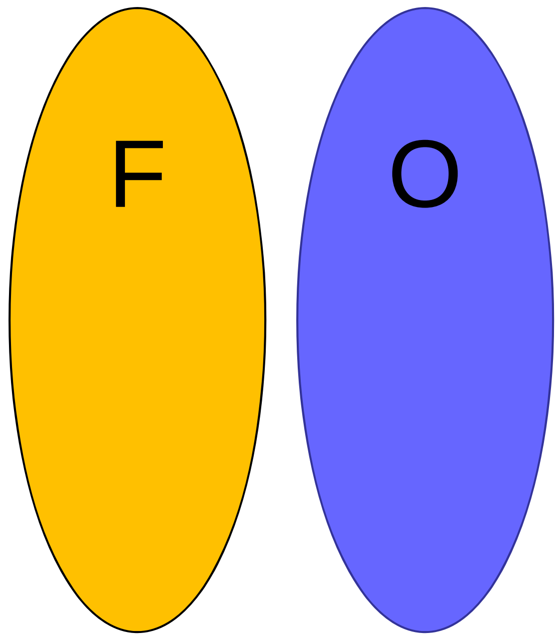
- Identify multiple ***verification attributes*** that can provide answers to the questions of interest
- Select ***measures and graphics*** that appropriately measure and represent the attributes of interest
- Identify a ***standard of comparison*** that provides a reference level of skill (e.g., persistence, climatology, old model)

WHAT MAKES A FORECAST GOOD?

Allan Murphy (1993) distinguished three types of "goodness":

- ***Consistency*** - the degree to which the forecast corresponds to the forecaster's best judgement about the situation, based upon his/her knowledge base
- ***Quality*** - the degree to which the forecast corresponds to what actually happened
- ***Value*** - the degree to which the forecast helps a decision maker to realize some incremental economic and/or other benefit

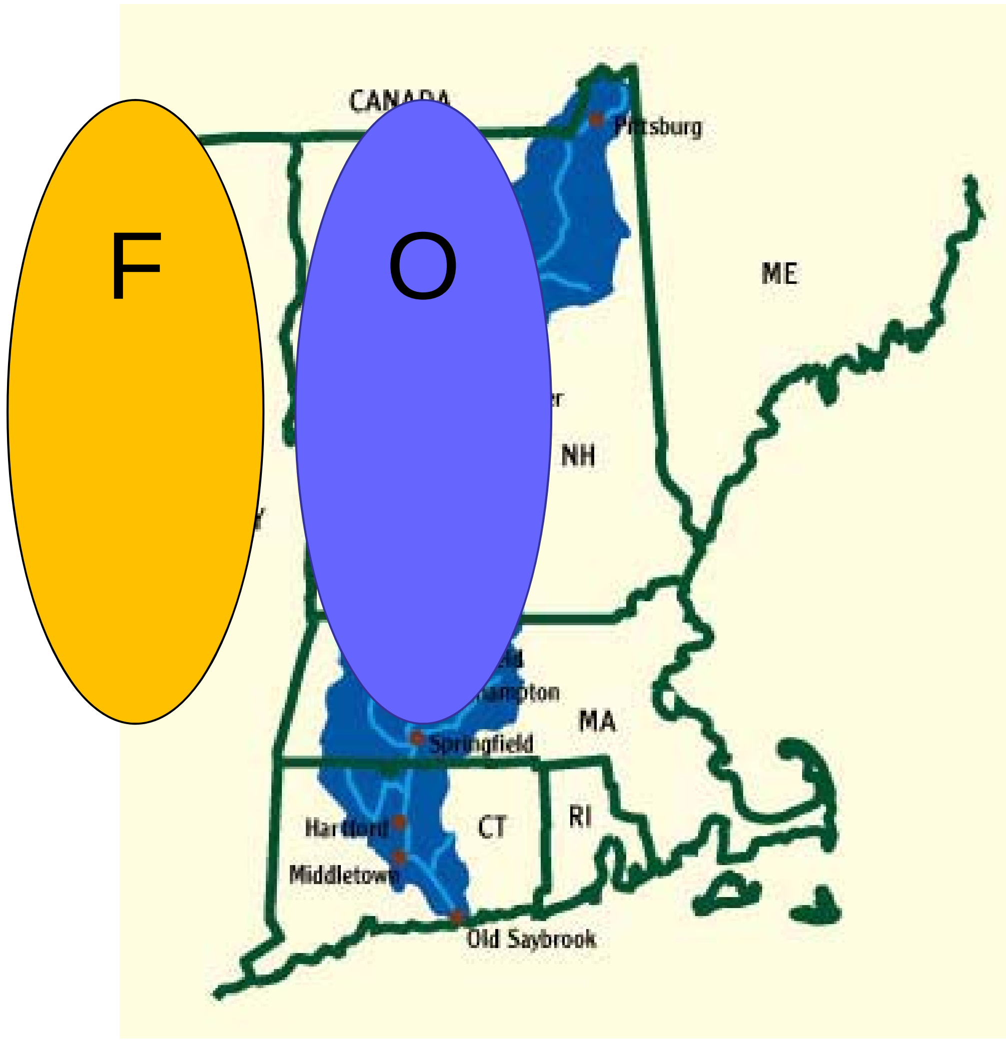
GOOD FORECAST OR BAD FORECAST?



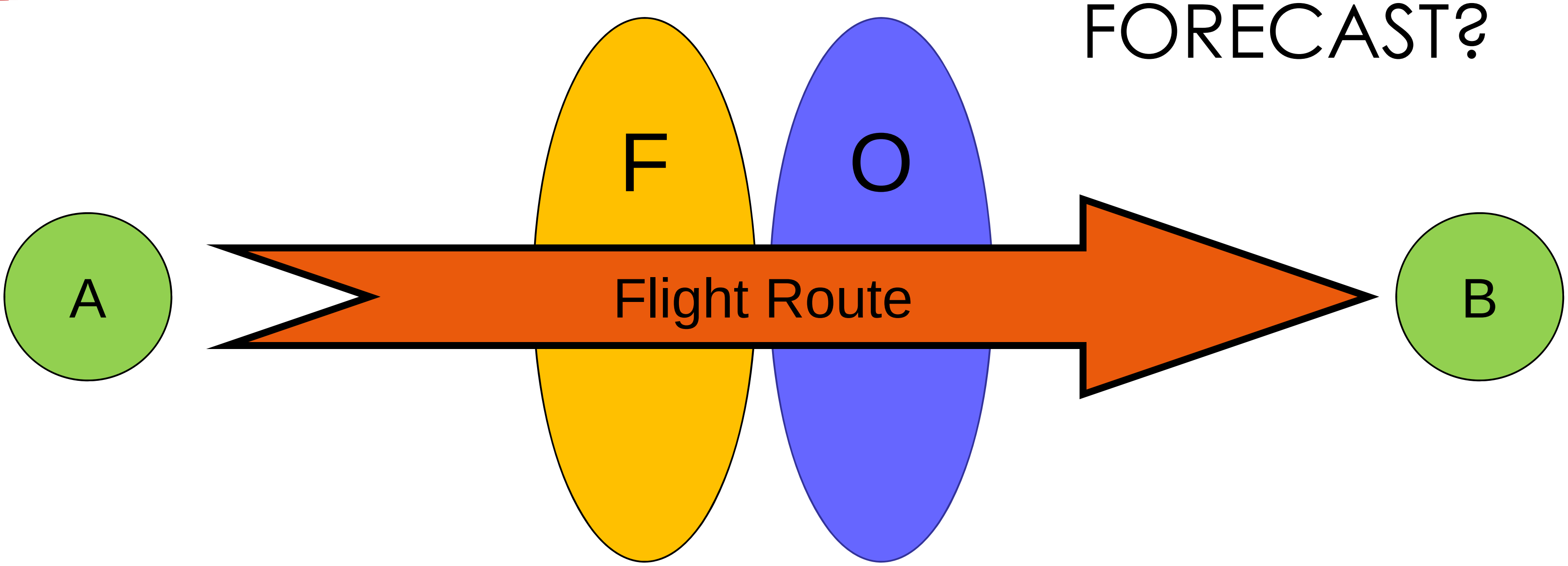
Many verification approaches would say that this forecast has NO skill and is very inaccurate.

GOOD FORECAST OR BAD FORECAST?

If I'm a water manager for this watershed, it's a pretty bad forecast…



GOOD FORECAST OR BAD FORECAST?



If I’m an aviation traffic strategic planner…

It might be a pretty good forecast

Different users have different ideas about what makes a forecast good

Different verification approaches can measure different types of “goodness”

FORECAST "GOODNESS"

- Forecast **quality** is only one aspect of forecast "goodness"
- Forecast **value** is related to forecast quality through complex, non-linear relationships
- ***However*** - Some approaches to measuring forecast quality can help understand goodness
  - Examples
    - Diagnostic verification approaches
    - New features-based approaches
    - Use of multiple measures to represent more than one attribute of forecast performance
    - Examination of multiple thresholds

OUTLINE

- General ideas about verification: why verify?, what verify?
- Finding “your score”. Steps in verification
- **Reference data set**
- Scores for Spatial verification
- Probabilistic verification

WHAT IS TRUTH WHEN VERIFYING A FORECAST?

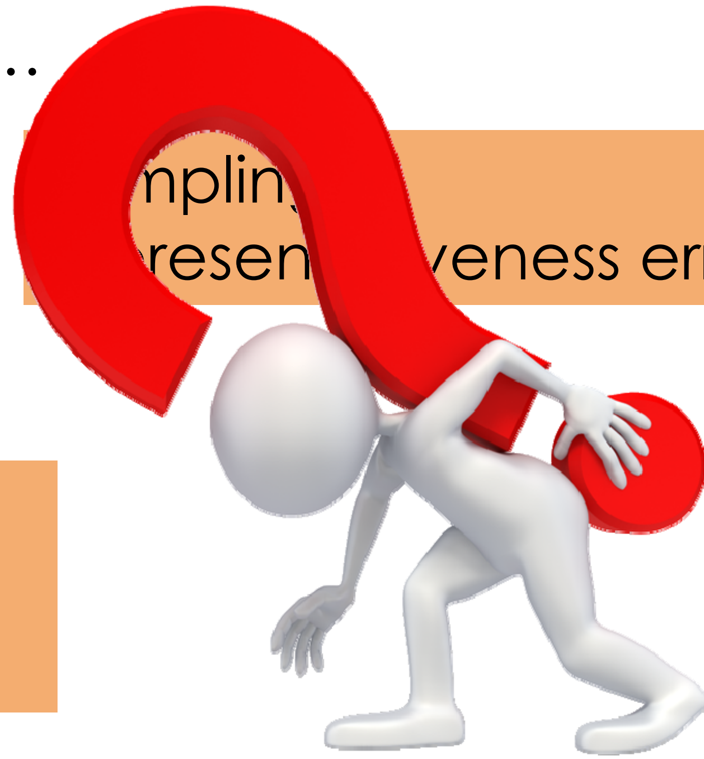
The "truth" generally comes from observational data

but…

Errors in the measurement

…nplin…
…resen…veness errors

We also have uncertainty in the verification!!



Uncertainty in scores and measures should be estimated whenever possible!

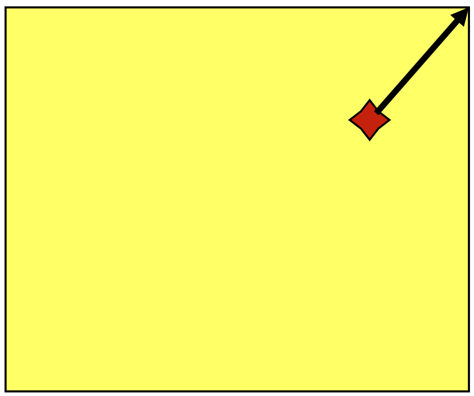
- Uncertainty arises from
  - Sampling variability
  - Observation error
  - Representativeness differences
  - Others?
- Erroneous conclusions can be drawn regarding improvements in forecasting systems and models
- Methods for *confidence intervals* and *hypothesis tests*
  - Parametric (i.e., depending on a statistical model)
  - Non-parametric (e.g., derived from re-sampling procedures, often called "bootstrapping")

MATCHING FORECASTS AND OBSERVATIONS

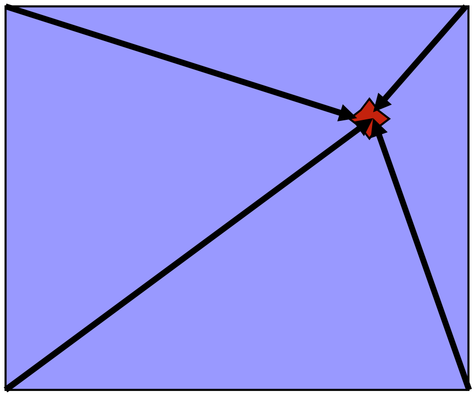
- May be the ***most difficult*** part of the verification process!
- Many factors need to be taken into account
  - Identifying observations that represent the forecast event
  - For a gridded forecast there are many options for the matching process
    - Point-to-grid
      - Match obs to closest gridpoint
    - Grid-to-point
      - Interpolate?
      - Take largest value?

MATCHING FORECASTS AND OBSERVATIONS

- Point-to-Grid and
  Grid-to-Point



- Matching approach can impact the results of the verification



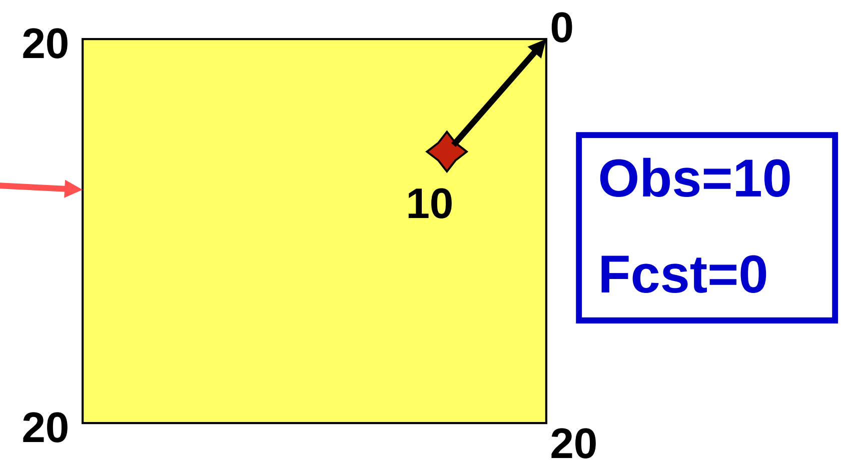
MATCHING FORECASTS AND OBSERVATIONS

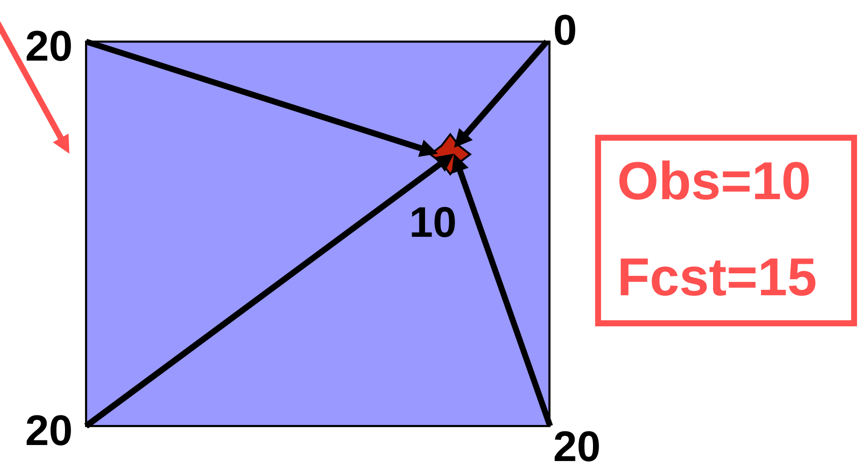
Example:

- Two approaches:
  - Match rain gauge to nearest gridpoint ***or***
  - Interpolate grid values to rain gauge location
    - Crude assumption: equal weight to each gridpoint
- Differences in results associated with matching:

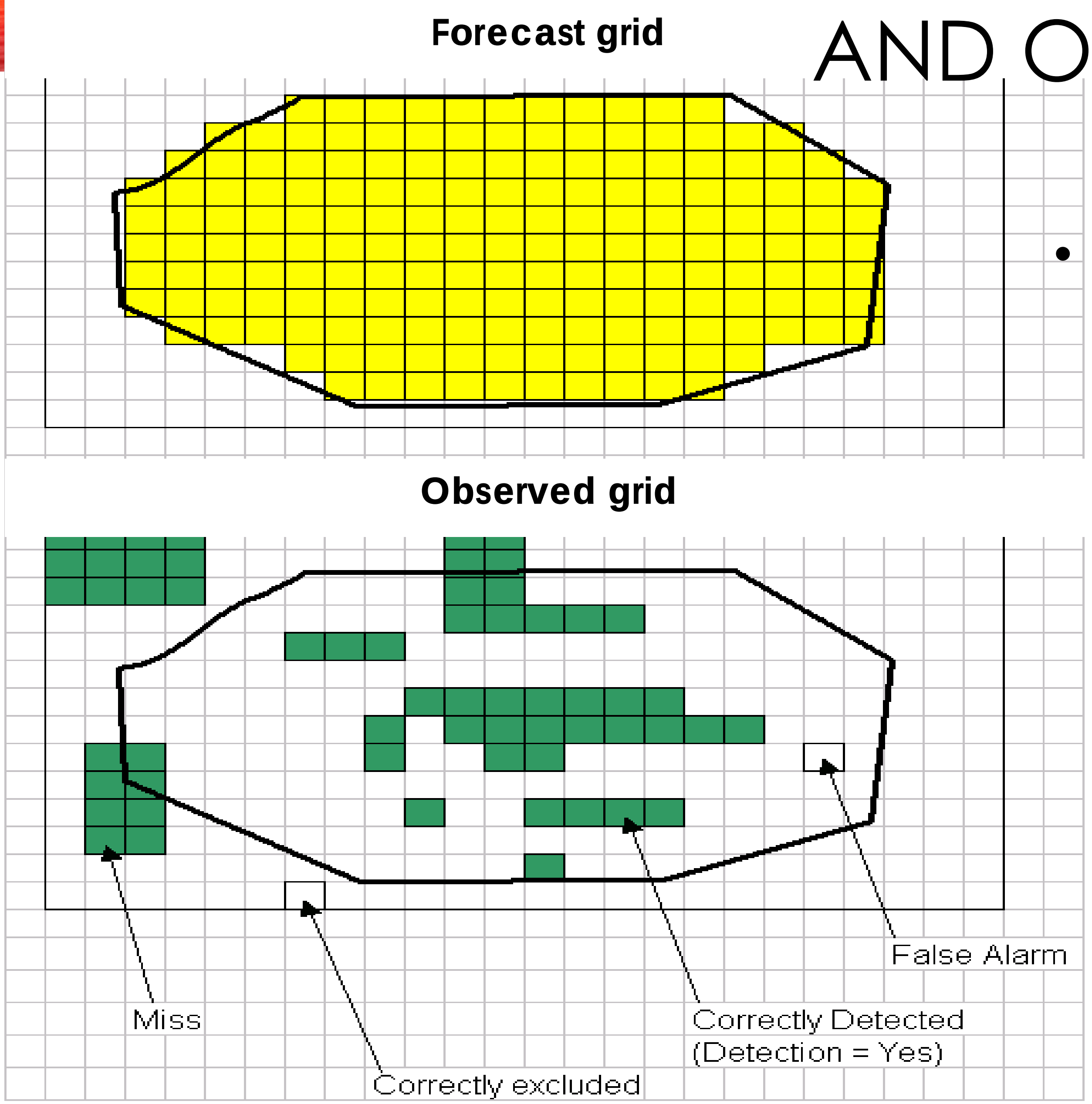
"Representativeness" difference

Will impact most verification scores





MATCHING FORECASTS AND OBSERVATIONS



- Grid to grid approach
  - Overlay forecast and observed grids
  - Match each forecast and observation

MATCHING FORECASTS AND OBSERVATIONS

<u>Final point:</u>

- It is not advisable to use the model analysis as the verification "observation"

- Why not??

- **Issue: Non-independence!!**

OUTLINE

- General ideas about verification: why verify?, what verify?
- Finding “your score”. Steps in verification
- Reference data set
- **Scores for Spatial verification**
- Probabilistic verification

SKILL SCORES

- A skill score is a measure of *relative performance*
  - ***<u>Ex</u>***: *How much more accurate are my temperature predictions than climatology? How much more accurate are they than the model's temperature predictions?*
  - *Provides a comparison to a* ***standard***
- Generic skill score definition:

$$\frac{M - M_{ref}}{M_{perf} - M_{ref}}$$

  Where M is the verification measure for the forecasts, M_{ref} is the measure for the reference forecasts, and M_{perf} is the measure for perfect forecasts
- Positively oriented (larger is better)
- Choice of the standard matters (*a lot*!)

METHODS FOR VERIFYING SPATIAL FORECASTS
VISUAL ("EYEBALL") VERIFICATION

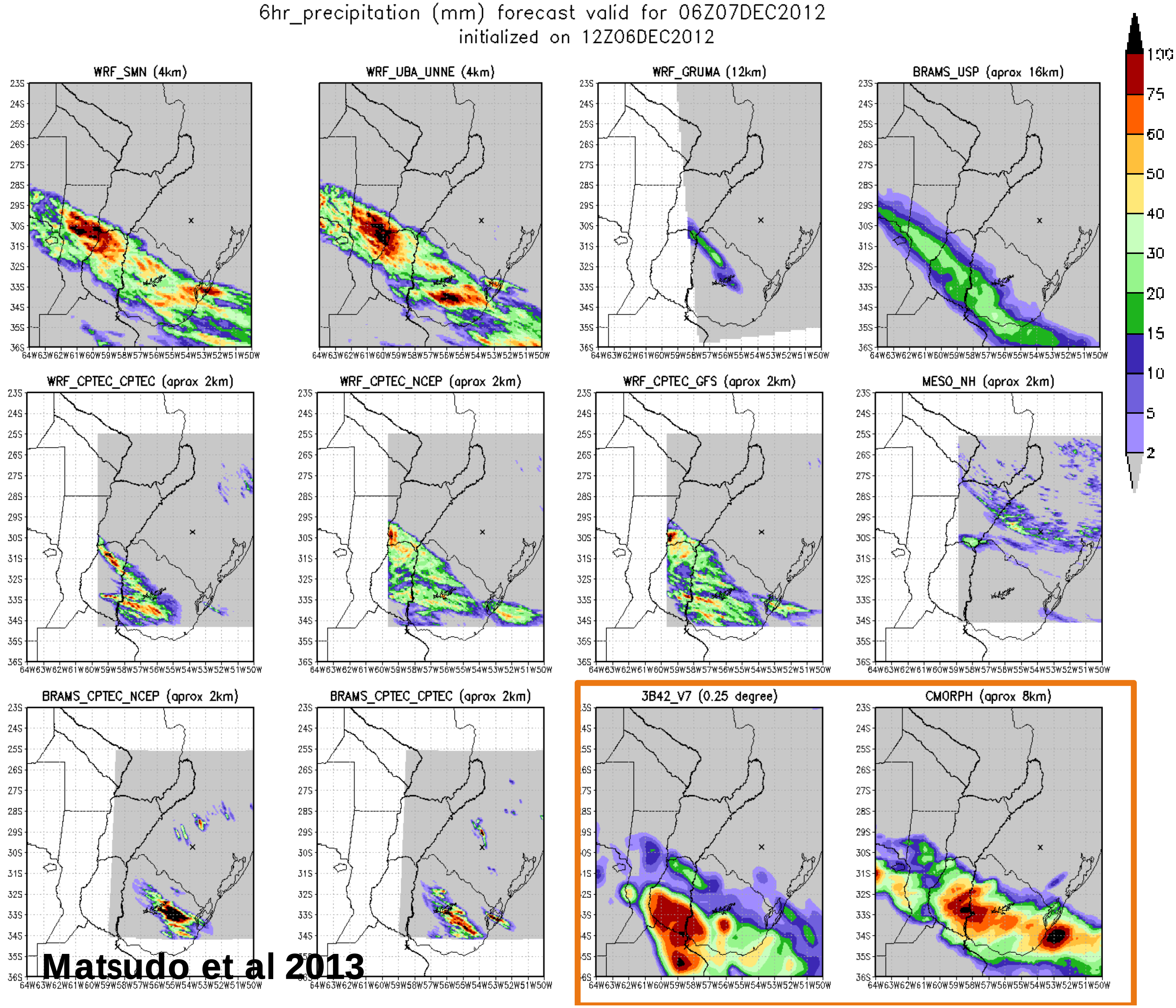
Visually compare maps of forecast and observations

Advantage: "A picture tells a thousand words…"

Disadvantages: Labor intensive, not quantitative, subjective

Chuva Case: 05 Dec 2012

18-hr forecast



Matsudo et al 2013

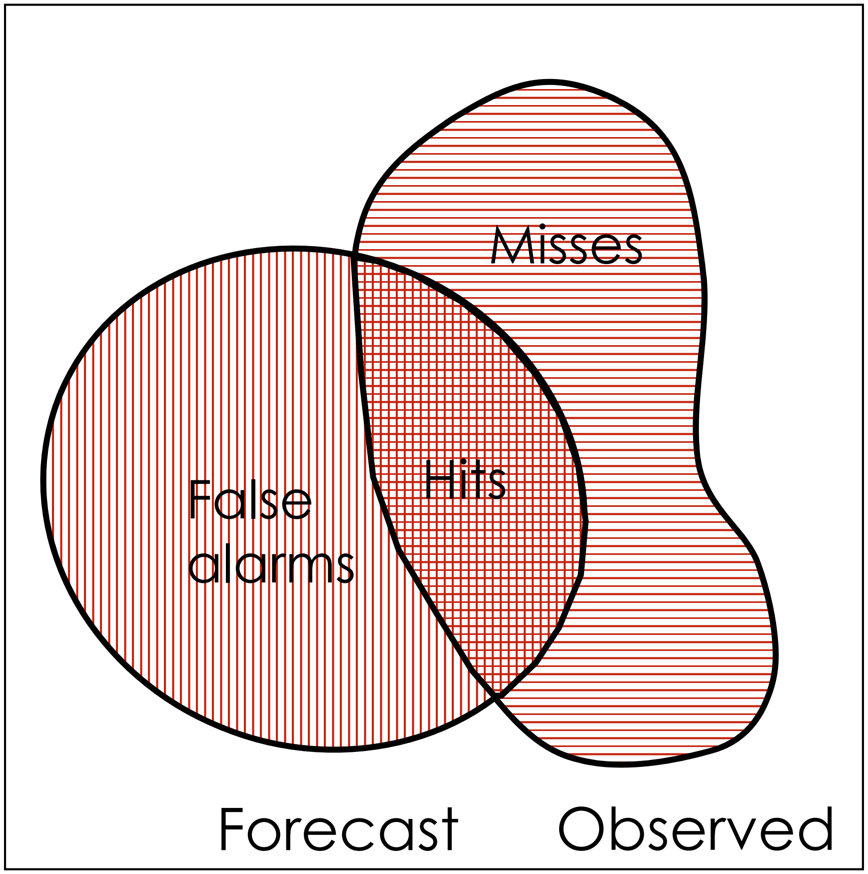
TRADITIONAL VERIFICATION APPROACHES

Compute statistics on forecast-observation pairs

- Continuous values (e.g., precipitation amount, temperature, NWP variables):
  - mean error, MSE, RMSE, correlation
  - anomaly correlation, S1 score
- Categorical values (e.g., precipitation occurrence):
  - Contingency table statistics (POD, FAR, CSI, equitable threat score,…)

TRADITIONAL SPATIAL VERIFICATION USING CATEGORICAL SCORES

Contingency Table



		Observed yes	**Observed** no
Predicted	yes	*hits*	*false alarms*
	no	*misses*	*correct negatives*

$$BIAS = \frac{hits + false\ \ alarms}{hits + misses}$$

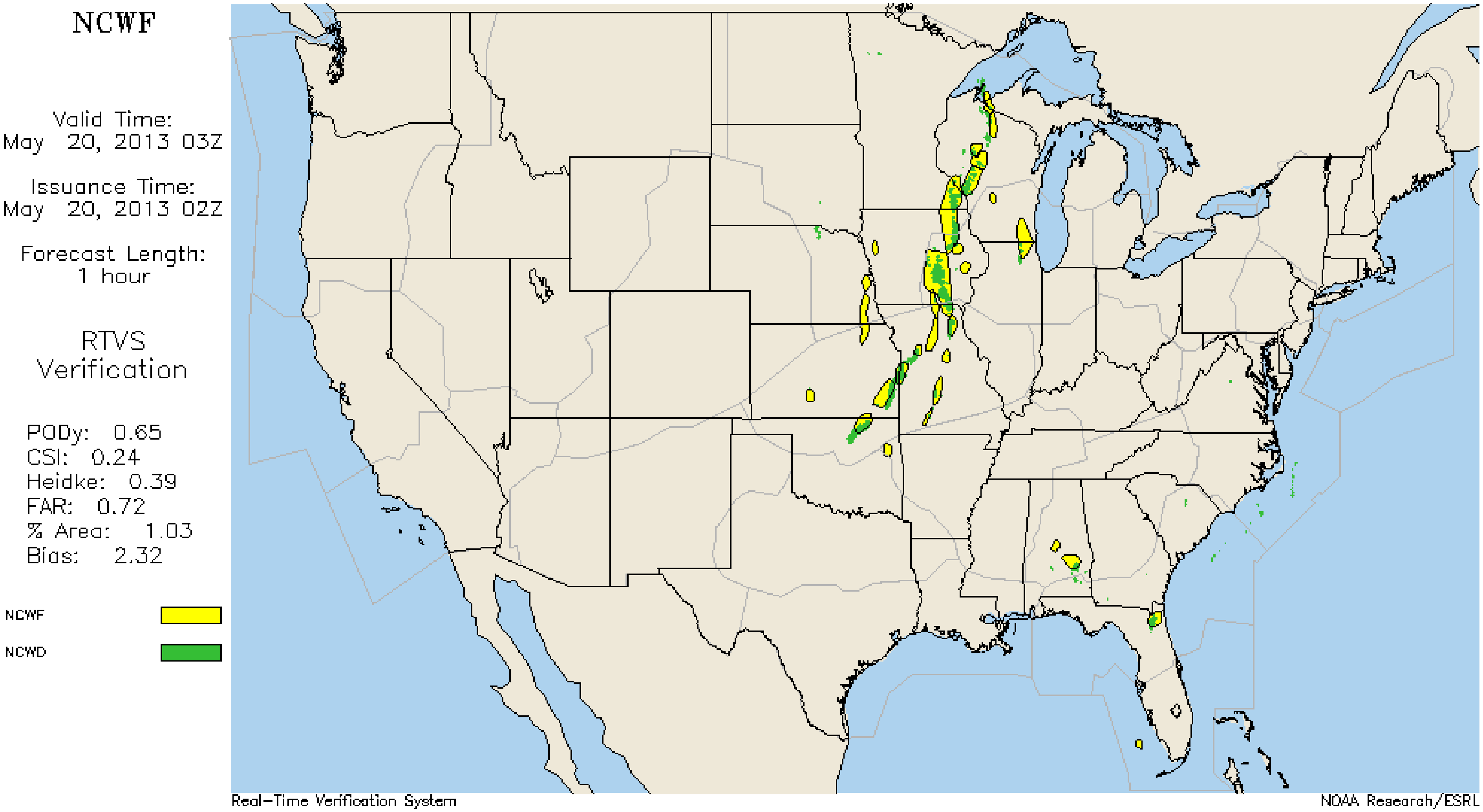
$$POD = \frac{hits}{hits + misses} \qquad FAR = \frac{false\ alarms}{hits + false\ alarms}$$

$$TS = \frac{hits}{hits + misses + false\ alarms} = CSI$$

$$ETS = \frac{hits - hits_{random}}{hits + misses + false\ alarms - hits_{random}}$$

National Convective Weather Forecast Product (NCWF)

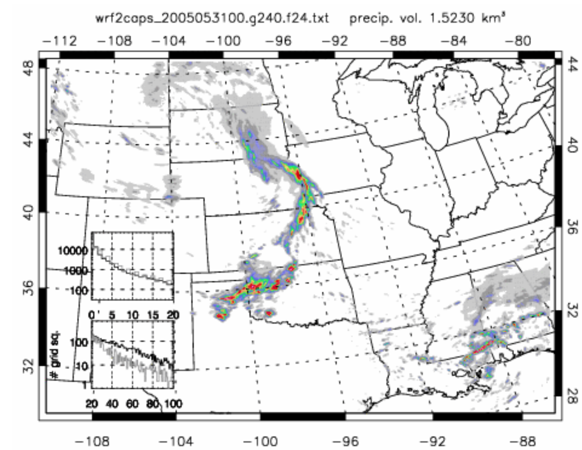
POD = 0.65   CSI= 0.24   FAR = 0.72   BIAS = 2.32



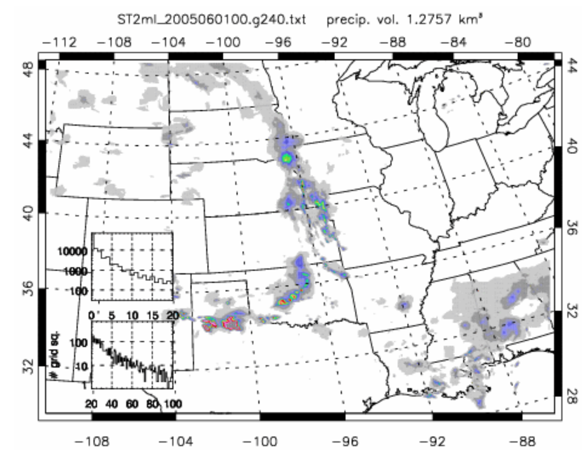
SPATIAL FORECASTS

Weather variables defined over spatial domains have **coherent spatial structure and features**

WRF model



Stage II radar



New spatial verification techniques aim to:

- account for field spatial structure
- provide information on error in physical terms
- account for uncertainties in location (and timing)

NEW SPATIAL VERIFICATION APPROACHES

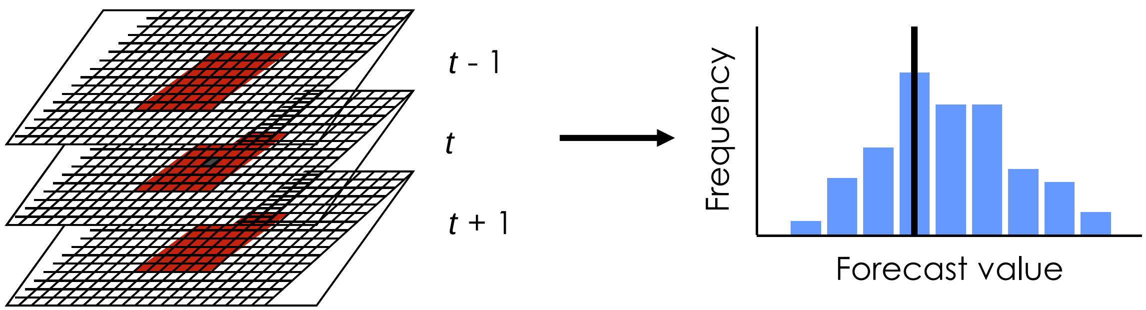
- Neighborhood (fuzzy) verification methods
  - give credit to "close" forecasts
- Scale decomposition methods
  - measure scale-dependent error
- Object-oriented methods
  - evaluate attributes of identifiable features
- Field verification
  - evaluate phase errors

NEIGHBORHOOD (FUZZY) VERIFICATION METHODS
→ GIVE CREDIT TO "CLOSE" FORECASTS

NEIGHBORHOOD VERIFICATION METHODS

- Don't require an exact match between forecasts and observations
  - Unpredictable scales
  - Uncertainty in observations

- Look in a space / time neighborhood around the point of interest



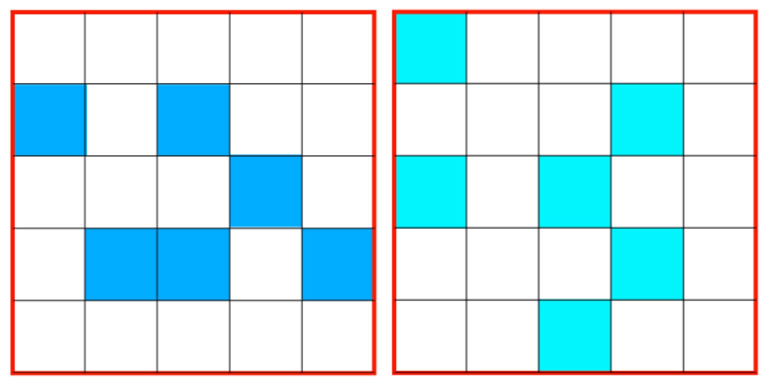
  - Evaluate using categorical, continuous, probabilistic scores / methods

FRACTIONS SKILL SCORE
(ROBERTS AND LEAN, *MWR*, 2008)

- We want to know
  - How forecast skill varies with neighborhood size
  - The smallest neighborhood size that can be can be used to give sufficiently accurate forecasts
  - Does higher resolution NWP provide more accurate forecasts on scales of interest (e.g., river catchments)

Compare forecast fractions with observed fractions (radar) in a *probabilistic* way over different sized neighbourhoods

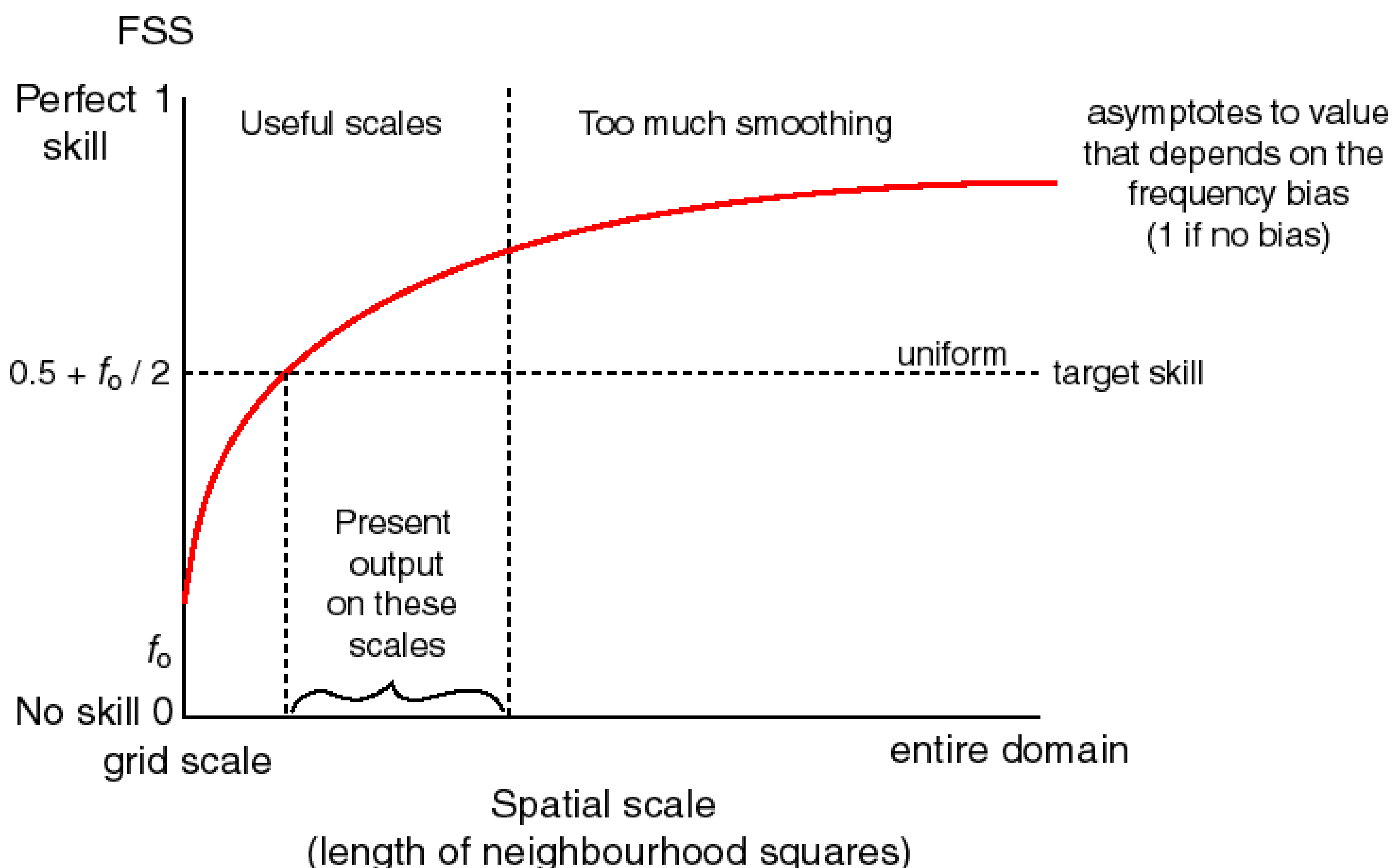
$$FSS = 1 - \frac{\frac{1}{N}\sum_{i=1}^{N}(P_{fcst} - P_{obs})^2}{\frac{1}{N}\sum_{i=1}^{N}P_{fcst}^2 + \frac{1}{N}\sum_{i=1}^{N}P_{obs}^2}$$



Fraction = 6/25 = 0.24 observed

Fraction = 6/25 = 0.24 forecast

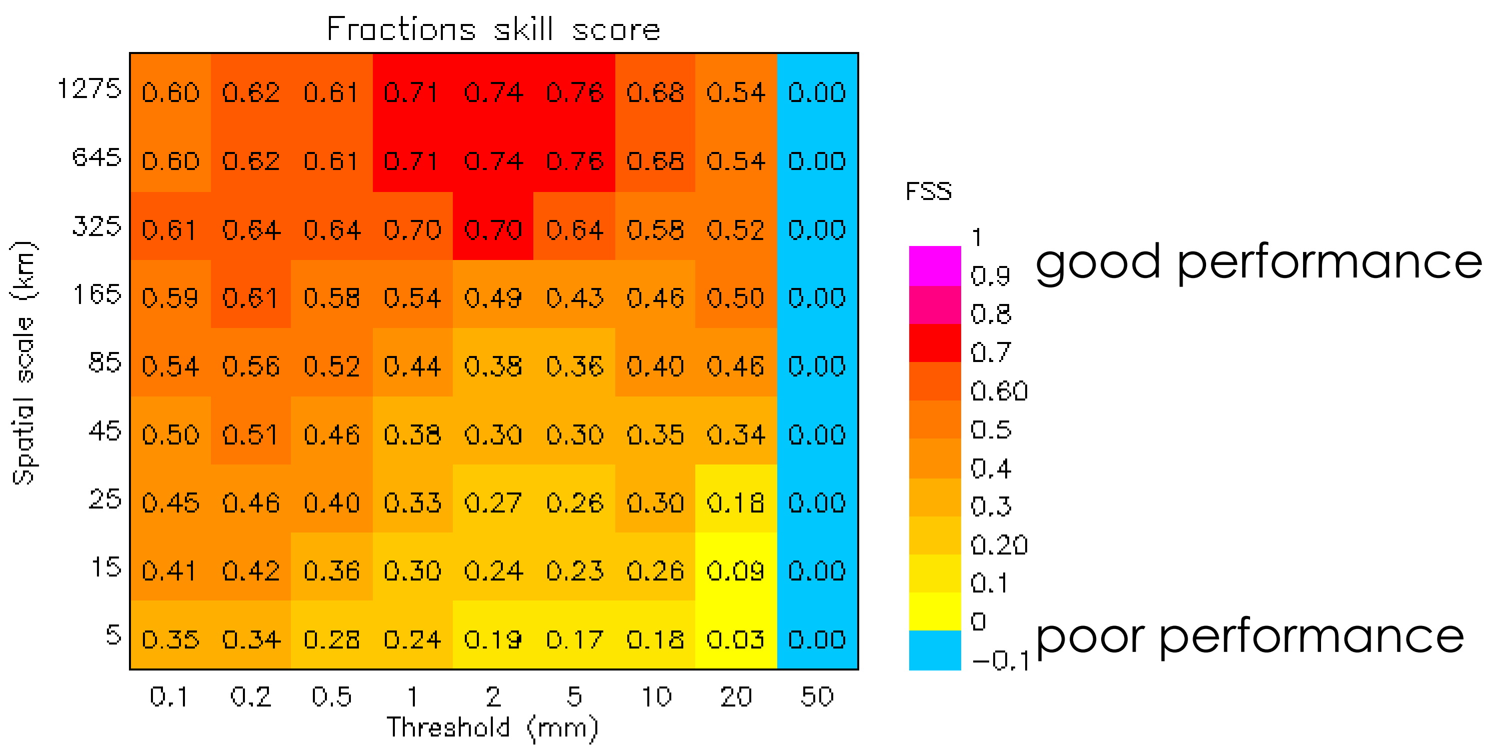
FRACTIONS SKILL SCORE
(ROBERTS AND LEAN, *MWR*, 2008)



f_o=domain obs fraction

MULTI-SCALE, MULTI-INTENSITY APPROACH

- Forecast performance depends on the scale and intensity of the event



OUTLINE

- General ideas about verification: why verify?, what verify?
- Finding “your score”. Steps in verification
- Reference data set
- Scores for Spatial verification
- **Probabilistic verification**

VERIFICATION OF PROBABILISTIC FORECASTS

- The forecast now is a function of $f(x,y,z,t,e)$
- How do we deal with added dimension when
  - interpreting, verifying and diagnosing EPS output?

Transition from deterministic (yes/no) to probabilistic

ASSESSING THE QUALITY OF A PROBABILISTIC FORECAST

- The forecast indicated 10% probability for rain
- It did rain on the day
- Was it a good forecast?
  - □ Yes
  - □ No
  - □ I don't know
- Single probabilistic forecasts are never completely wrong or right (unless they give 0% or 100% probabilities)
- To evaluate a forecast system we need to look at a (large) number of forecast–observation pairs

ASSESSING THE QUALITY OF A PROBABILISTIC FORECAST

- Brier score: *Accuracy*
- Brier skill score: *Skill*
- Reliability Diagrams measure:
  - **Reliability**: Can I trust the probabilities to mean what they say?
  - **Sharpness**: How much do the forecasts differ from the climatological mean probabilities of the event?
  - **Resolution**: How much do the forecasts differ from the climatological mean probabilities of the event, and the systems gets it right?

BRIER SCORE

- The Brier score is a measure of the accuracy of probability forecasts

- Considering *N* forecast – observation pairs the BS is defined as:

$$BS = \frac{1}{N}\sum_{n=1}^{N}(p_n - o_n)^2$$

with *p*: forecast probability (fraction of members predicting event)
o: observed outcome (1 if event occurs; 0 if event does not occur)

- BS varies from 0 (perfect deterministic forecasts) to 1 (perfectly wrong!)

- BS corresponds to RMS error for deterministic forecasts

BRIER SKILL SCORE

- In the usual skill score format: proportion of improvement of accuracy over the accuracy of a standard forecast, climatology or persistence.

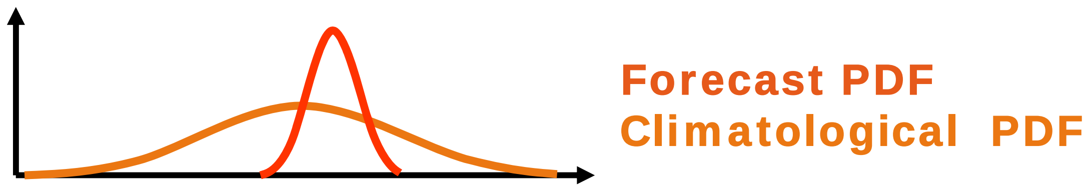
$$BSS = -\frac{\mathrm{BS} - \mathrm{BS}_{\mathrm{ref}}}{\mathrm{BS}_{\mathrm{ref}}}$$

- IF the sample climatology is used, can be expressed as:

$$BSS = -\frac{\mathrm{Res} - \mathrm{Rel}}{\mathrm{Unc}}$$

RELIABILITY

- A forecast system is **reliable** if:
  - statistically the predicted probabilities agree with the observed frequencies, i.e.
  - taking all cases in which the event is predicted to occur with a probability of x%, that event should occur exactly in x% of these cases; not more and not less.
- A **reliability diagram** displays whether a forecast system is reliable (unbiased) or produces over-confident / under-confident probability forecasts
- A reliability diagram also gives information on the resolution (and sharpness) of a forecast system



RELIABILITY DIAGRAM: HOW TO DO IT

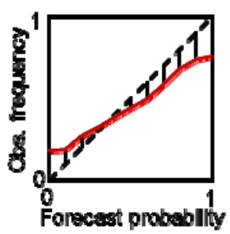
1. Decide number of categories (bins) and their distribution:
   - Depends on sample size, discreteness of forecast probabilities
   - Don't all have to be the same width – within bin sample should be large enough to get a stable estimate of the observed frequency.
2. Bin the data
3. Compute the observed conditional frequency in each category (bin) *k*
   - $obs.\ relative\ frequency_k = obs.\ occurrences_k \ / \ num.\ forecasts_k$
4. Plot observed frequency vs forecast probability
5. Plot sample climatology ("no resolution" line) (The sample base rate)
   - $sample\ climatology = obs.\ occurrences \ / \ num.\ forecasts$
6. Plot "no-skill" line halfway between climatology and perfect reliability (diagonal) lines
7. Plot forecast frequency histogram to show sharpness (or plot number of events next to each point on reliability graph)

RELIABILITY DIAGRAM

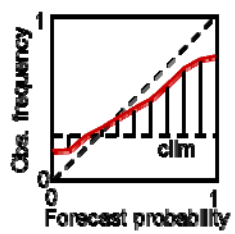
Reliability: Proximity to diagonal

Resolution: Variation about horizontal (climatology) line

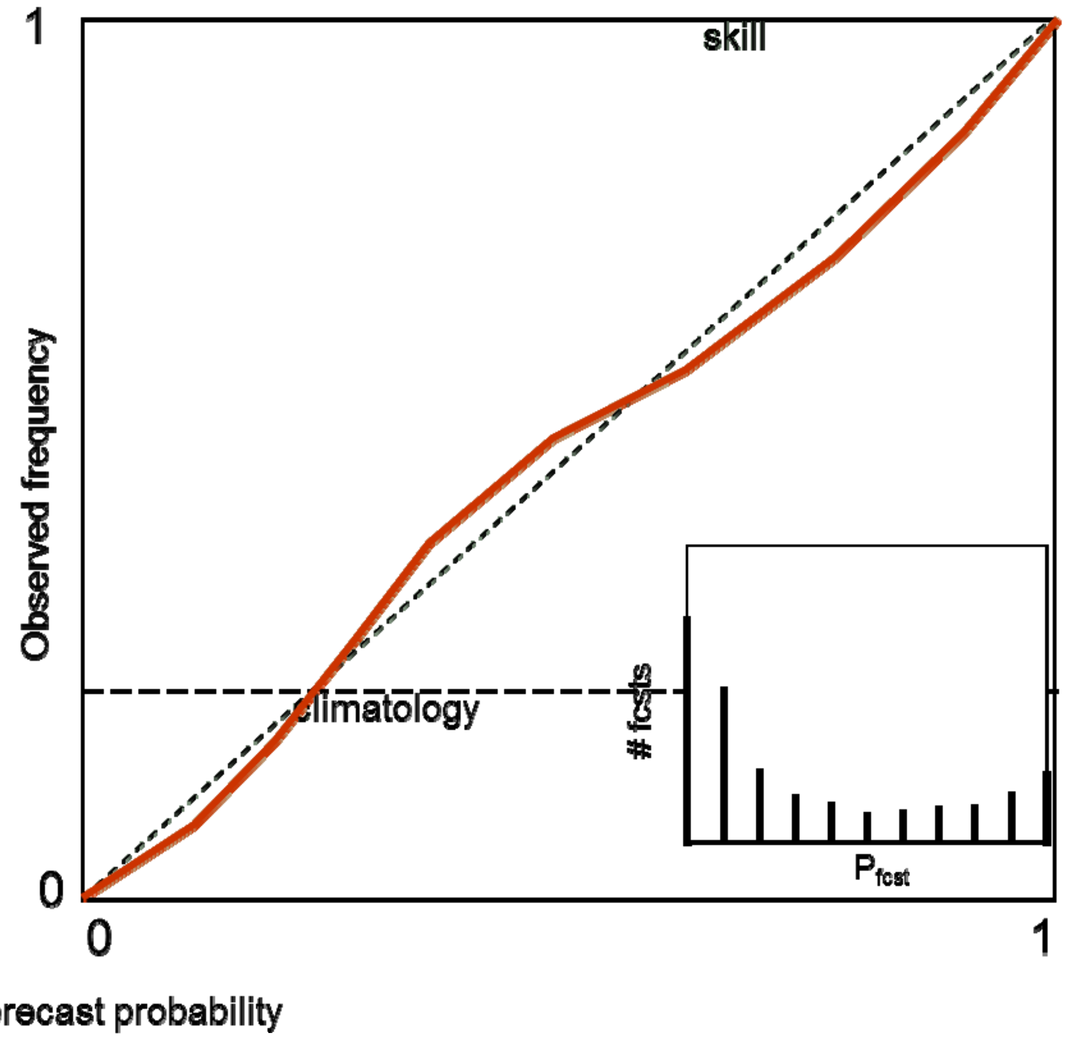
No skill line: Where reliability and resolution are equal – Brier skill score goes to 0



Reliability



Resolution

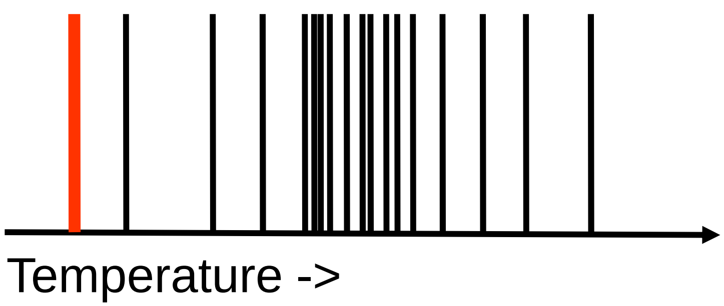


➢ Resolution: ability to issue reliable forecasts close to 0% or 100%

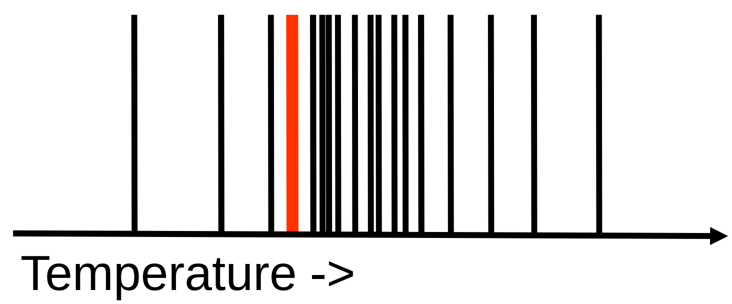
RANK HISTOGRAM

- Rank Histograms asses whether the ensemble spread is consistent with the assumption that the observations are statistically just another member of the forecast distribution
  - Check whether observations are equally distributed amongst predicted ensemble
  - Sort ensemble members in increasing order and determine where the observation lies with respect to the ensemble members

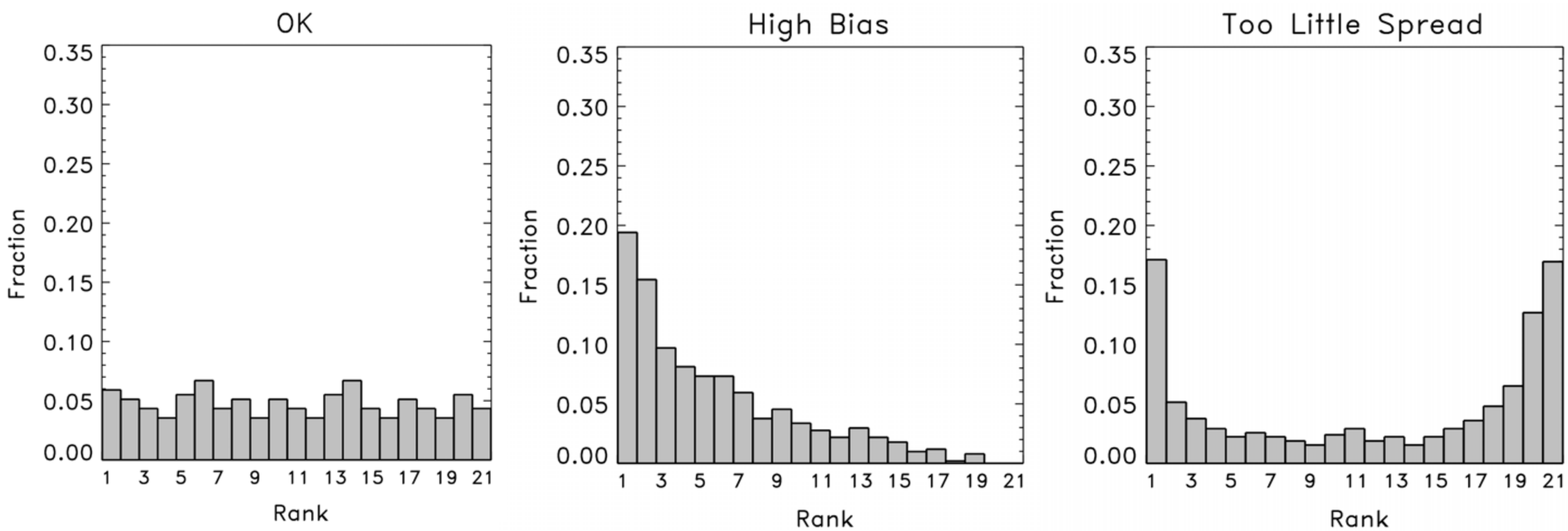
Rank 1 case



Rank 4 case



RANK HISTOGRAMS



OBS is indistinguishable from any other ensemble member

OBS is too often below the ensemble members (biased forecast)

OBS is too often outside the ensemble spread

A uniform rank histogram is a necessary but not sufficient criterion for determining that the ensemble is reliable (see also: T. Hamill, 2001, MWR)