Basic Verification Concepts

Tressa L. Fowler

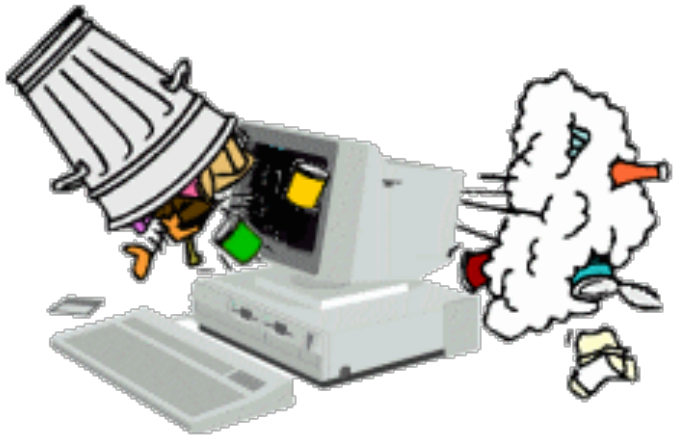
National Center for Atmospheric Research
Boulder Colorado USA

Basic concepts - outline

- What is verification?
- Why verify?
- Identifying verification goals
- Forecast "goodness"
- Designing a verification study
- Types of forecasts and observations
- Matching forecasts and observations
- Verification attributes
- Miscellaneous issues
- **Questions to ponder: Who? What? When? Where? Which? Why?**

How do you do verification?

- Using MET is the easy part, scientifically speaking.
- Good verification depends mostly on what you do before and after MET.
  - What do you want to know?
  - Good forecasts.
  - Good observations.
  - Well matched.
  - Appropriate selection of methods
  - Thorough and correct interpretation of results.



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

Why verify?

- Purposes of verification (traditional definition)

  - Administrative purpose
    - Monitoring performance
    - Choice of model or model configuration (has the model improved?)

  - Scientific purpose
    - Identifying and correcting model flaws
    - Forecast improvement

  - Economic purpose
    - Improved decision making
    - "Feeding" decision models or decision support systems

Why verify?

- What are some other reasons to verify weather forecasts?
  - 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

Identifying verification goals

What *questions* do we want to answer?

- Examples:
  - ✓ In what locations does the model have the best performance?
  - ✓ Are there regimes in which the forecasts are better or worse?
  - ✓ Is the probability forecast well calibrated (i.e., reliable)?
  - ✓ Do the forecasts correctly capture the natural variability of the weather?

Other examples?

Identifying verification goals (cont.)

- What forecast performance *attribute* should be measured?
  - Related to the *question* as well as the type of forecast and observation
- Choices of verification statistics, measures, graphics
  - Should match the type of forecast and the attribute of interest
  - Should measure the quantity of interest (i.e., the quantity represented in the question)

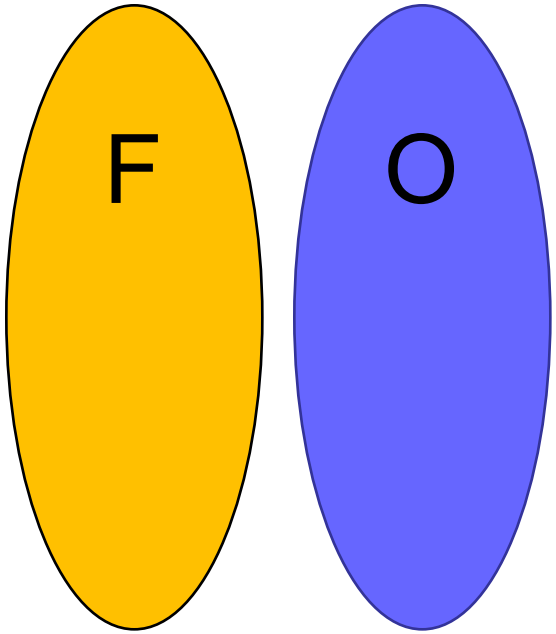
Forecast "goodness"

- Depends on the quality of the forecast

AND

- The user and his/her application of the forecast information

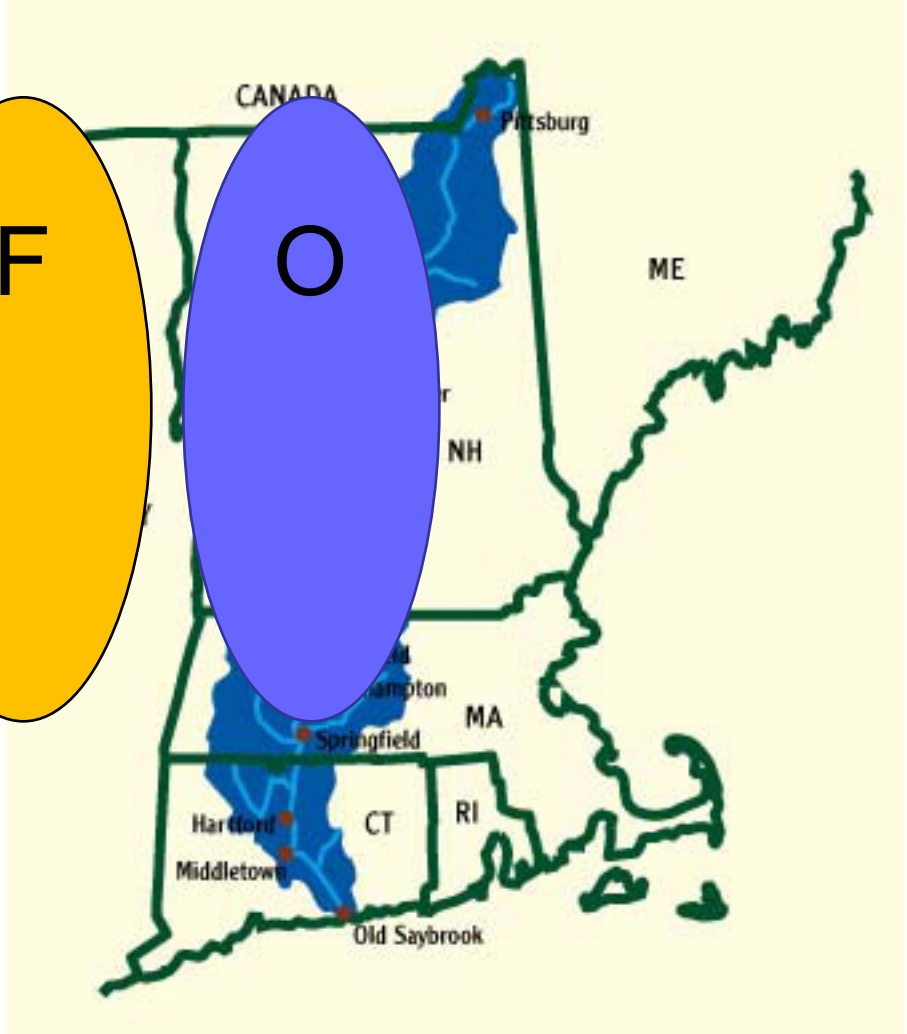
Good forecast or bad forecast?



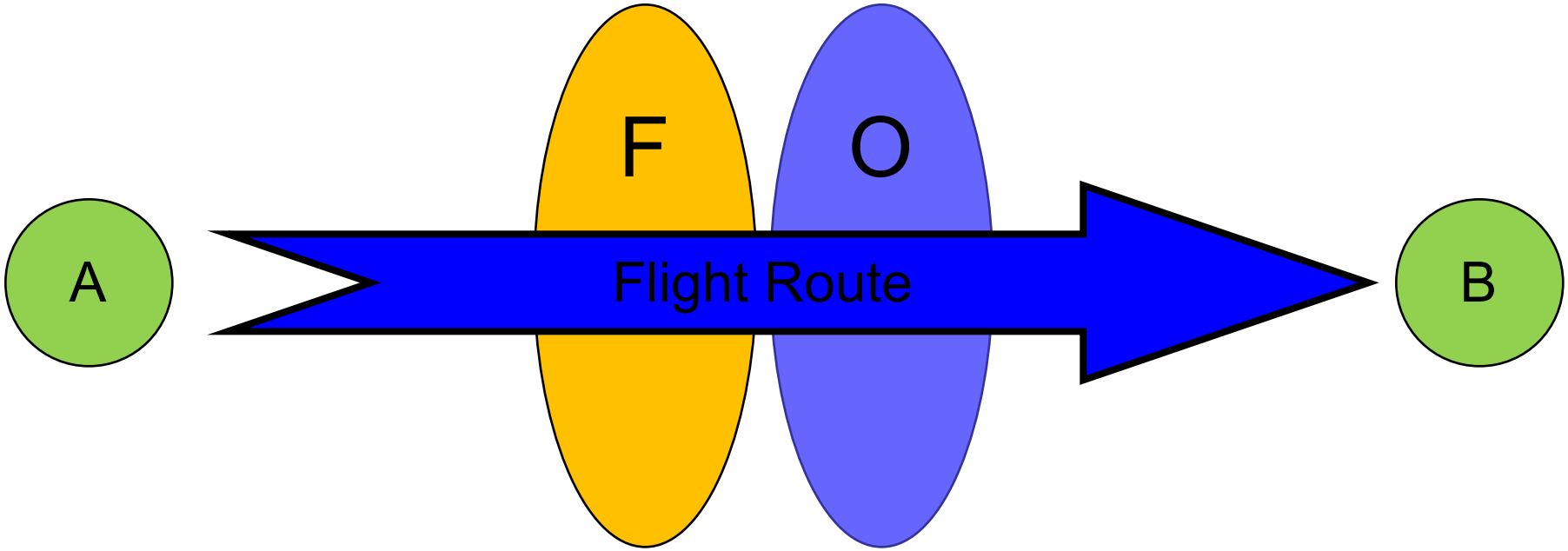
Many verification approaches would say that this forecast has NO skill and is very inaccurate.

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
  - In some cases, *improvements in forecast quality (according to certain measures) may result in a degradation in forecast value for some users!*
- ***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

Basic guide for developing verification studies

Consider the users…

- … of the forecasts
- … of the verification information

- What aspects of forecast quality are of interest for the user?
  - Typically (always?) need to consider multiple aspects

Develop verification questions to evaluate those aspects/attributes

- *Exercise*: What verification questions and attributes would be of interest to …
  - … operators of an electric utility?
  - … a city emergency manager?
  - … a mesoscale model developer?
  - … aviation planners?

Basic guide for developing verification studies

Identify ***observations*** that represent the ***event*** being forecast, including the

- Element (e.g., temperature, precipitation)
- Temporal resolution
- Spatial resolution and representation
- Thresholds, categories, etc.

Observations are not truth

- We can't know the complete "truth".
- Observations generally are more "true" than a model analysis (at least they are relatively more independent)
- Observational uncertainty should be taken into account in whatever way possible
  - ✓In other words, how well do adjacent observations match each other?

Observations might be garbage if

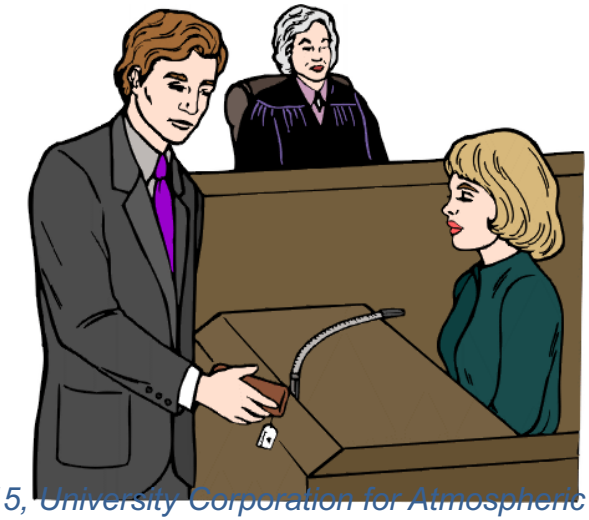
- Not Independent (of forecast or each other)
- Biased
  - Space
  - Time
  - Instrument
  - Sampling
  - Reporting
- Measurement errors
- Not enough of them

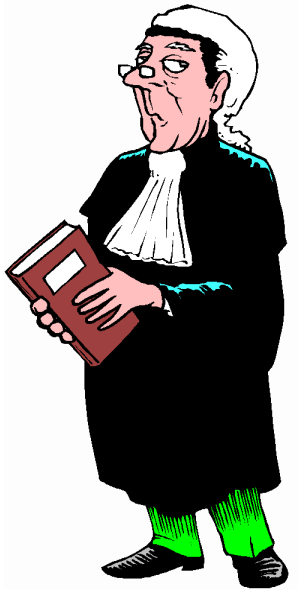
Basic guide for developing verification studies

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

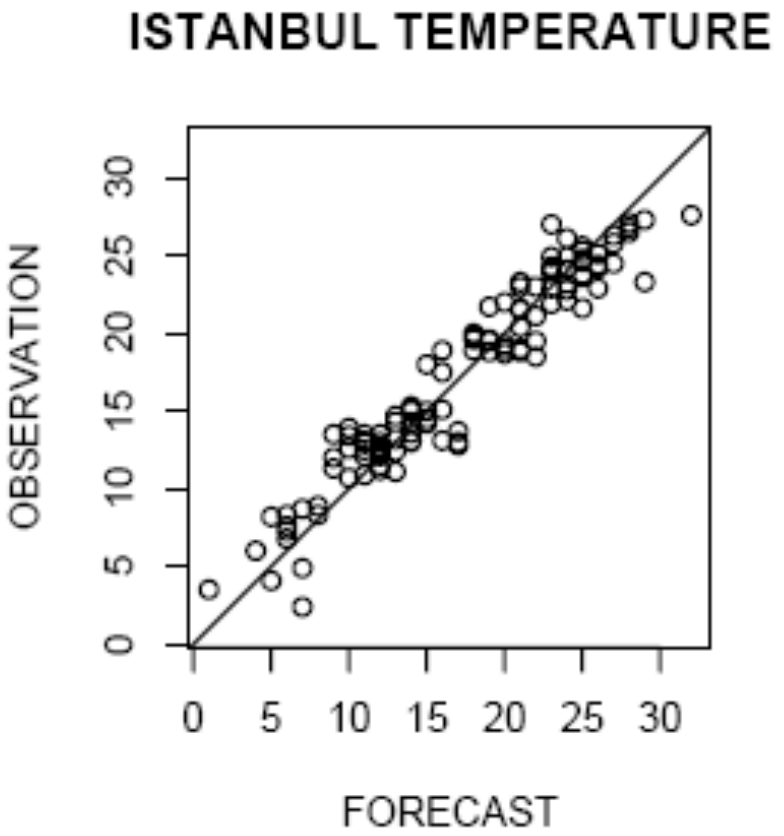
Identify a ***standard of comparison*** that provides a reference level of skill (e.g., persistence, climatology, old model)





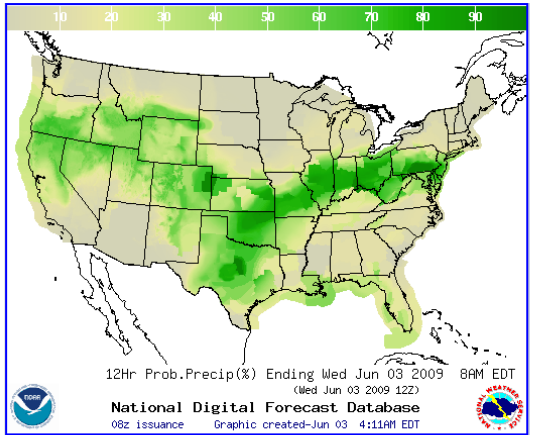
Types of forecasts, observations

- Continuous
  - Temperature
  - Rainfall amount
  - 500 mb height
- Categorical
  - Dichotomous
    - ✓ Rain vs. no rain
    - ✓ Strong winds vs. no strong wind
    - ✓ Night frost vs. no frost
    - ✓ Often formulated as Yes/No
  - Multi-category
    - ✓ Cloud amount category
    - ✓ Precipitation type
  - May result from *subsetting* continuous variables into categories
    - ✓ *Ex: Temperature categories of 0-10, 11-20, 21-30, etc.*



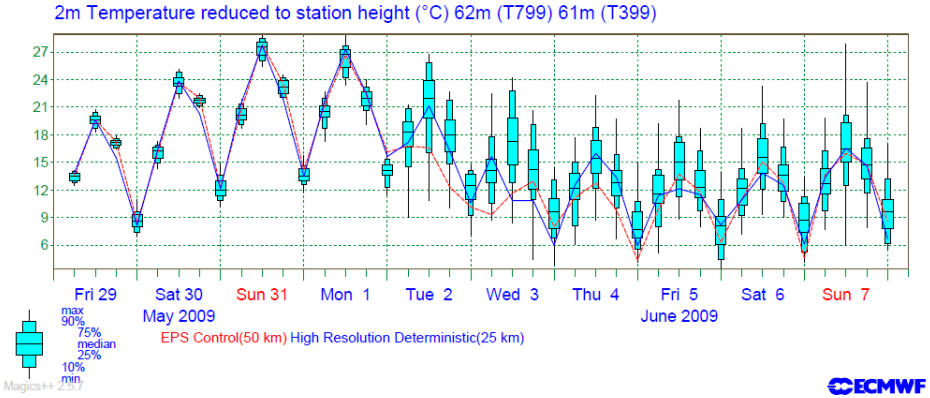
Types of forecasts, observations

- Probabilistic
  - Observation can be dichotomous, multi-category, or continuous
    - Precipitation occurrence – Dichotomous (Yes/No)
    - Precipitation type – Multi-category
    - Temperature distribution - Continuous
  - Forecast can be
    - Single probability value (for dichotomous events)
    - Multiple probabilities (discrete probability distribution for multiple categories)
    - Continuous distribution
  - For dichotomous or multiple categories, probability values may be limited to certain values (e.g., multiples of 0.1)



2-category precipitation forecast (PoP) for US

- Ensemble
  - Multiple iterations of a continuous or categorical forecast
    - May be transformed into a probability distribution
  - Observations may be continuous, dichotomous or multi-category



ECMWF 2-m temperature meteogram for Helsinki

Matching forecasts and observations

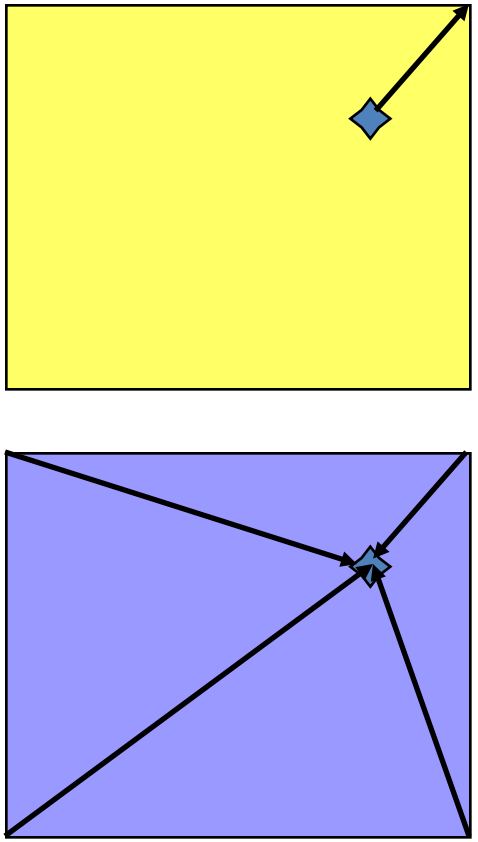
- May be the *most difficult* part of the verification process!
- Many factors need to be taken into account
  - Identifying observations that represent the forecast event
    - ✓Example: Precipitation accumulation over an hour at a point
  - For a gridded forecast there are many options for the matching 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



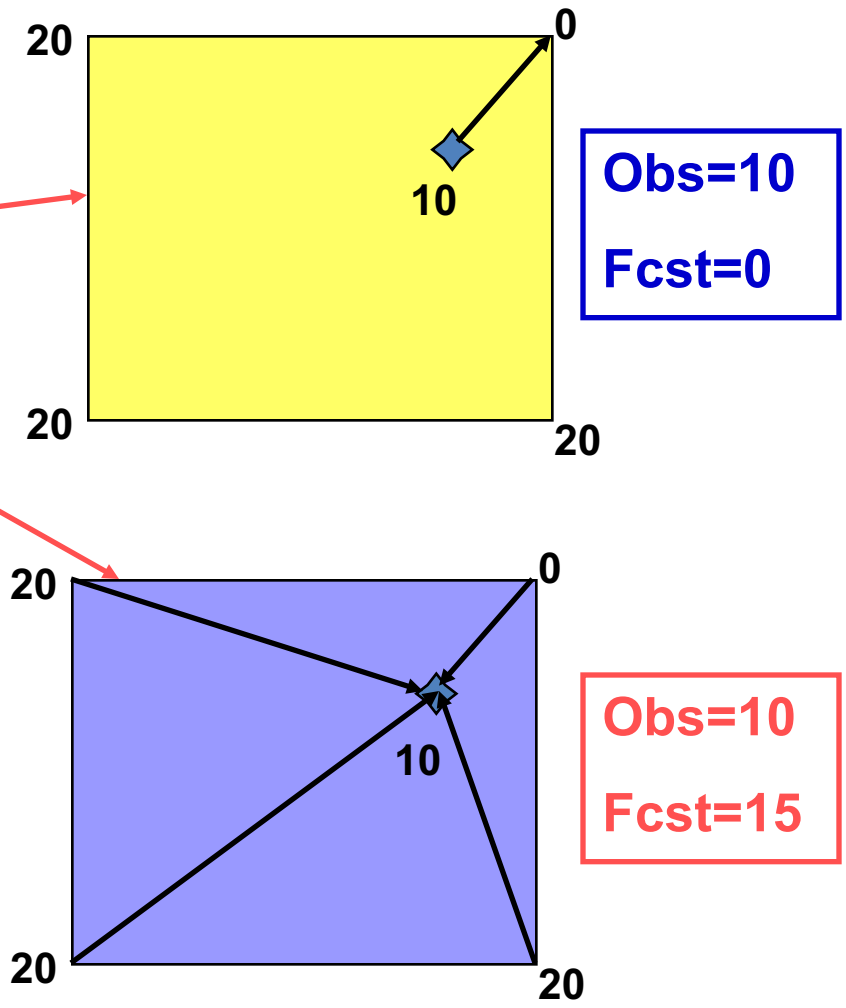
Matching forecasts and observations

Example:

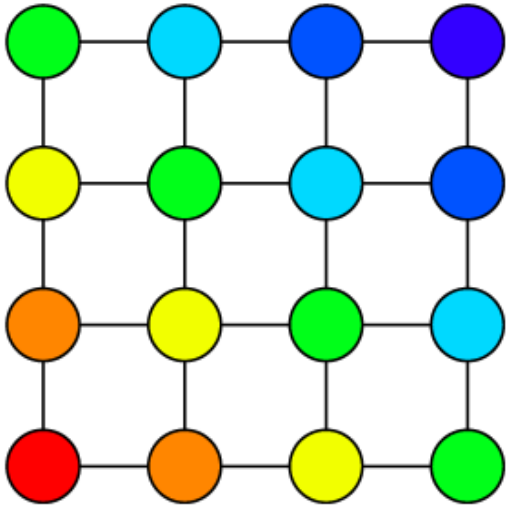
- Two approaches:
  - Match rain gauge to nearest gridpoint ***or***
  - Interpolate grid values to rain gauge location
    - Crude assumption: equal weight to each gridpoint
- Differences in results associated with matching:

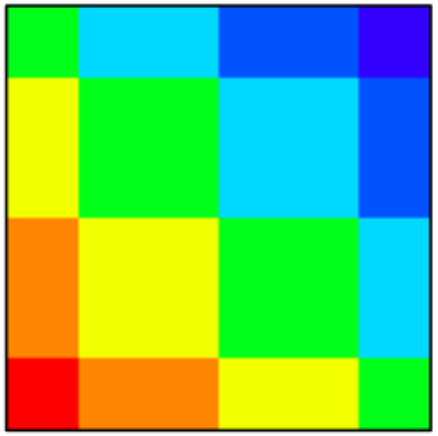
"Representativeness" difference

Will impact most verification scores

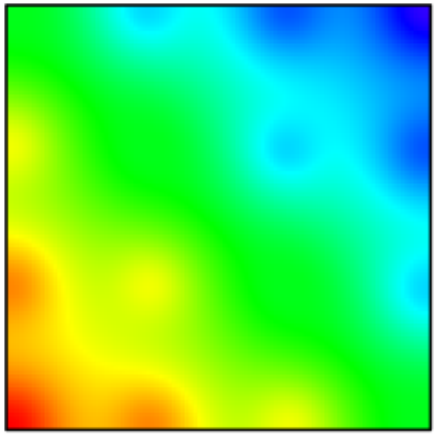


Interpolation Examples

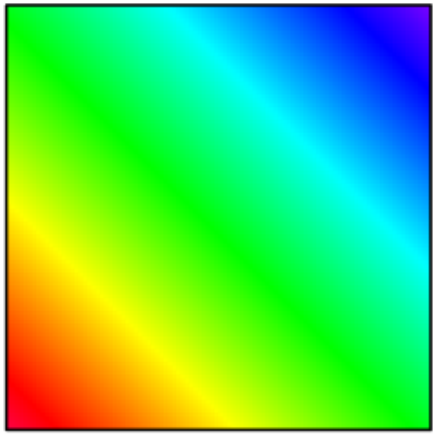




Nearest Neighbor



Distance Weighted Mean



Least Squares

Matching forecasts and observations

Final point:

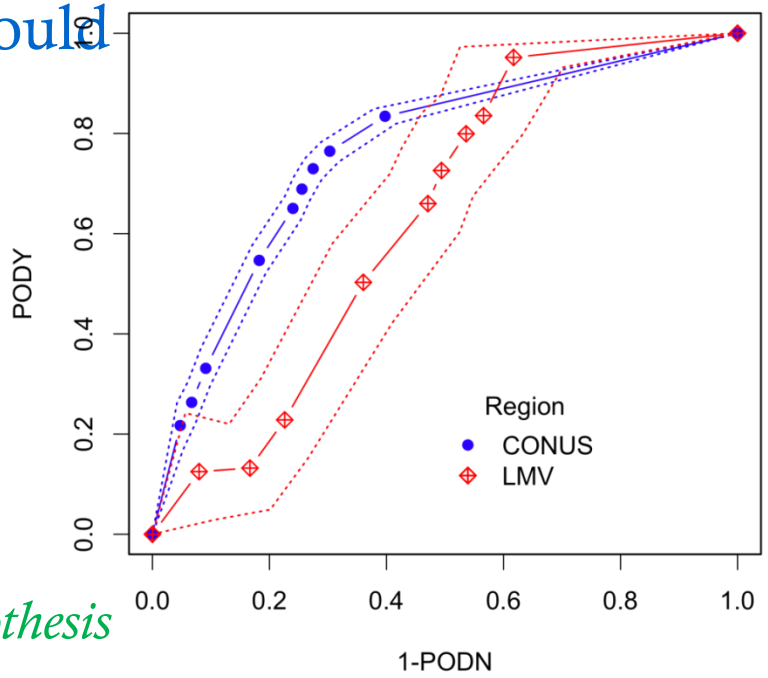
- It is not advisable to use the model analysis as the verification "observation".

- Why not??
- Issue: Non-independence!!

Comparison and inference

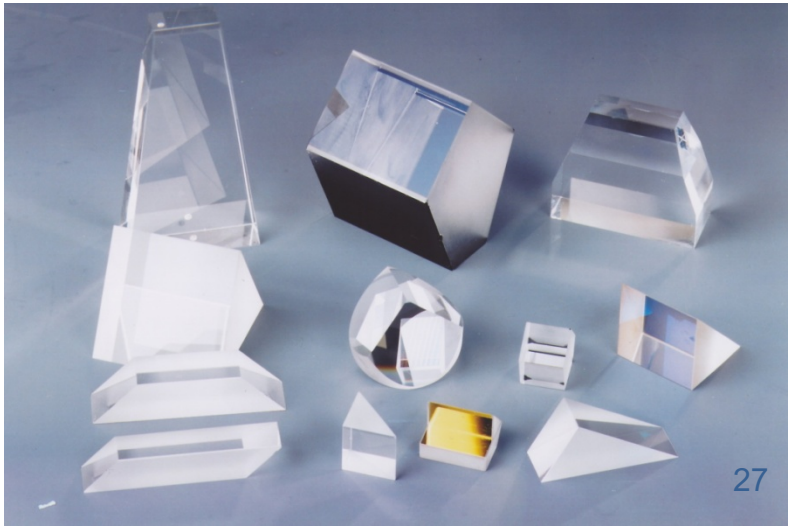
Uncertainty in scores and measures should be estimated whenever possible!

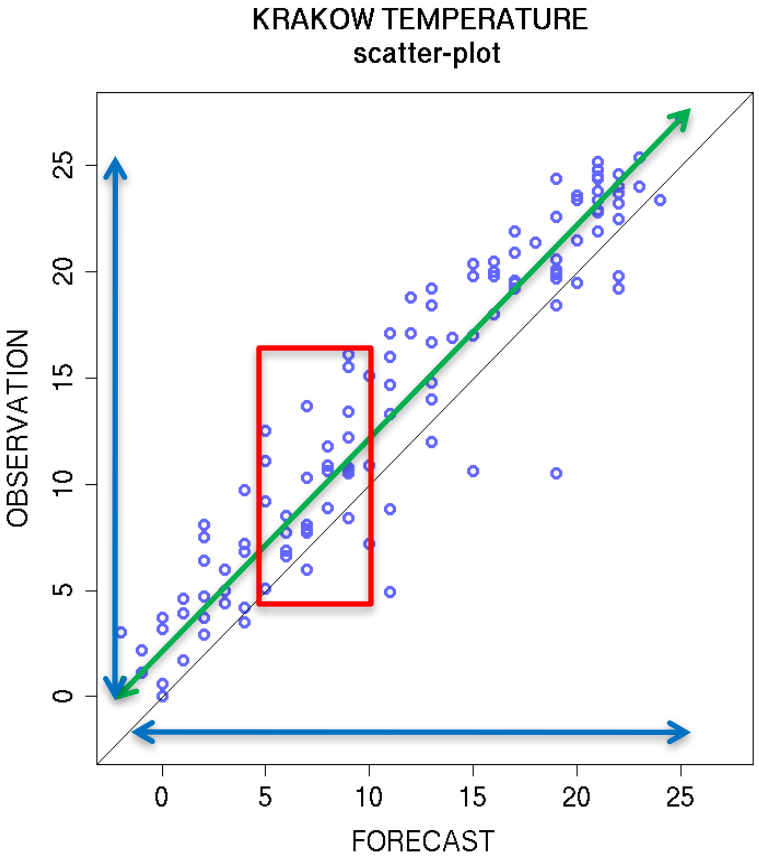
- Uncertainty arises from
  - Sampling variability
  - Observation error
  - Representativeness differences
  - Others?
- Erroneous conclusions can be drawn regarding improvements in forecasting systems and models
- Methods for *confidence intervals* and *hypothesis tests*
  - Parametric (i.e., depending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



Verification attributes

- Verification attributes measure different aspects of forecast quality
  - Represent a range of characteristics that should be considered
  - Many can be related to joint, conditional, and marginal distributions of forecasts and observations





Joint : The probability of two events in conjunction.

Pr (Tornado forecast AND Tornado observed) = 30 / 2800 = 0.01

Conditional : The probability of one variable given that the second is already determined.

Pr (Tornado Observed | Tornado Fcst) = 30/50 = 0.60

Marginal : The probability of one variable without regard to the other.

Tornado forecast	Tornado Observed		
	yes	no	Total fc
yes	30	70	100
no	20	2680	2700
Total obs	50	2750	2800

Pr(Yes Forecast) = 100/2800 = 0.04
Pr(Yes Obs) = 50 / 2800 = 0.02

Verification attribute examples

- Bias
  - (Marginal distributions)
- Correlation
  - Overall association (Joint distribution)
- Accuracy
  - Differences (Joint distribution)
- Calibration
  - Measures conditional bias (Conditional distributions)
- Discrimination
  - Degree to which forecasts discriminate between different observations (Conditional distribution)

Miscellaneous issues

- In order to be *verified*, forecasts must be formulated so that they are *verifiable*!
  - Corollary: All forecasts should be verified – if something is worth forecasting, it is worth verifying
- Stratification and aggregation
  - Aggregation can help increase sample sizes and statistical robustness but can also hide important aspects of performance
    - ✓Most common regime may dominate results, mask variations in performance.
  - Thus it is very important to *stratify results into meaningful, homogeneous sub-groups*

Some key things to think about …

Who…
- …wants to know?

What…
- … does the user care about?
- … kind of parameter are we evaluating? What are its characteristics (e.g., continuous, probabilistic)?
- … thresholds are important (if any)?
- … forecast resolution is relevant (e.g., site-specific, area-average)?
- … are the characteristics of the obs (e.g., quality, 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…

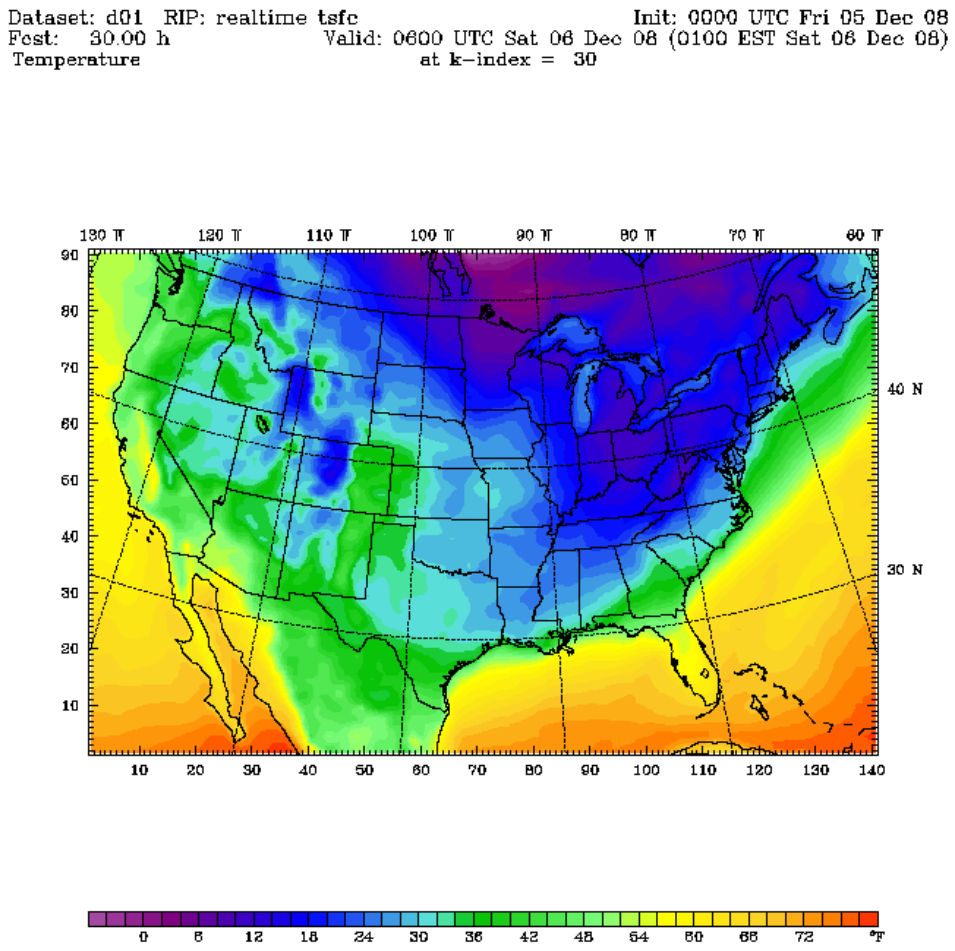
- …methods and metrics are appropriate?
- … methods are required (e.g., bias, event frequency, sample size)

What you can do with MET verification software depends on what type of data you have.

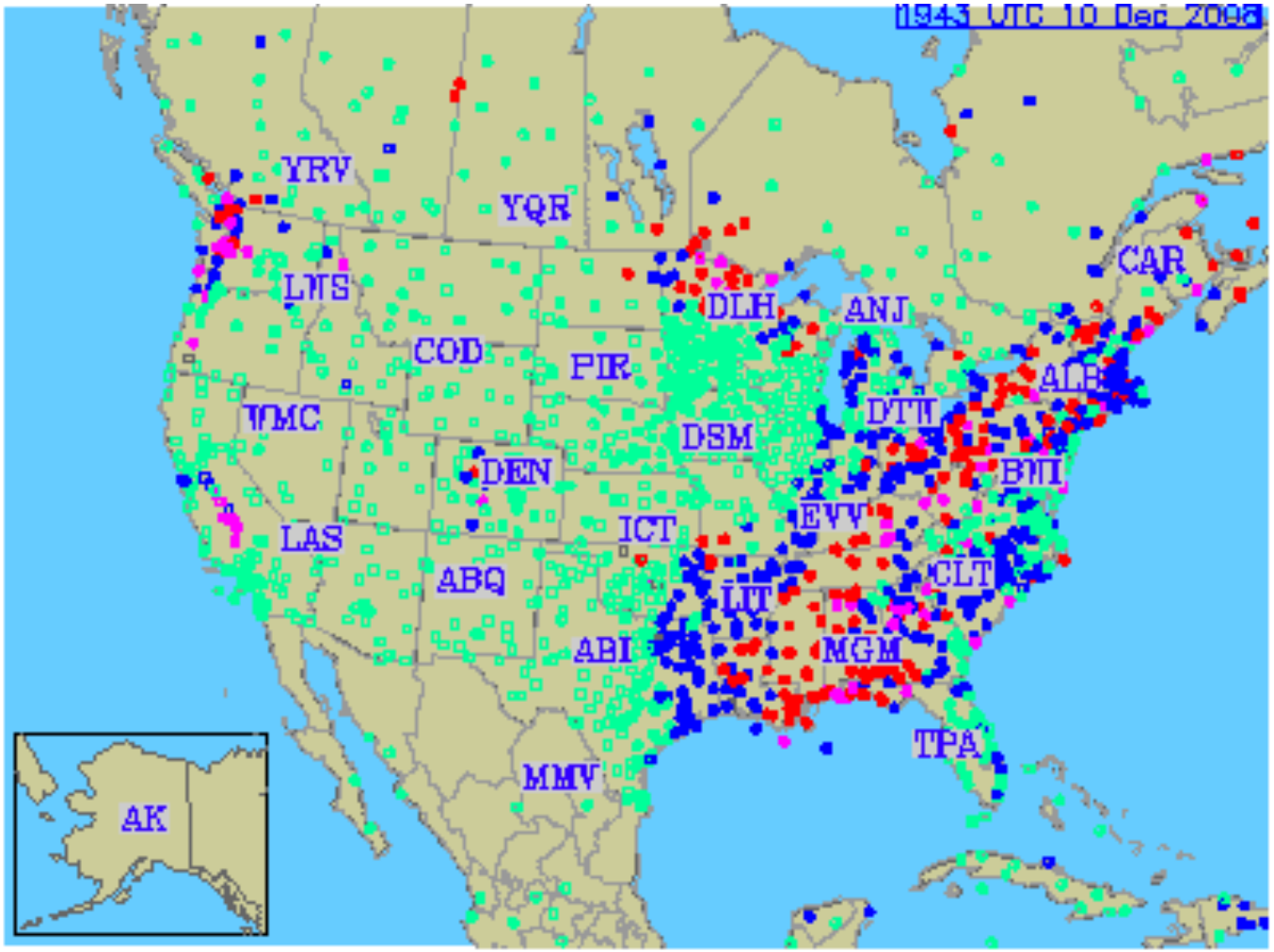
The **format** (grid, point) of your data determines your MET tool(s).

The **type** (continuous, binary) of your data determines the analyses to use within each tool.

Gridded Forecasts
(2D or 3D)

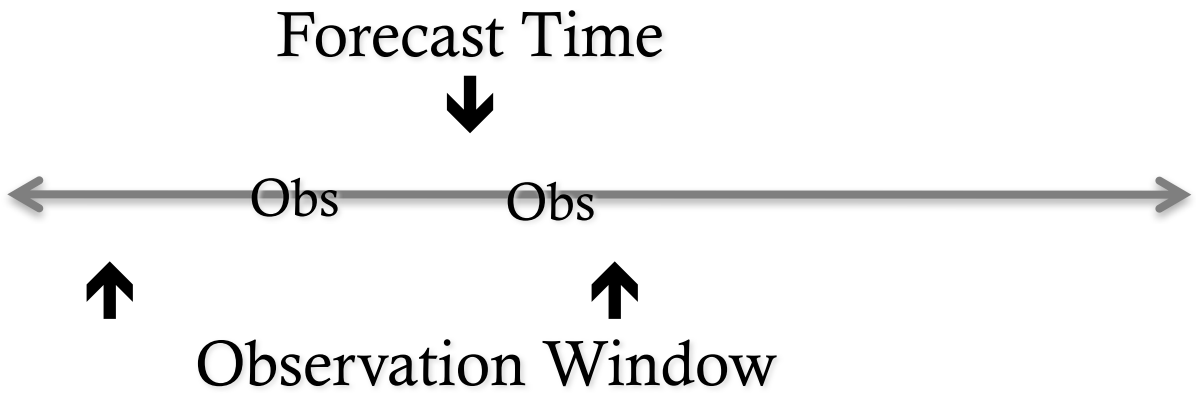


Point Observations
(2D or 3D)

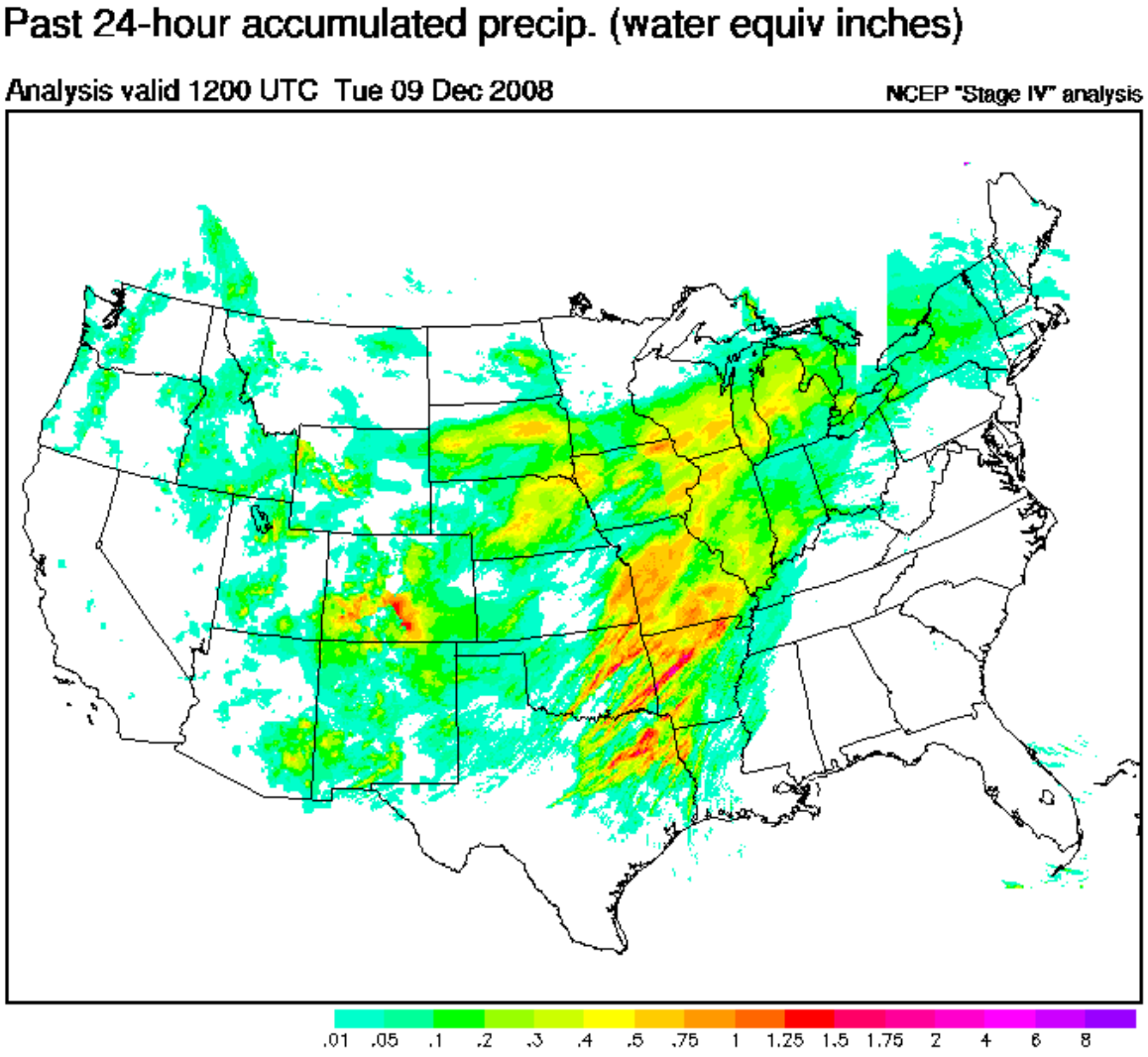


Time

- If your forecasts and observations are not at the same time, you may need to define a time window for your observations.



Gridded Observations (2D or 3D)



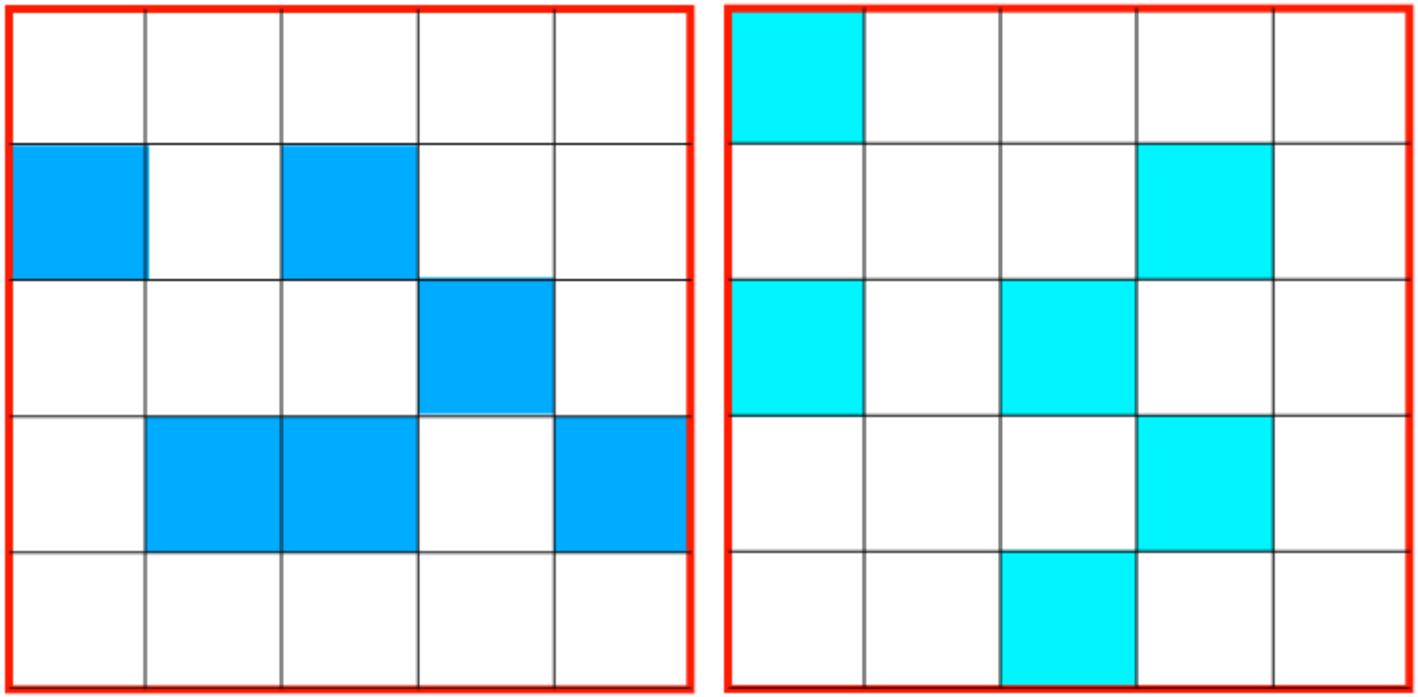
Matching Grids to Grids

- Must use some converter to put forecasts and observations on the same grid.
  - Example: copygb

(High resolution) Gridded Data for use with Neighborhood Methods

observed

forecast



Fraction = 6/25 = 0.24

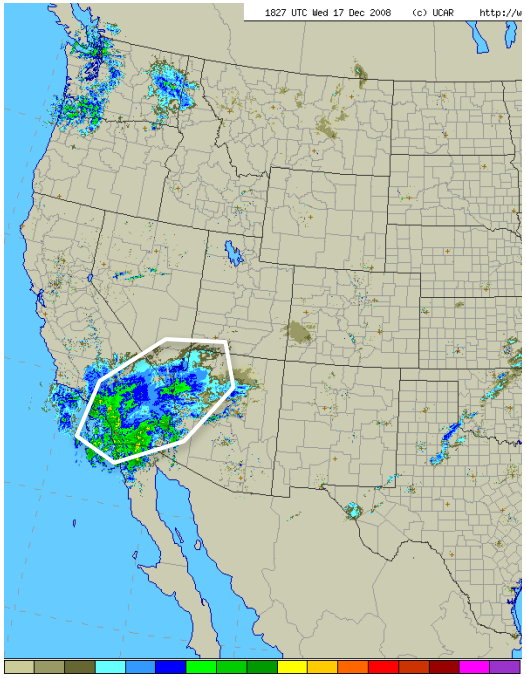
Fraction = 6/25 = 0.24

Intensity threshold exceeded where squares are blue

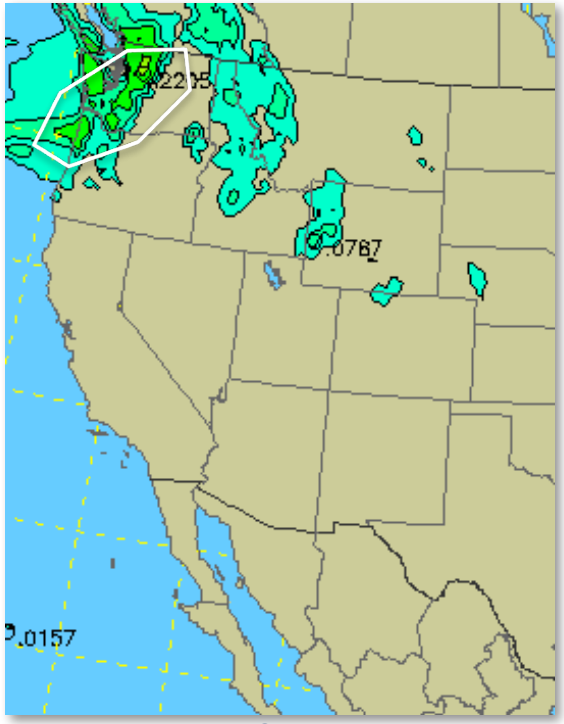
slide from Mittermaier

Gridded data to transform into **Objects**

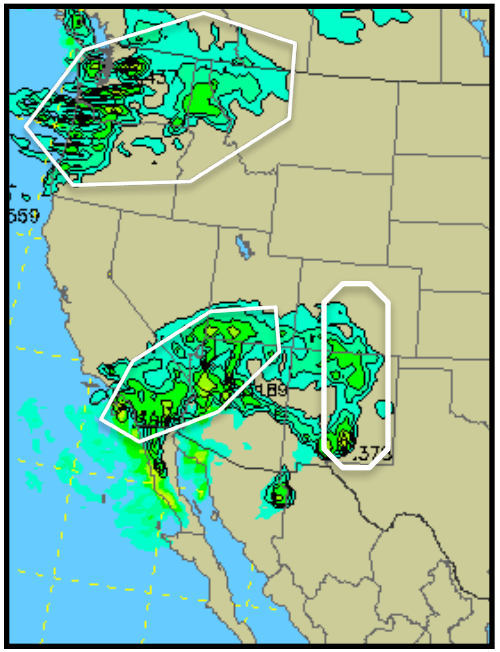
REAL - observed



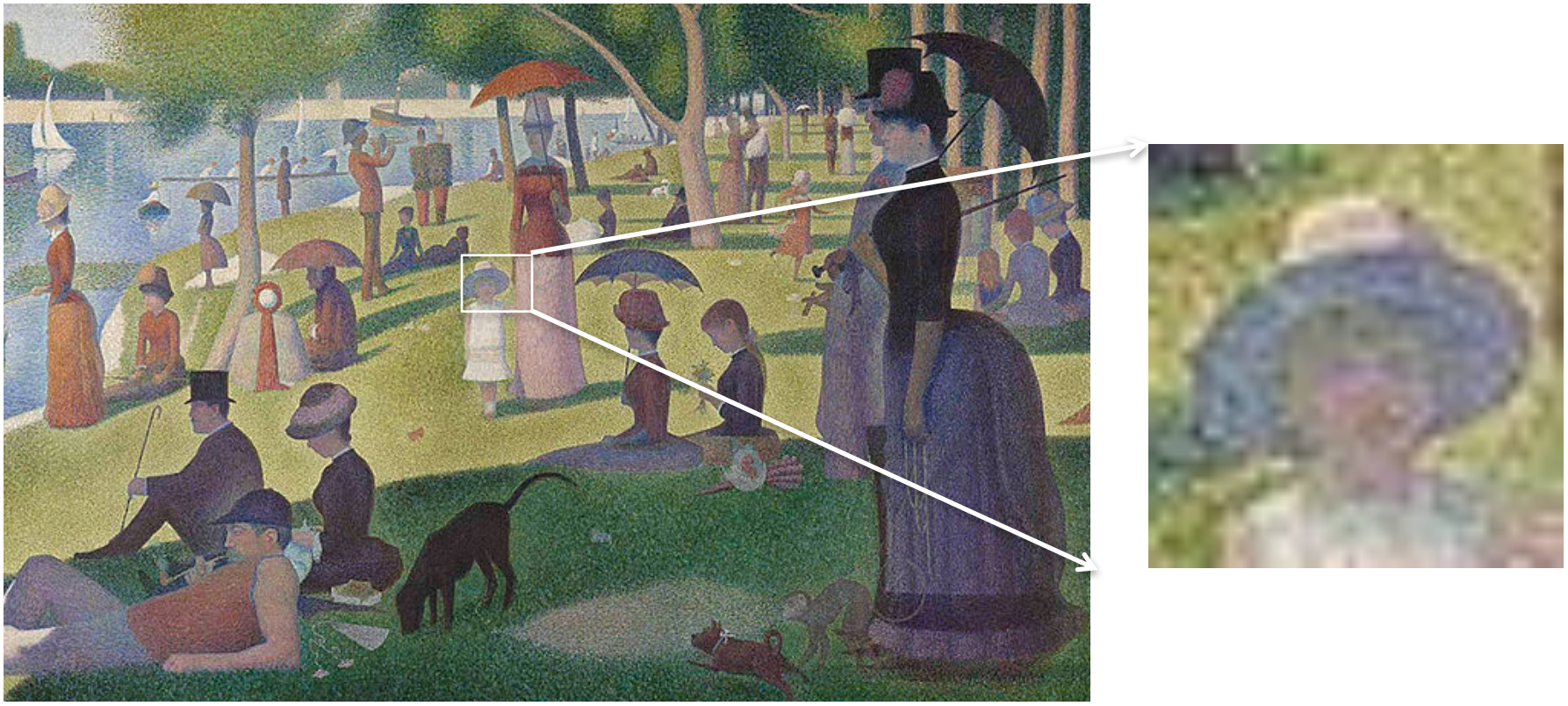
Forecast 1



Forecast 2



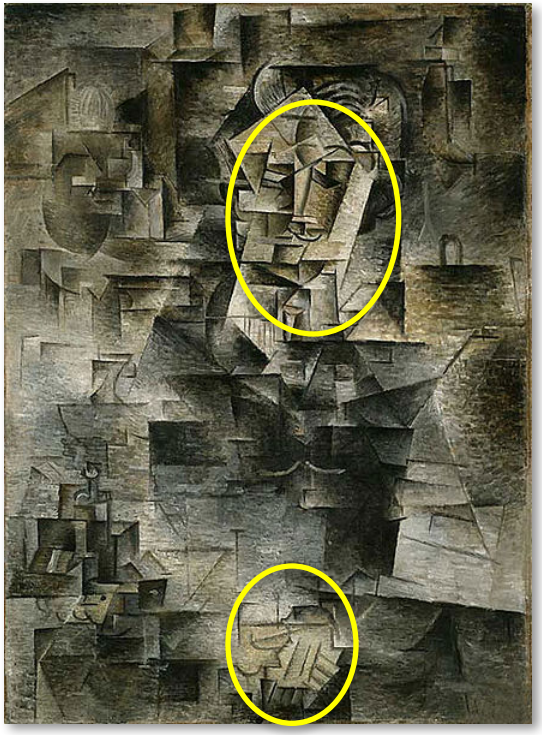
Pixels (traditional Verification)
or
Pictures (Object Verification)?



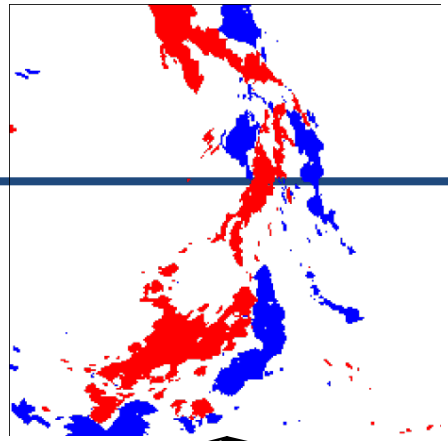
- Humans can pick out which objects exist and go together.
- In object based verification, we use software to mimic this process.

REAL - observed

Forecast 1

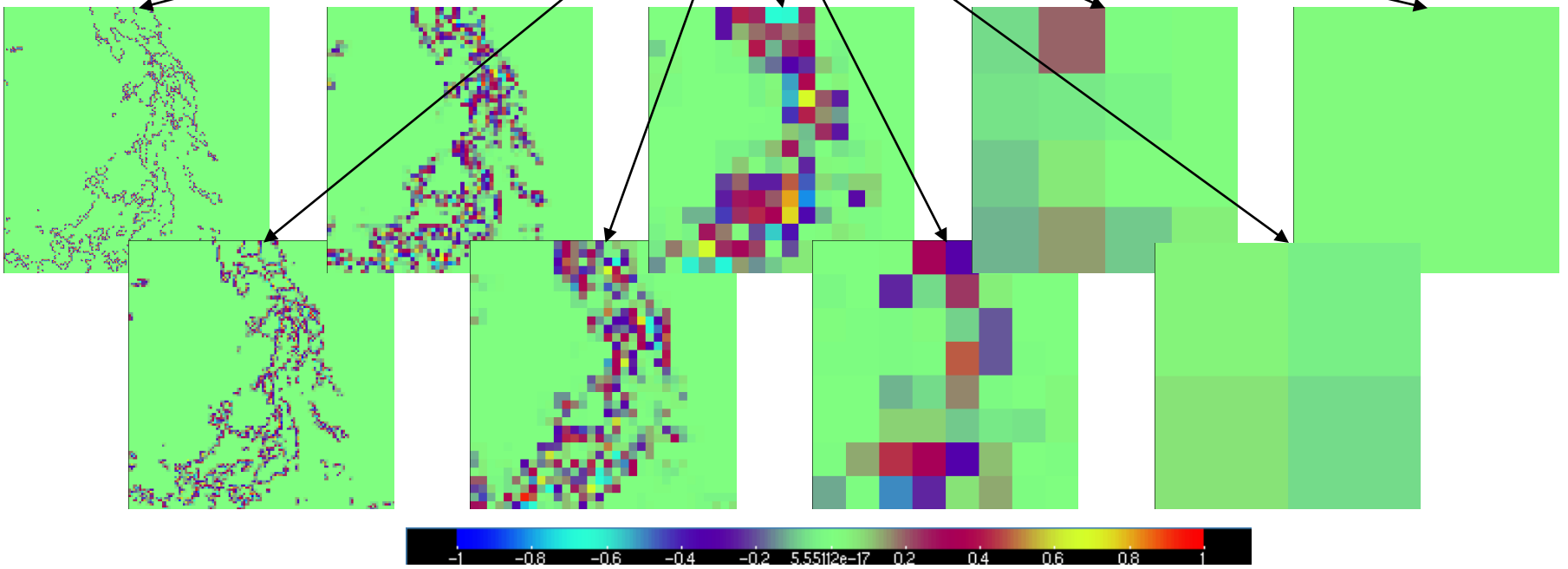


Forecast 2



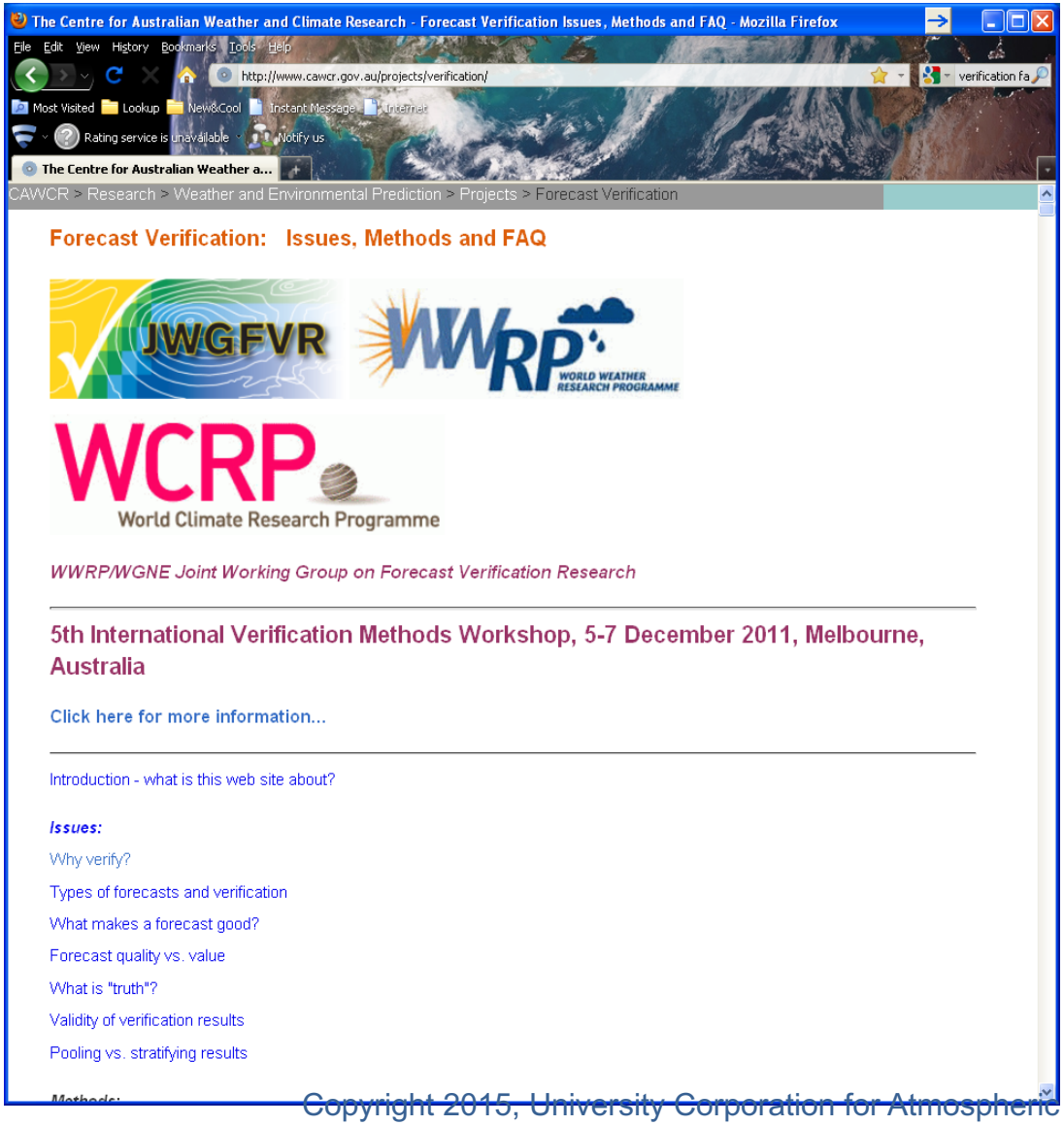
Examine spatial error field at different scales using wavelets

Decompose with Wavelet



Data	MET Tool
Gridded Forecasts Gridded Observations	Grid stat (traditional or neighborhood) Series Analysis Wavelet Stat MODE Ensemble Tool
Gridded Forecasts Point Observations	Point Stat Ensemble Tool
Tropical Cyclone A decks and B decks (both point observations)	MET - TC

Resources



Verification Methods FAQ: http://www.cawcr.gov.au/projects/verification/

Verification Discussion Group: Subscribe at http://mail.rap.ucar.edu/mailman/listinfo/vx-discuss