Basic Concepts in Forecast Evaluation/Verification

Barbara Brown
National Center for Atmospheric Research
Boulder Colorado USA

bgb@ucar.edu

June 2021
Reading University Virtual Summer School

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
- Statistical basis for verification
- Comparison and inference
- Verification attributes
- Miscellaneous issues
- Metaverification: What attributes make a “good” verification measure?
- Questions to ponder

SOME BASIC IDEAS

What is verification?

Verify: **ver·i·fy**

Pronunciation: 'ver-&-"fI

1 : to confirm or substantiate in law by oath

2 : to establish the truth, accuracy, or reality of <*verify* the claim>

synonym see **CONFIRM**

- 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 (broader) term is "***evaluation***"

Why verify?

- Purposes of verification (traditional definition; Brier and Allen 1951)
  - Economic
  - Administrative
  - Scientific

Why verify?

- Economic purpose
  - Understanding forecast value
  - Justification for investment in weather/climate services
  - Improved decision making
  - “Feeding” decision models or decision support systems
- Administrative purpose
  - Monitoring performance
  - Choice of model or model configuration (has the model improved?)
- Scientific purpose
  - Identifying and correcting model flaws
  - Forecast improvement

Why verify?

- What are some other reasons to verify hydrometeorological 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(s)* 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)

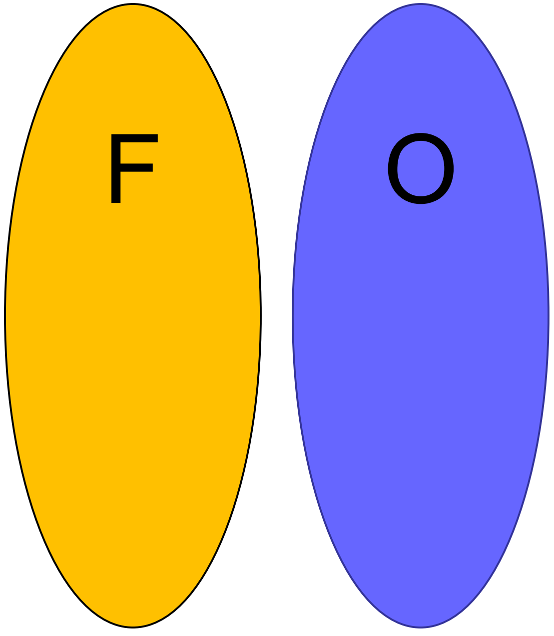
Forecast “goodness”

- Depends on the quality of the forecast

AND

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

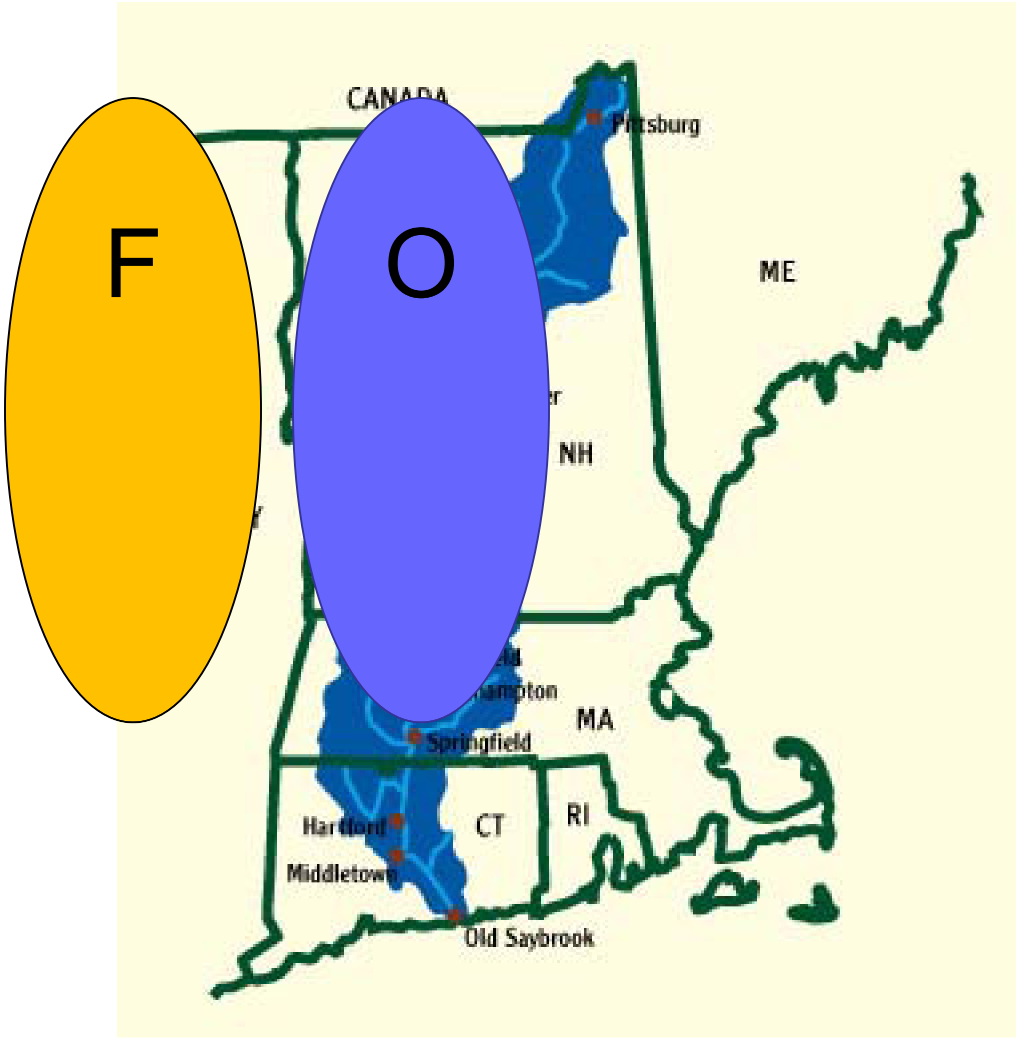
Good forecast or bad forecast?



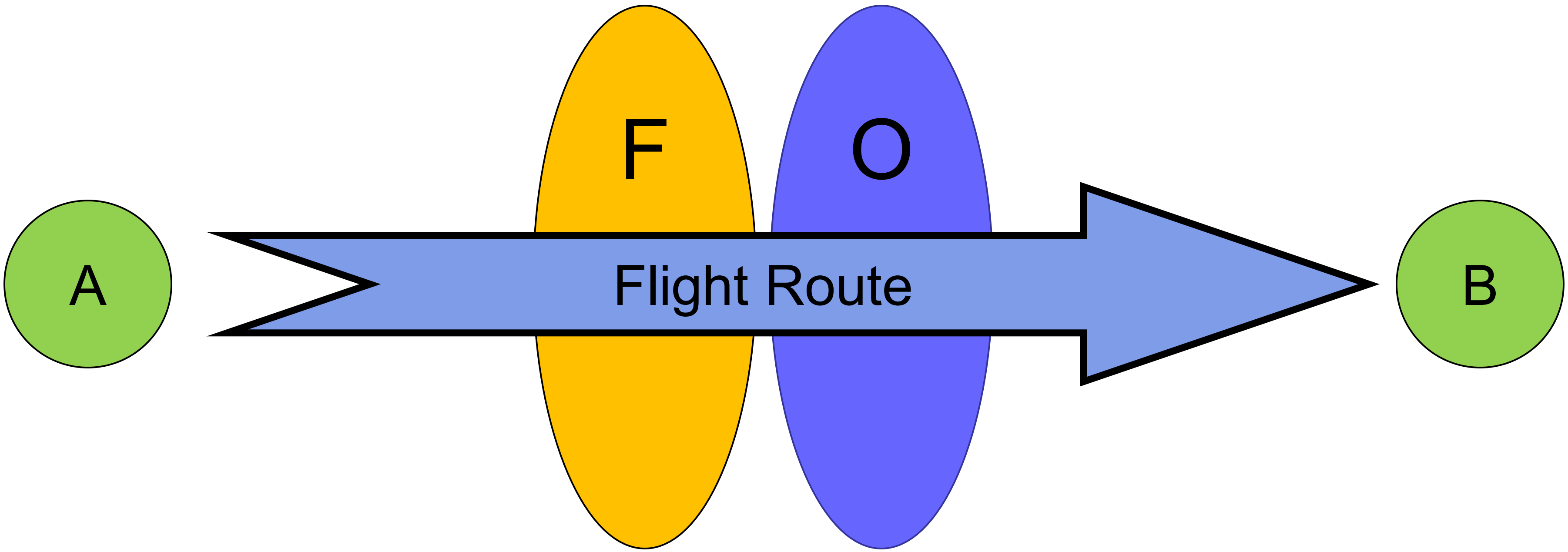
Many verification approaches would say that this forecast has NO skill and is very inaccurate.

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
  - *Diagnostic* verification approaches
  - Features-based or other *spatial 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 we need to consider *multiple aspects*

Develop verification questions to evaluate those aspects/attributes

For example: 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 (cont.)

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

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

Identify multiple ***verification attributes*** that can provide answers to the questions of interest

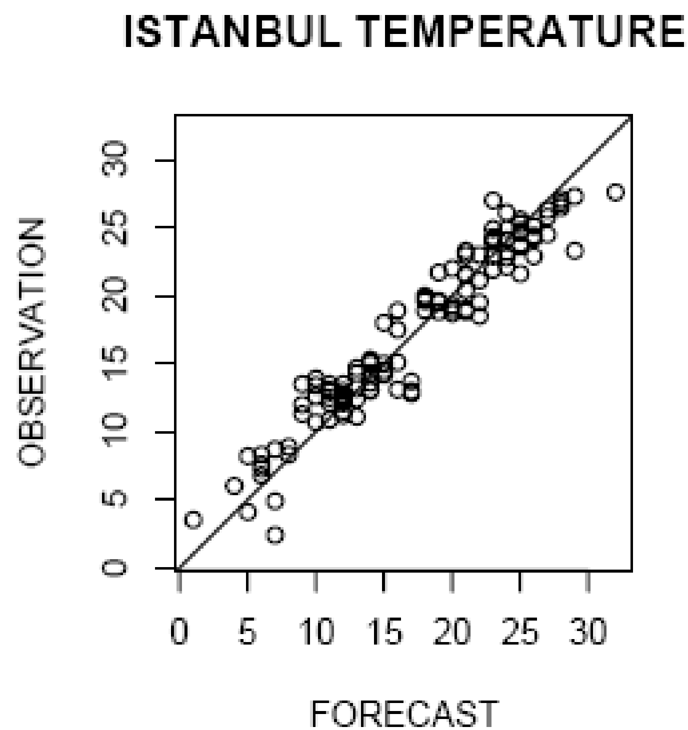
Select ***measures and graphics*** that appropriately measure and represent the attributes of interest

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

FORECASTS AND OBSERVATIONS

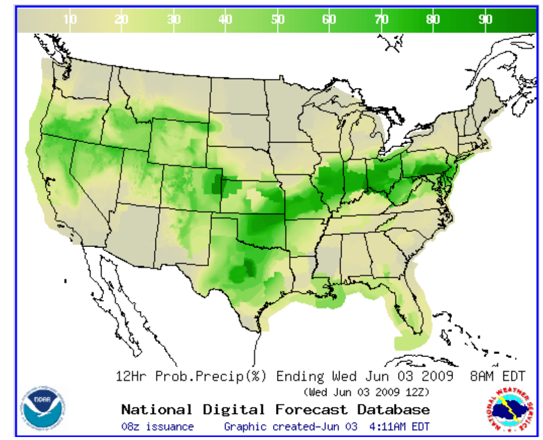
Types of forecasts, observations

- Continuous
  - Temperature
  - Rainfall amount
  - 500 mb height
- Categorical
  - Dichotomous
    - Rain vs. no rain
    - Strong wind vs. no strong wind
    - 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.*



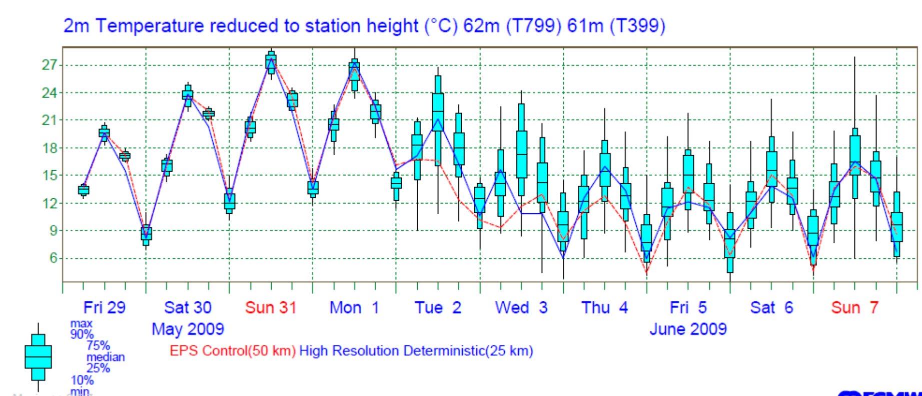
Types of forecasts, observations

- Probabilistic
  - Observation can be
    - Dichotomous (e.g., Yes/No, precipitation occurrence)
    - Multi-category (e.g., precipitation type)
    - Continuous (e.g., temperature distribution)
  - Forecast can be
    - Single probability value (for dichotomous events)
    - Multiple probabilities (discrete probability distribution for multiple categories)
    - Continuous distribution
  - 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

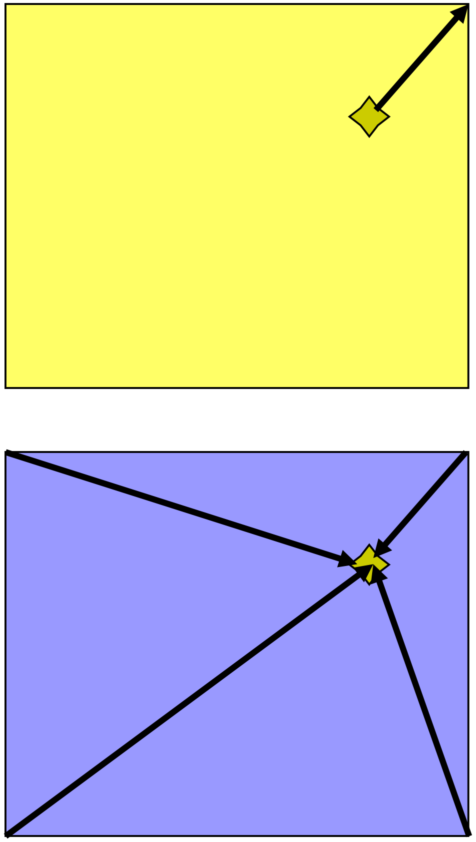
Matching forecasts and observations

- Often the *most difficult* part of the verification process!
- Important factors:
  - Identifying observations that represent the forecast event
    - <u>Example</u>: Precipitation accumulation over an hour at a point
  - For a gridded forecast there are many matching options, including…
    - 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



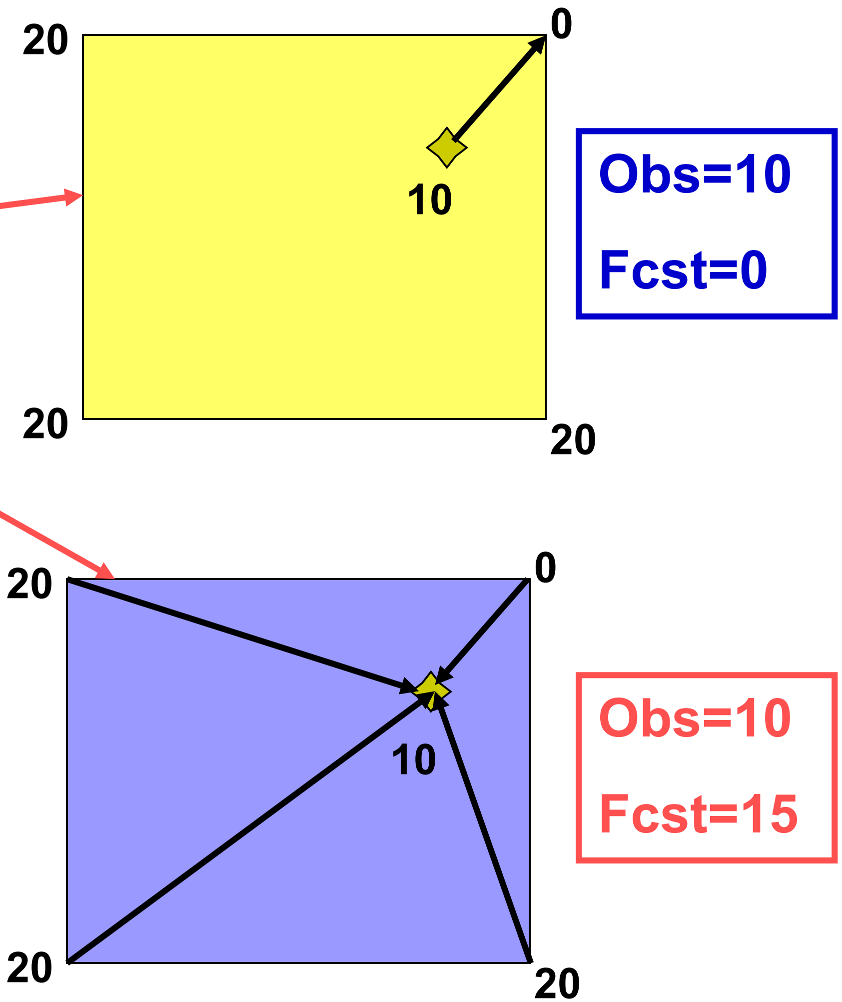
Matching forecasts and observations

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

Final point:

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

- Why not??

  Issue: Non-independence!!

- What would be the impact of non-independence?

  "Better" scores… (not representative)

Observations are *NOT* perfect!

- **Observation error** vs **predictability** and forecast error/uncertainty
- **Different observation types** of the same parameter (manual or automated) can impact results
- Typical **instrument errors** are:
  - *For temperature: +/- 0.1^{o}C*
  - *For wind speed: speed dependent errors but ~ +/- 0.5 m/s*
  - *For precipitation (gauges): +/- 0.1 mm (half tip) but up to 50%*
- Additional **issues: Siting issues (e.g., shielding/exposure)**
- In some instances "forecast" errors are very similar to instrument limits

Impacts of observation errors

- Observations are not perfect due to siting, instrument, and other uncertainties
- Observation errors add uncertainty to the verification results
- Example effects on verification results
  - RMSE – overestimated
  - Spread – more obs outliers make ensemble look under-dispersed
  - Reliability – poorer
  - Resolution – can be better or worse depending on the measure
  - CRPS – poorer mean values
- Some ways of coping
  - Quantify actual observation errors as much as possible
  - Some basic and new methods are available to take into account the effects of observation error
  - More samples can help (reliability of results)

STATISTICAL BASIS FOR VERIFICATION

Statistical basis for verification

Any verification activity should begin with a thorough examination of the statistical properties of the forecasts and observations.

- Many statistical/verification tools are based on assumptions of normality (Gaussian distribution). *Does this assumption hold for the dataset in question?*
- *Does the forecast capture the observed range?*
- *Do the forecast and observed distributions match/agree?*
- *Do they have the same mean behavior, variation etc?*

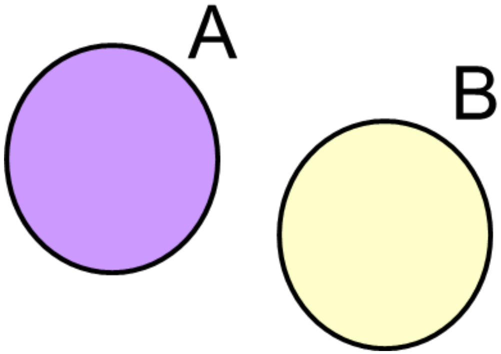
Statistical basis for verification

Beyond the need to assess the characteristics of the data…

Joint, marginal, and conditional distributions are useful for understanding the statistical basis for forecast verification

- These distributions can be related to specific summary and performance measures used in verification
- Specific attributes of interest for verification are measured by these distributions

Statistical basis for verification



Basic (**marginal**) probability

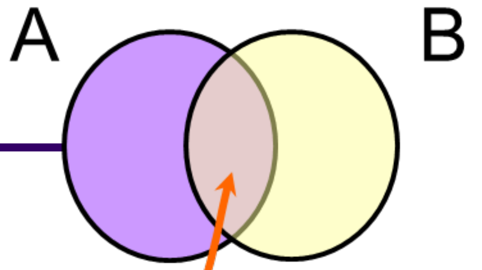
$$p_x = \Pr(X = x)$$

is the probability that a random variable, X, will take on the value x

<u>*Example*</u>*:*

For June in Reading, what is the probability that the observed minimum temperature is between 10 and 15° C?

Basic probability



Joint probability

$$p_{x,y} = \Pr(X = x, Y = y)$$

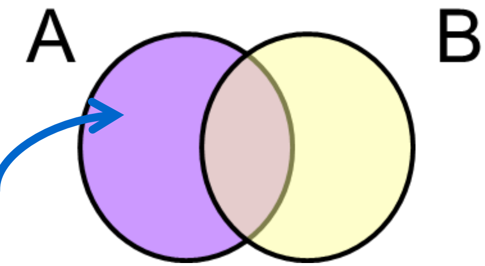
= probability that ***both*** events x and y occur

Example: What is the probability that the daily minimum temperature in June in Reading is between 10 and 15° ($X =$ "*10-15*") *AND* precipitation occurs ($Y =$ "*Yes*")

$= \mathbf{Pr}\ (\boldsymbol{X = 10\text{-}15,\ Y = Yes})$

Basic probability

Conditional probability



$$p_{x,y} = \Pr(X = x \mid Y = y)$$

= probability that event x is true (or occurs) given that event y is true (or occurs)

Example: If it is raining, what is the likelihood that that the minimum temperature is between 10 and 15°C:

$$= \Pr\,(X = \mathit{10\text{-}15} \mid Y = \mathit{Yes})$$

What does this have to do with verification?

Verification can be represented as the process of evaluating the joint distribution of forecasts and observations, $p(f,x)$

- All of the information regarding the forecasts, observations, and their relationship is represented by this distribution
- Furthermore, the joint distribution can be factored into two pairs of conditional and marginal distributions:

$$p(f,x) = p(F=f \mid X=x)\,p(X=x)$$

$$p(f,x) = p(X=x \mid F=f)\,p(F=f)$$

Decompositions of the joint distribution

- Many forecast verification attributes can be derived from the conditional and marginal distributions
- Likelihood-base rate decomposition

$$p(f,x) = \underbrace{p(F=f \mid X=x)}_{\text{Likelihood}}\,\underbrace{p(X=x)}_{\text{Base rate}}$$

- Calibration-refinement decomposition

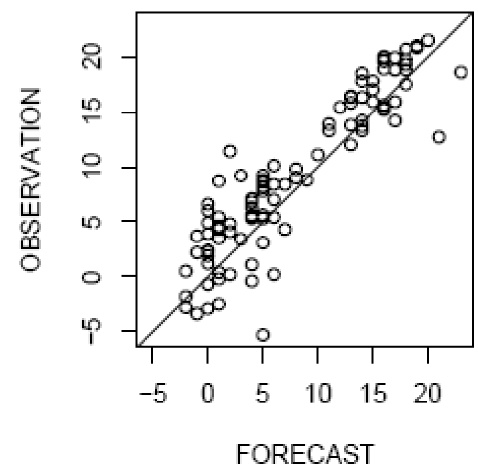
$$p(f,x) = \underbrace{p(X=x \mid F=f)}_{\text{Calibration}}\,\underbrace{p(F=f)}_{\text{Refinement}}$$

Graphical representation of distributions

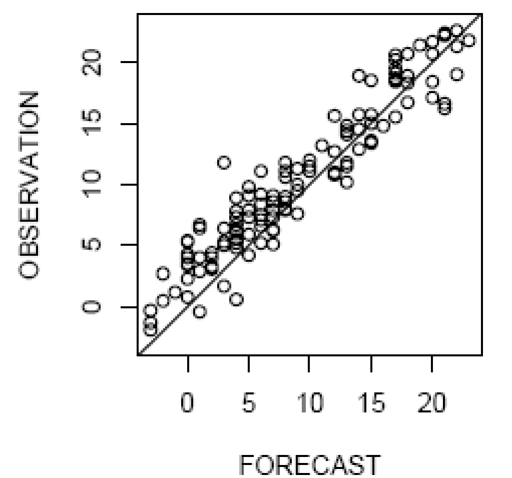
Joint distributions

- Scatter plots
- Density plots
- 3-D histograms
- Contour plots

OSLO TEMPERATURE



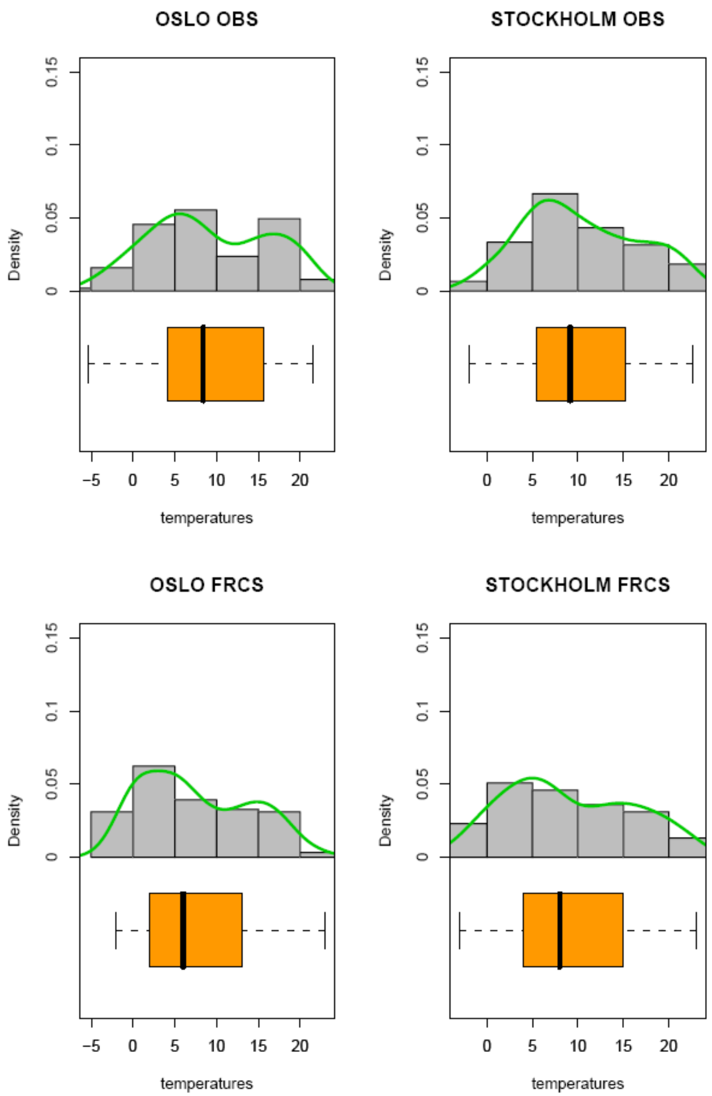
STOCKHOLM TEMPERATURE



Graphical representation of distributions

Marginal distributions

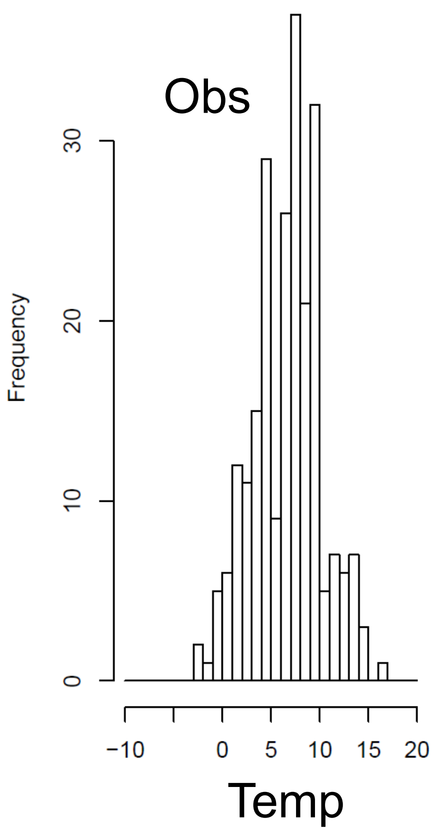
- Stem and leaf plots
- Histograms
- Box plots
- Cumulative distributions
- Quantile-Quantile plots

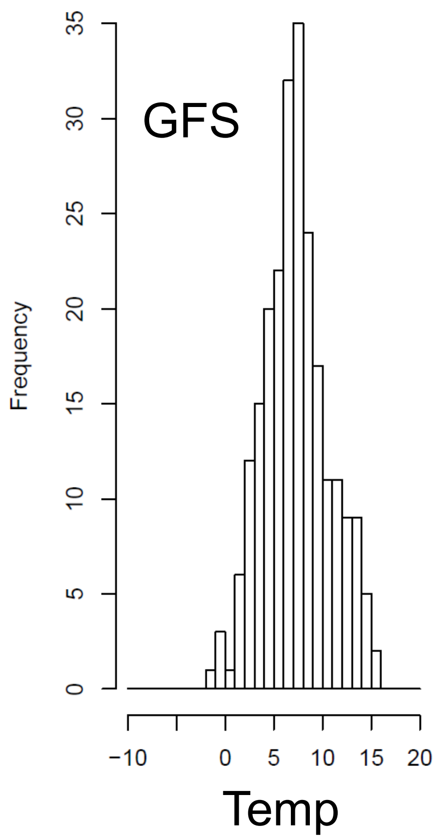


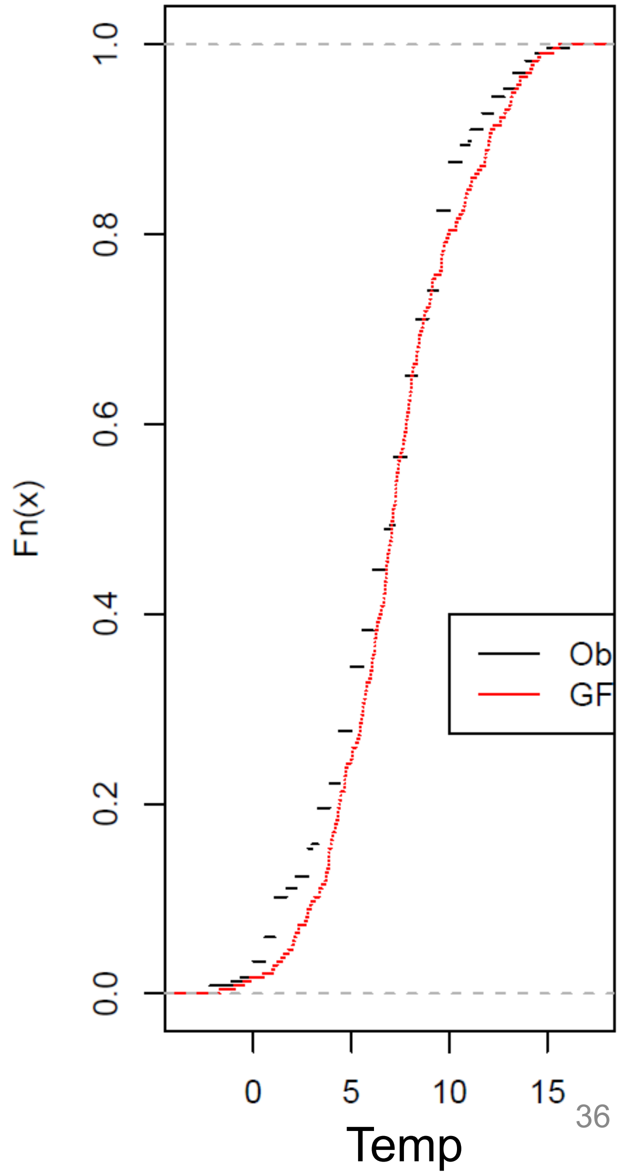
Graphical representation of distributions

Marginal distributions

- Density functions
- Cumulative distributions



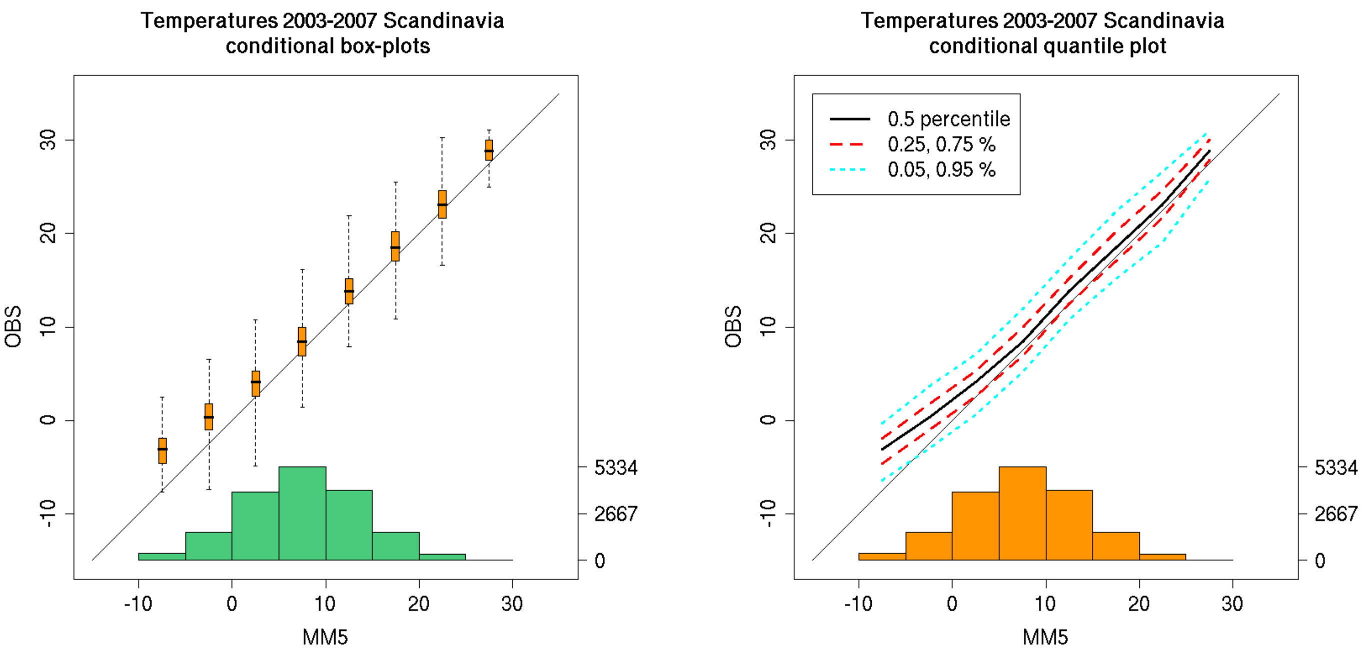




Graphical representation of distributions

Conditional distributions

- Conditional quantile plots
- Conditional boxplots
- Stem and leaf plots



COMPARISON AND INFERENCE

Comparison and inference

Skill scores

- Measure *relative performance*
  - ***Ex**: 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***
- Measures percent improvement over the standard
- Positively oriented (larger is better)
- Choice of the standard matters (*a lot*!)

Question: Which standard of comparison would be more difficult to "beat": climatology or persistence

For

- A 72-hour precipitation forecast?
- A 6-hour ceiling forecast?

Skill scores

Generic skill score definition:

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

Where $\boldsymbol{M}$ is the verification measure for the forecasts, $\boldsymbol{M_{ref}}$ is the measure for the reference forecasts, and $\boldsymbol{M_{perf}}$ is the measure for perfect forecasts

Example: for Mean-squared error (MSE)

$$Skill_{MSE} = \frac{MSE_{fcst} - MSE_{ref}}{0 - MSE_{ref}} = \frac{MSE_{ref} - MSE_{fcst}}{MSE_{ref}}$$

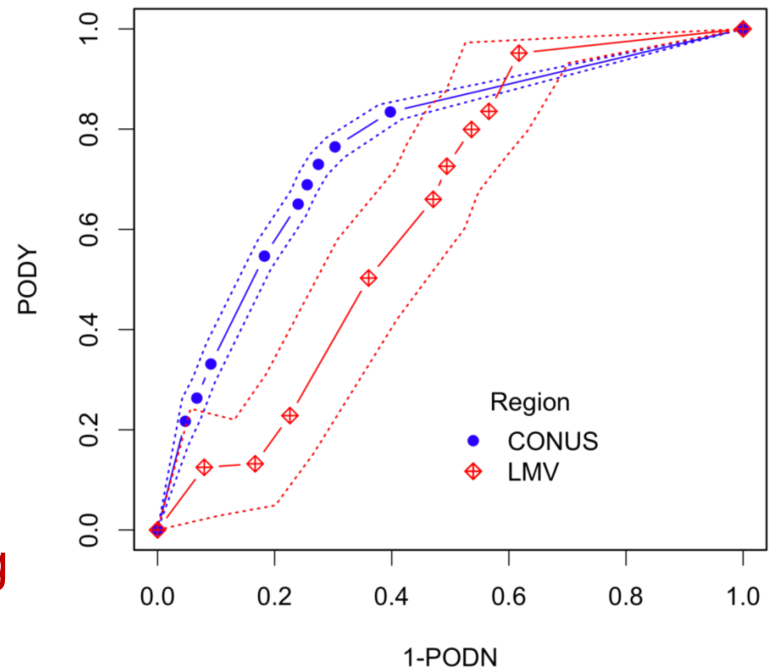
Types of references

Type	Example	Properties
Random	Equitable Threat Score	• Well understood statistical benchmark • Not physically meaningful
Persistence	Constructed skill score	• Measure of predictability (predictability is low when persistence is a poor forecast) • Show value added by running NWP model
Sample climate	Constructed skill score	• One step further removed than persistence, i.e. smoothed • Retains predictability element due to regime dependence
Long-term climatology	Constructed skill score, extremes	• Easiest reference to beat, smoothest • Care required with respect to representativeness, pooling issues, climate change trends

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 can be either *parametric* or *non-parametric*
  - Parametric depend on a statistical model
  - Non-parametric are typically derived from re-sampling procedures (e.g.,“bootstrapping”)



More on this topic to be presented tomorrow by Ian Jolliffe

VERIFICATION ATTRIBUTES

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

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)

Desirable characteristics of verification measures

- Statistical validity
- Properness (probability forecasts)
  - "Best" score is achieved when forecast is consistent with forecaster's best judgments
  - "Hedging" is penalized
  - Example: Brier score
- Equitability
  - Constant and random forecasts should receive the same score
  - <u>*Example*</u>: Gilbert skill score (2x2 case); Gerrity score
  - No scores achieve this in a more rigorous sense
    - Ex: Most scores are sensitive to bias, event frequency

Metaverification[1]: Desirable properties of verification measures

- Statistical validity
- Consistency (Murphy, 1993; *WAF*)
  - Appropriate representation of forecast uncertainty (e.g., via use of probabilities)
- Propriety
  - Forecaster rewarded for forecast probabilities that correspond to his/her "true" beliefs (i.e., not *hedging*)
- Equitability
  - Unskilled forecasts (e.g., constant, random) receive the same expected score
  - No scores achieve this in a more rigorous sense

    Ex: Most scores are sensitive to bias, event frequency

[1]First used by Murphy (1996; WAF)

SUMMARY

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*

Verification issues cont.

- Observations
  - No such thing as “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
    - e.g., how well do adjacent observations match each other?

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)

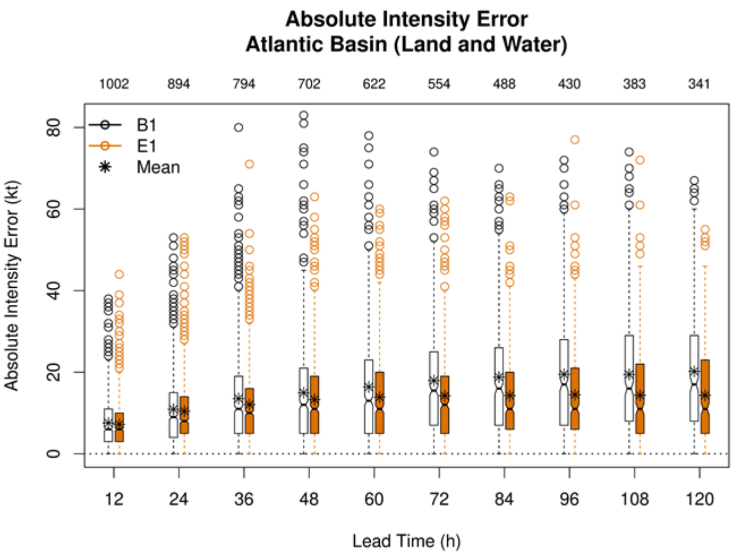
A few relevant references and links

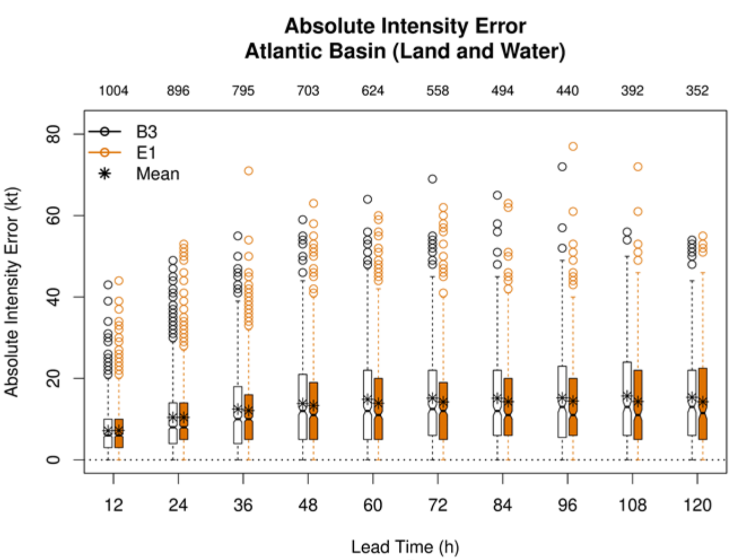
- Jolliffe, I.T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science*. Wiley-Blackwell, 274 pp.
- Murphy, A.H., 1993: What is a good forecast: An essay on the nature of goodness in forecasting. *Weather and Forecasting*, **8**, 281-293.
- Murphy, A.H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev*., **115**, 1330–1338.
- Wilks. D.S., 2019: Forecast verification. *Statistical Methods in the Atmospheric Sciences*, 4th ed., Elsevier, 369–483.
- JWGFVR Forecast verification web page: https://www.cawcr.gov.au/projects/verification/

A few optional exercises

- Interpretation of box plots of error distributions
- A “contingency table” example
- A “discrimination” diagram example

Interpretation of error distributions





- The figures to the left each show box plots of tropical cyclone intensity (wind speed) errors measured in knots as a function of lead (aka integration) time for two pairs of models (B1 and E1 in upper diagram, B3 and E1 in lower diagram).
- The box plots show (from bottom to top) minimum; 0.25th, 0.50th and 0.75th quantile values; upper whisker representing non-outlier values; and outliers (open circles). The mean value for each error distribution is represented by *
- Questions:
  - How would you summarize – in words – the differences in performance for each pair of models?
  - What can you say about average performance of each of the models as a function of lead time?
  - What differences can you describe regarding the characteristics of the outliers for each of the pairs of models?

Examination of a 2x2 contingency table (aka "confusion matrix")

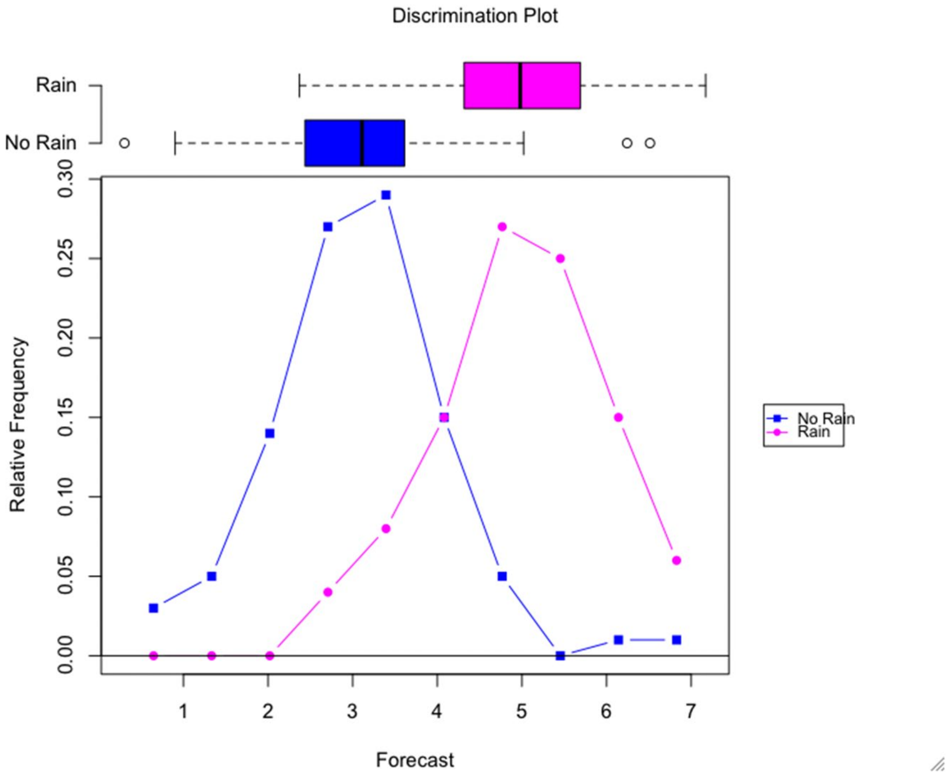
	Observed		
Forecast	**Tornado**	**No tornado**	**Total**
Tornado	28	72	100
No Tornado	23	2680	2703
Total	51	2752	2803

Summary table for a set of 2803 forecasts of the occurrence of tornados. The contingency table shows the frequencies associated with each possible forecast and outcome: (a) tornado was forecast and tornado occurred; (b) tornado was forecast and no tornado occurred; (c) no tornado was forecast and a tornado occurred; and (d) no tornado was forecast and no tornado occurred.

Questions:

- What can you deduce about the quality/performance of the tornado predictions?
- How would you summarize the performance of these forecasts?

Analyzing "discrimination"



This "discrimination" plot shows the empirical distributions of forecast probabilities of rain for times when measurable rain *did not* occur (blue) vs. the distribution of rain forecast probabilities when measurable rain *did* occur (pink). The box plots at the top also summarize these distributions.

- How would you interpret the performance of this set of forecasts to predict rainfall based on this distribution diagram?
- What kinds of information can you gain from this type of distribution plot?
- Would you say that this is a "good" set of forecasts? Why or why not?