

The verification using R & Issues of flatness and inferences with rank histograms

Matt Pocernich
pocernic@ucar.edu

R and the Verification Package

- R Language
 - Statistical programming environment
 - Open source, free, multi-platform
 - More than 1,000 contributed packages
 - Incorporate foreign code such as fortran and C++
 - May/October major version updates
 - Began ~ 1997.

The verification Package

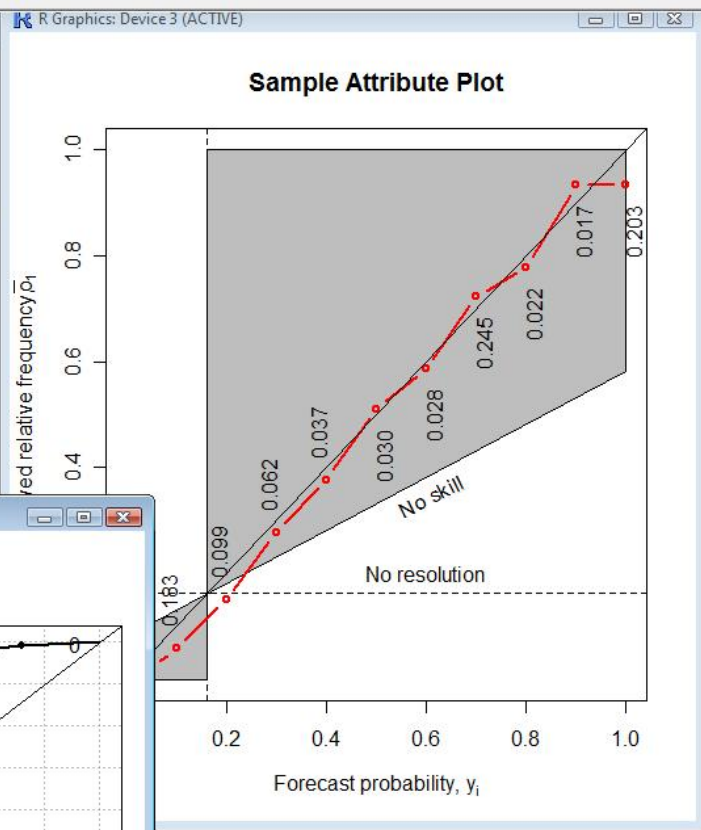
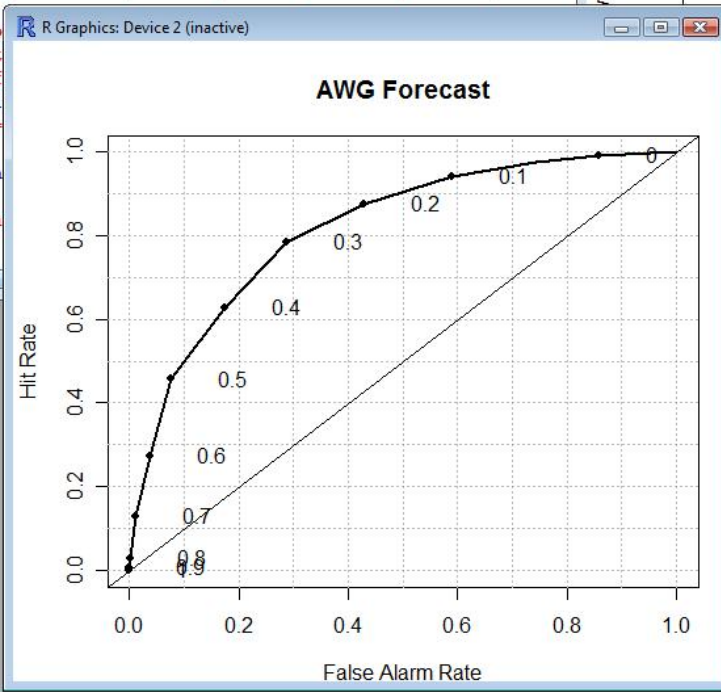
- Created 2003 to consolidate many independent verification functions.
- Basic functions such as roc areas, roc plots, attribute diagrams, conditional quantile plots, scores, some confidence intervals, ...
- Contributed functions – Casati's spatial intensity, Gneiting's circular rps, Brigg's measurement error.
- Feedback from users has been helpful.
- Used in verification tutorials.



```
R Console
> library(verification)
Loading required package: waveslim
Loading required package: fields
Loading required package: spam
Package 'spam' is loaded. Version 0.13-2 (2008-01-04).

Type demo( spam) for some demos, help( Spam) for an overview of

Try help(fields) for an overview of this library
Loading required package: boot
Loading required package: CircStats
Loading required package: MASS
> ## Data from Wilks, table 7.3 page 246.
> y.i <- c(0,0.05, seq(0.1, 1, 0.1))
> obar.i <- c(0.006, 0.019, 0.059, 0.15, 0.277, 0.377, 0.511,
+ 0.587, 0.723, 0.779, 0.934, 0.933)
> prob.y<- c(0.4112, 0.0671, 0.1833, 0.0986, 0.0616, 0.0366,
+ 0.0303, 0.0275, 0.245, 0.022, 0.017, 0.203)
> obar<- 0.162
> attribute(y.i, ob
> data(prob.frcs.dat
> A <- verify(prob.f
If baseline is not i
> roc.plot(A, main =
> device()
Error: could not fin
> x11()
> attribute(y.i, oba
> |
```



R: Forecast verification utilities. - Mozilla Firefox

File Edit View History Bookmarks Tools Help

file:///C:/Users/pocernic/Documents/R/R-2.6.2/library/verification/html/00Index.html

Getting Started Latest Headlines

The Comprehensive R Archive Net... RAL | RAL home R: Forecast verification utilities.

Forecast verification utilities.

Documentation for package 'verification' version 1.26

User Guides and Package Vignettes

Read [overview](#) or browse [directory](#).

Help Pages

analysis.dat	Spatial Observation Dataset.
attribute	Attribute plot
attribute.prob.bin	Attribute plot
brier	Brier Score
conditional.quantile	Conditional Quantile Plot
crps	Continuous Ranked Probability Score
crps.circ	CRPS Statistics for circular data
disc.dat	Discrimination plot dataset.
discrimination.plot	Discrimination plot
forecast.dat	Forecast forecast dataset.
int.scale.verify	Intensity-Scale Verification Model
IS	Alternative Intensity Scale Function
leps	Linear Error in Probability Space (LEPS)

April 10, 2006

DTC VERIFICATION WORKSHOP 2006

Limitations/ Next steps

- No ensemble verification functions
 - Plot rank histograms
 - For a given threshold, convert to probabilistic forecast.
- Use proper bootstrapping method (boot)
- Convert to S4 system – provides a stronger handshake between data and methods. (Important for spatial).
- Web based apps? See Andy Loughe's work.

Verifying ensemble forecasts

- Wind and temperature forecasts, surface and select levels of upper atmosphere.
- Four models some differ by initial parameterizations some by model physics.
- Data from October 2006 ~ 28 days.
- Many spatially correlated data points for each time period.
- Several sources of upper elevation observations.

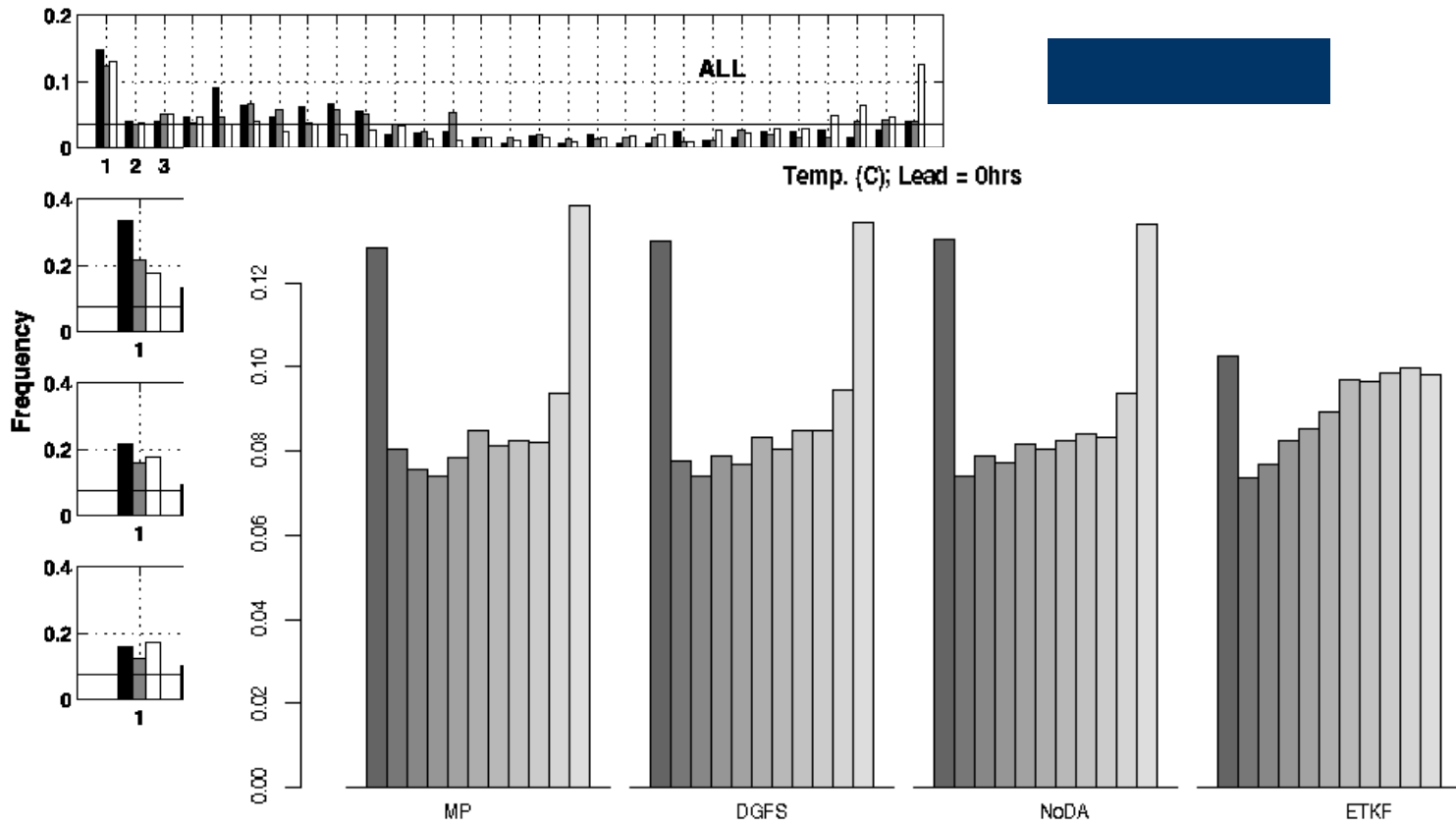
Processing steps

- Ensemble members each adjusted for bias, individually.
- Some models consist of anonymous members, others are identifiable.
- Excluded forecasts with missing members.
- For probability = $(\text{rankobs} - 1) / N$
- For ignorance scores, = $(\text{rank obs} + \frac{1}{2}) / (n.\text{ens} + 1)$

Comparing Flatness

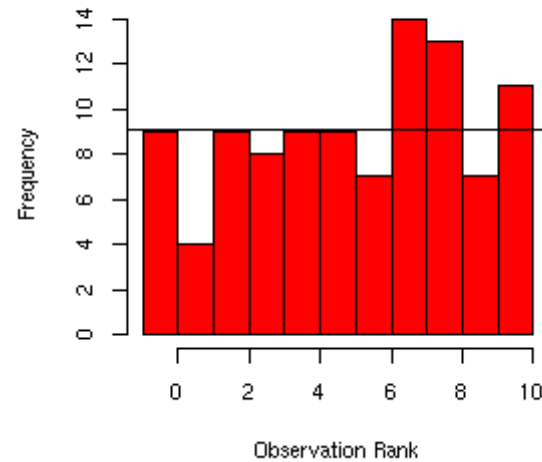
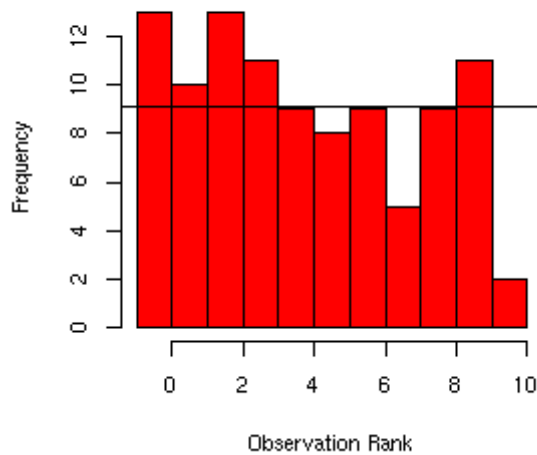
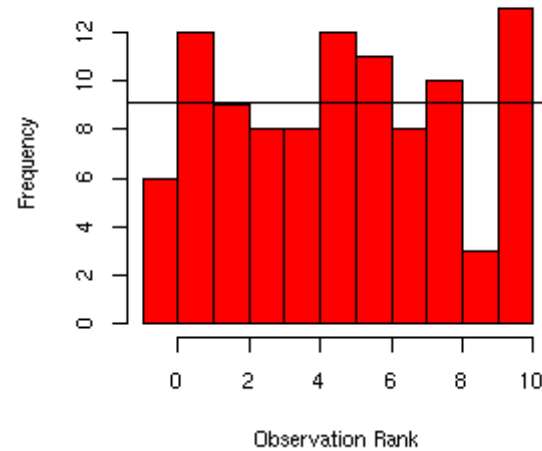
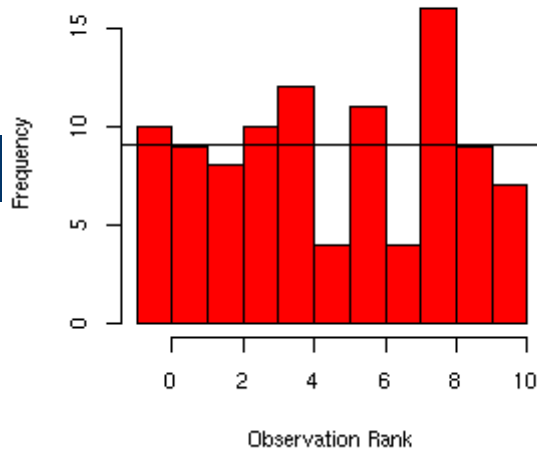
- Many articles plot series of rank histograms and make plots. With the exception of gross over or under dispersion, making a comparisons or comments is difficult.
- Rank histograms should not be used alone since they do not depict refinement or sharpness.

Examples of comparing histograms



“Perfectly Reliable Forecasts”

(Simulated, $N = 10$, $M = 100$)



April 16, 2008

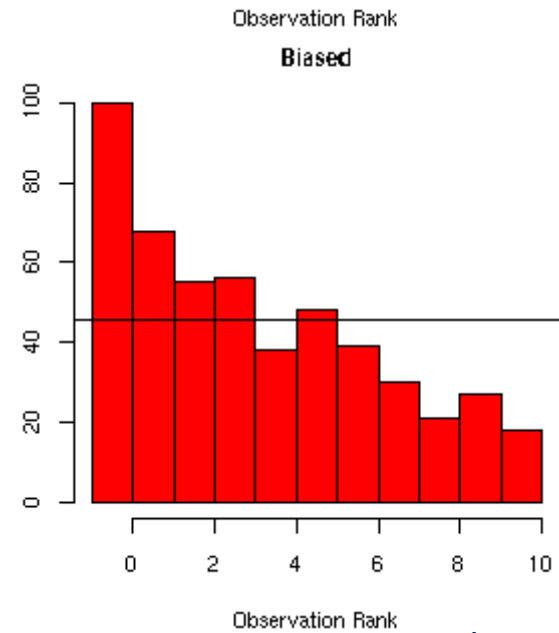
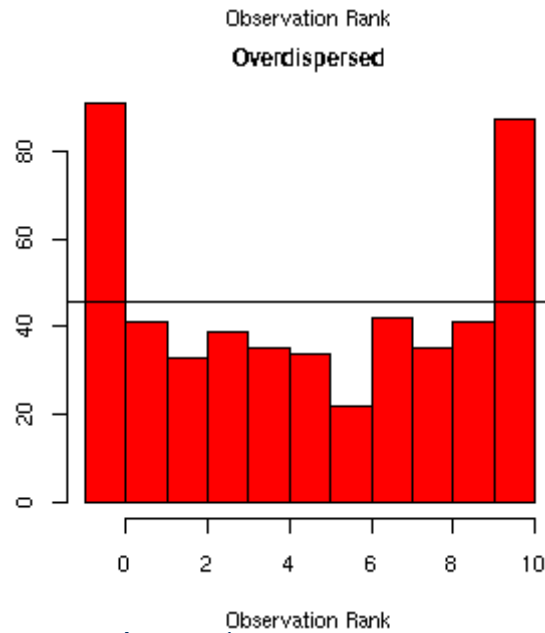
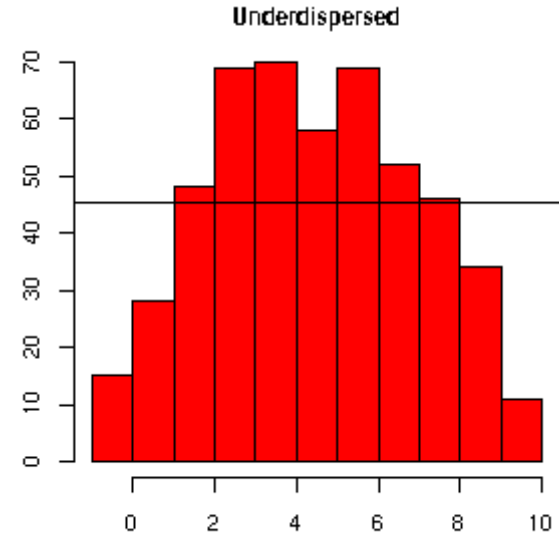
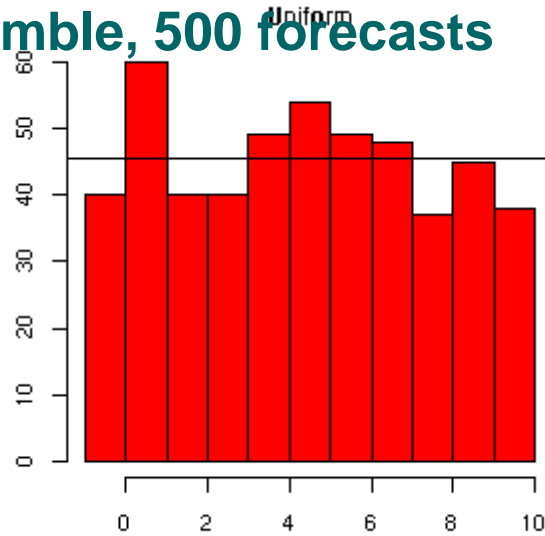
DTC Verification Workshop 2008

Outline

- Describe a several scores which summarize flatness.
- Show how these scores behave on contrived datasets.
- Some results and interpretation.

Simulated Ensemble Forecast Outcomes.

10 member ensemble, 500 forecasts



“Binned Probability Ensemble technique” — Anderson(1996)

- 1 = reliable
- >1 = not reliable
- <1 successive realizations are not independent
- S_k = number of outcomes in each bin.
- N = # of ens. Members
- M = # of realizations

$$\Delta = \sum_{k=1}^{N+1} \left(S_k - \frac{M}{N+1} \right)^2$$

$$\Delta_0 = \frac{MN}{N+1}$$

$$\delta = \frac{\Delta}{\Delta_0}$$

Reliability Index (Hacker 2006)

$$\text{RI} = \frac{\text{mean dist from ideal bin count}}{\text{ideal bin count}}$$
$$= \sum_{k=1}^{N+1} \left| \frac{S_k}{M} - \frac{1}{N+1} \right|$$

Lower RI -> more reliable

No information on shape

**Can be adjusted for differing # of ensembles.
(Not shown)**

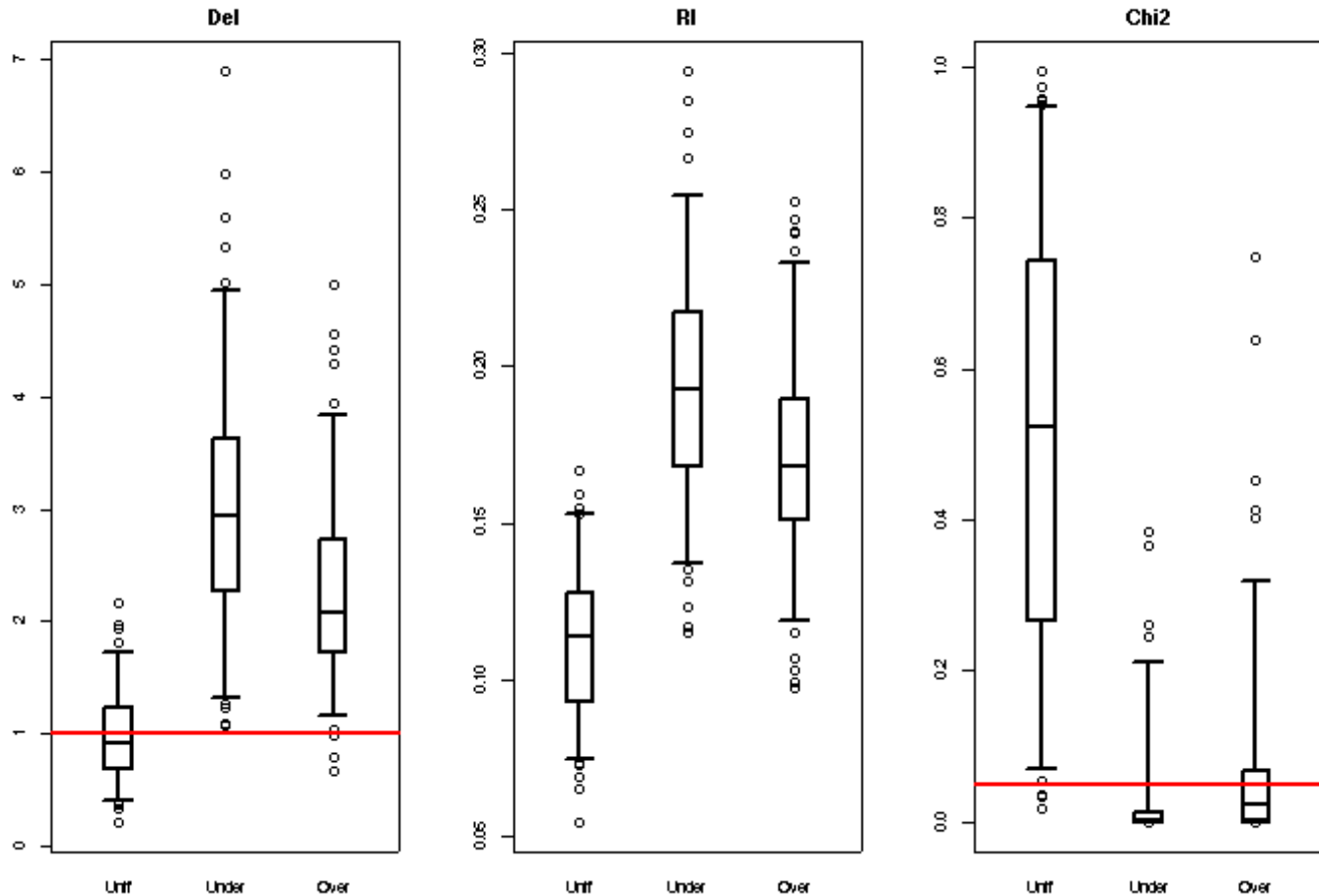
Goodness of Fit Tests

- Chi-squared
 - Null hypothesis data from specified distribution
 - Requires independent ensembles.
 - Likely to reject as N increases.

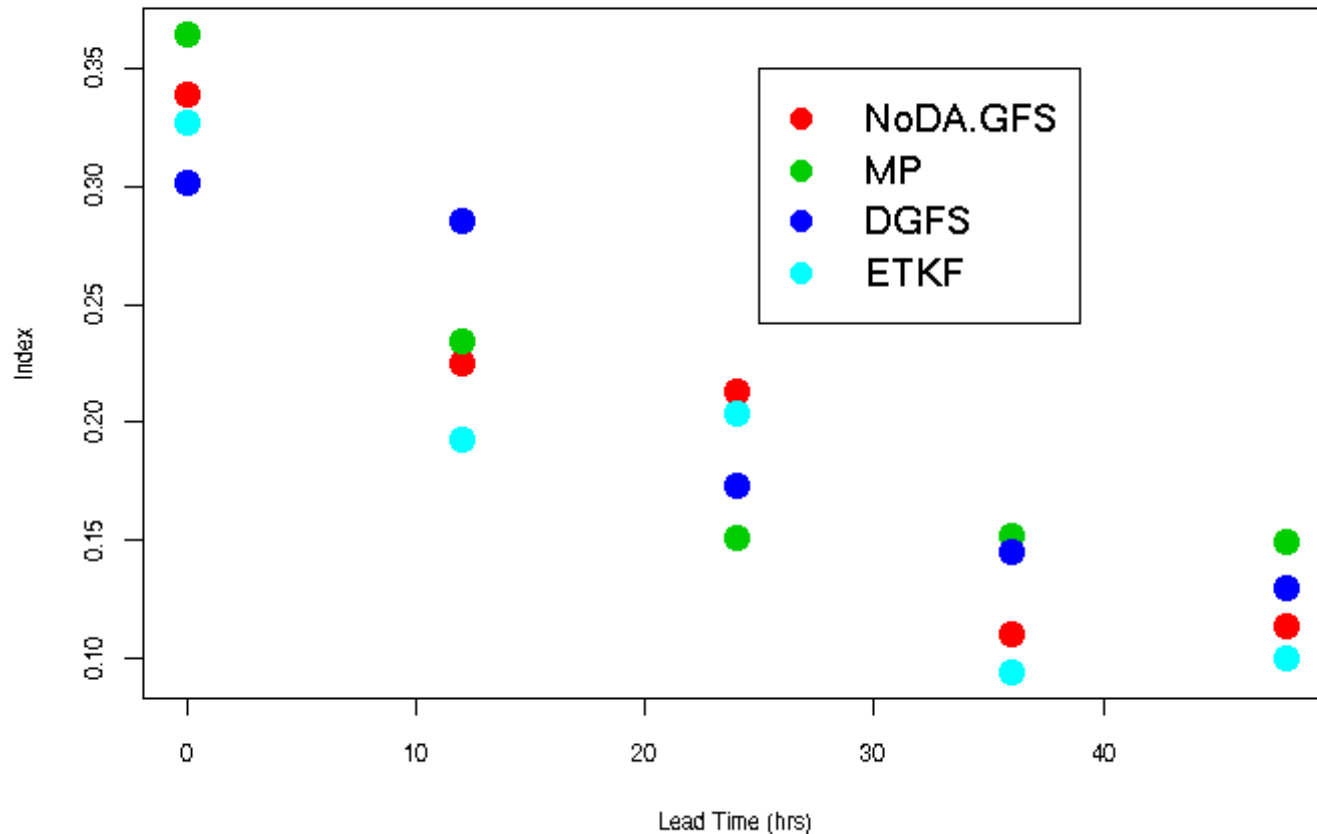
$$\chi^2 = \sum \frac{(\#Observed - \#Expected)^2}{\#Observed}$$

Making inference using tests

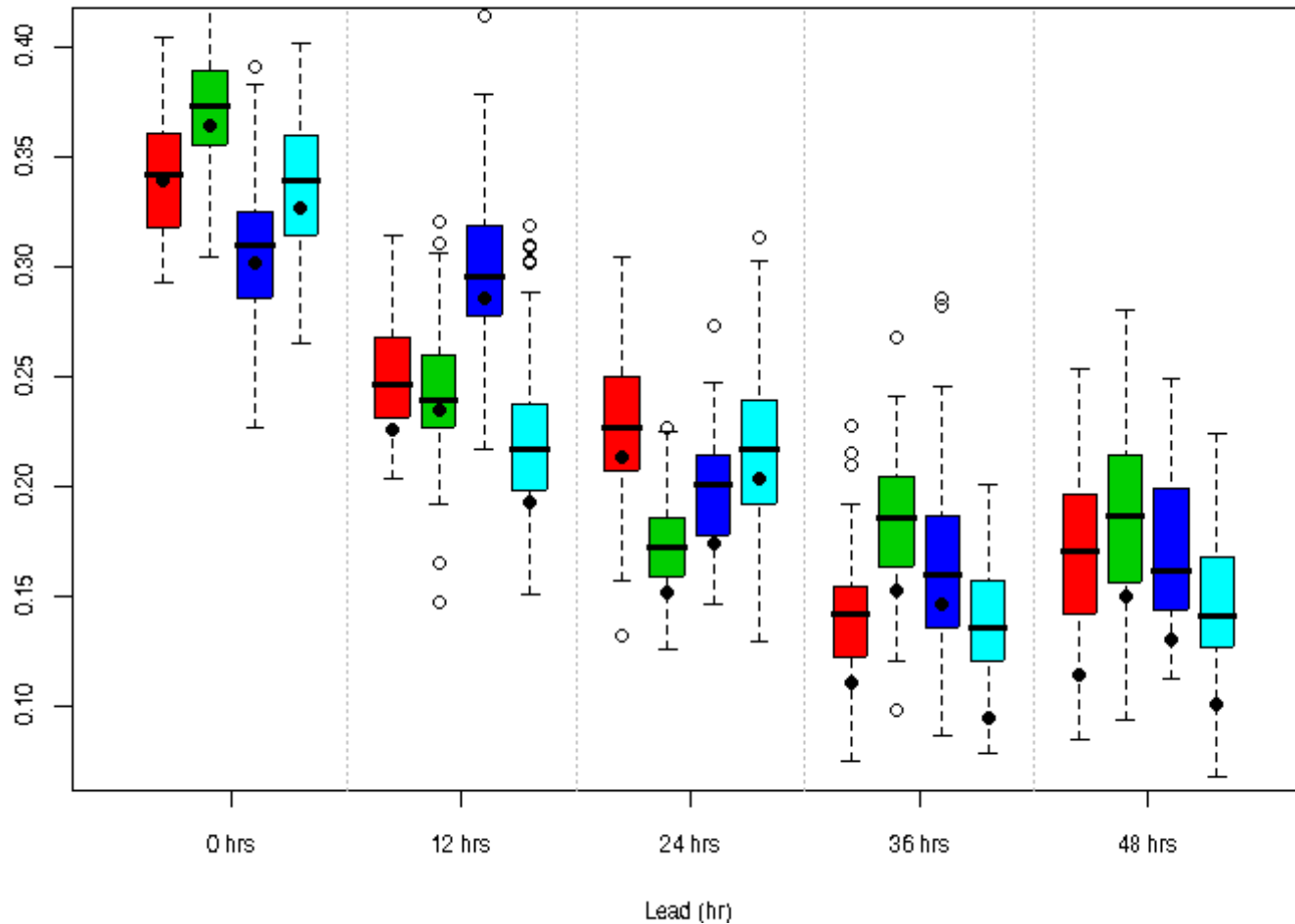
10 ensembles; 500 realizations



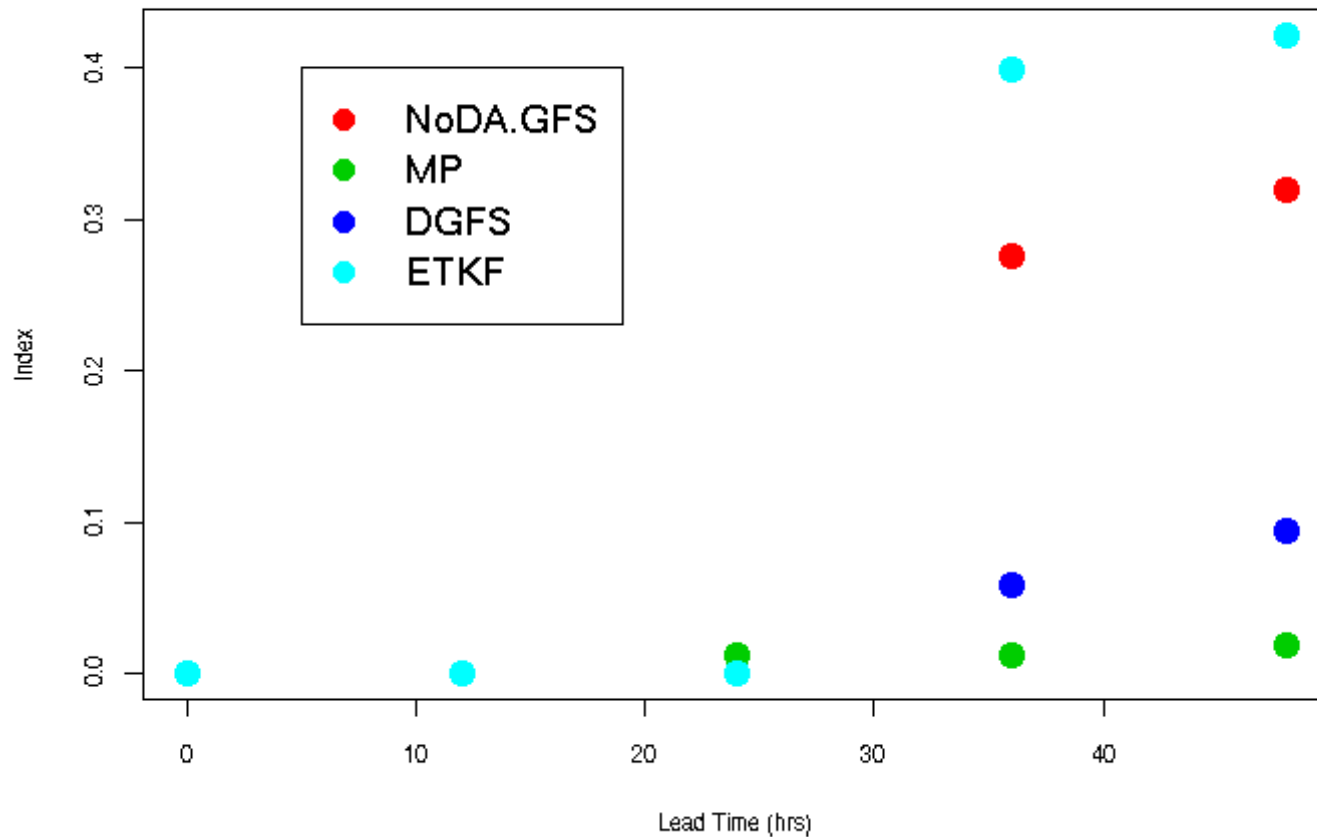
Point estimates of reliability index



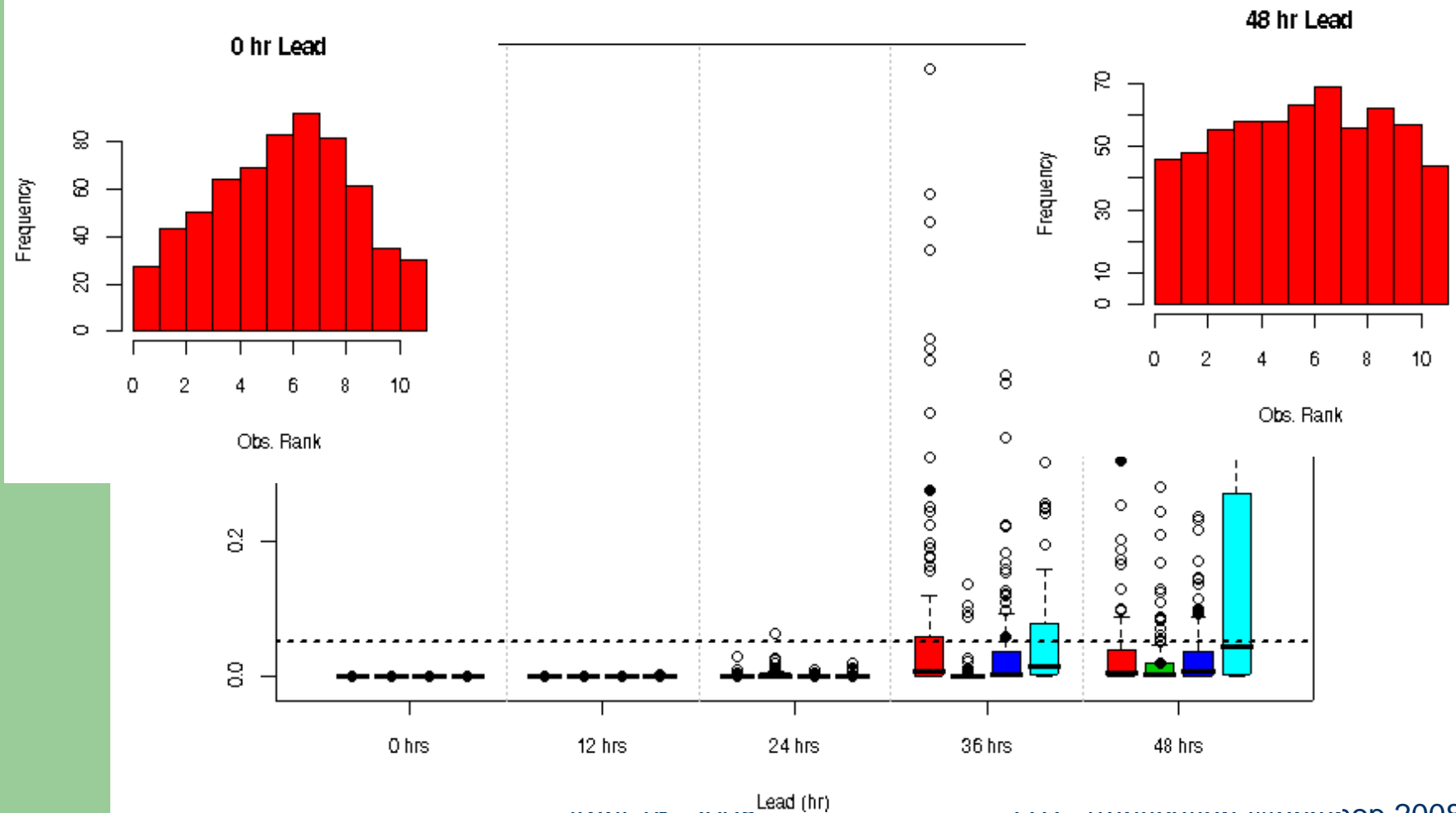
RI – summary of 100 resamples



Chi-Square – p-values



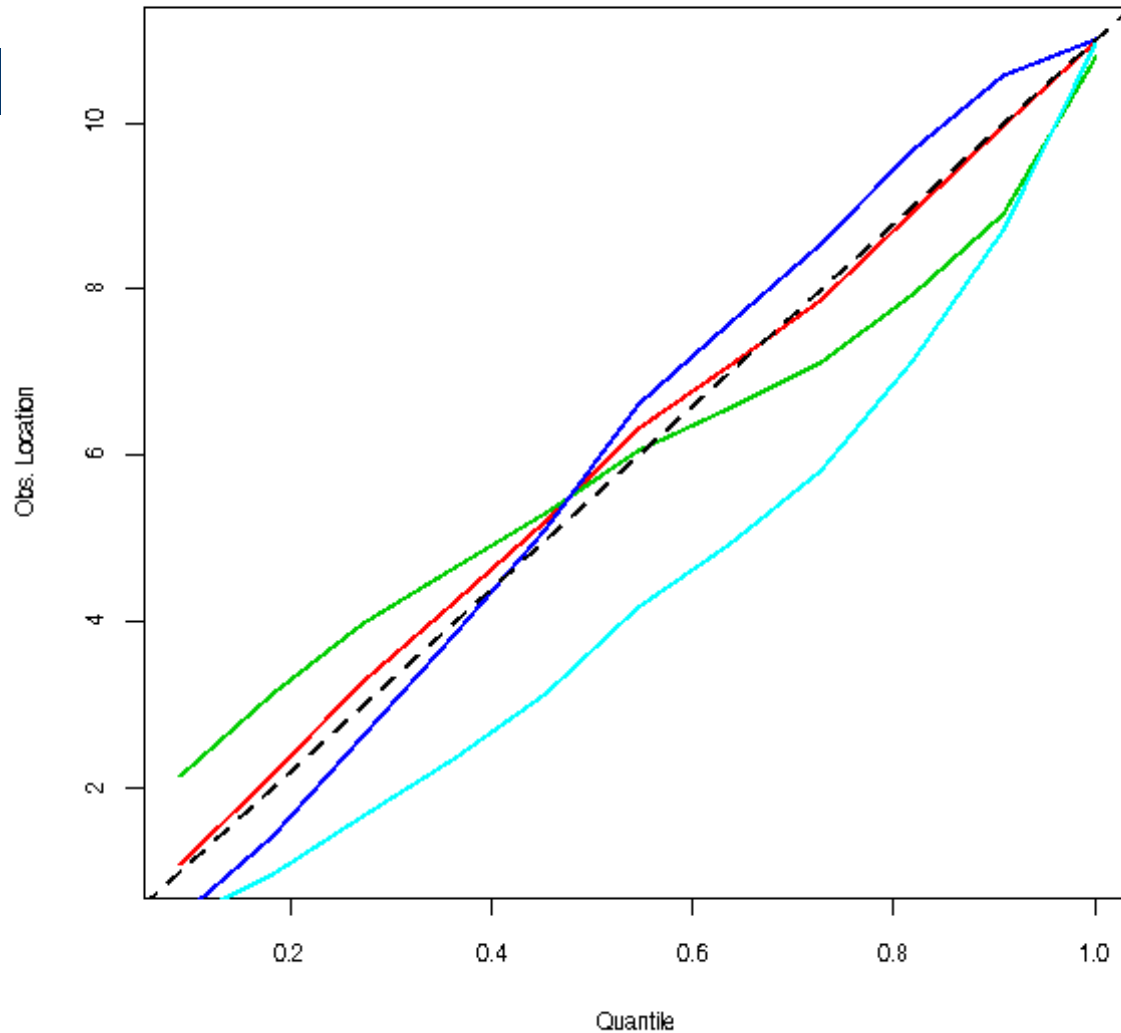
Chi Square – 100 resamples



How to summarize and draw make valid comparisons?

- For a given lead time,
 - Are any models different – ANOVA type comparison.
 - Comparison between models, for a given lead time, 6 individual comparisons?
- Considering all lead times.
 - Trends modeled independently. What trend is expected?

Describing Shape of rank histogram



Concluding remarks

- Work in progress
- Some functions from the ensemble work will be moved to R.
- Seek ability to succinctly describe differences as a function of limited data?
- More discussion on inferences tomorrow.