www.ec.gc.ca

Lawrence.wilson@ec.gc.ca

Ensemble forecast verification – Recent work, new ideas and issues

Laurence J. Wilson

3rd International Verification Methods Workshop
31/01/07

Environment Canada Environnement Canada

Canada

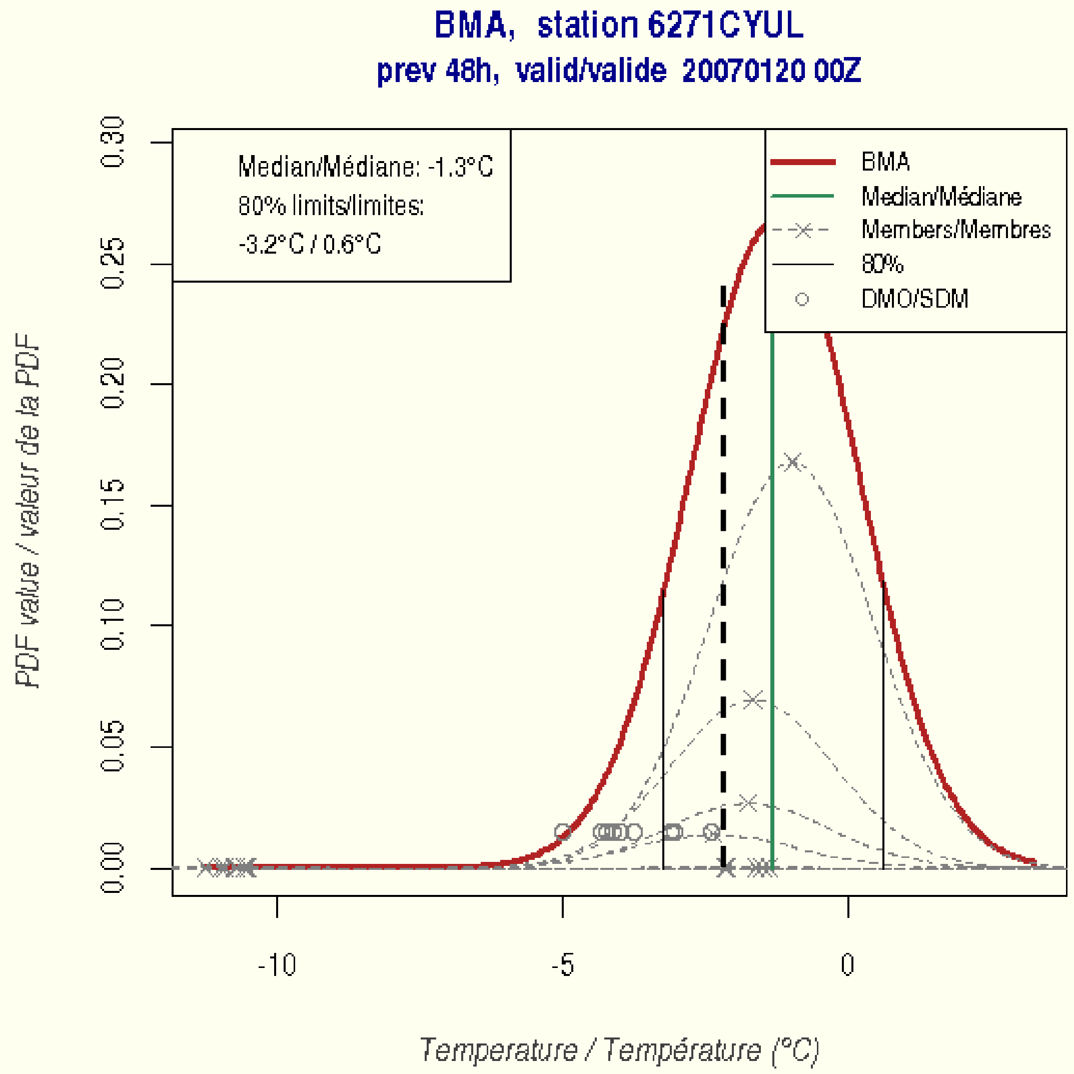
Ensemble Verification

Ensemble verification involves comparing single observations with ensemble distributions, or at least, multiple forecasts

What is a perfect ensemble forecast?

Is reliability enough?

Reliability: “For all instances where a pdf f(x) is forecast, the distribution of observations is equal to f”



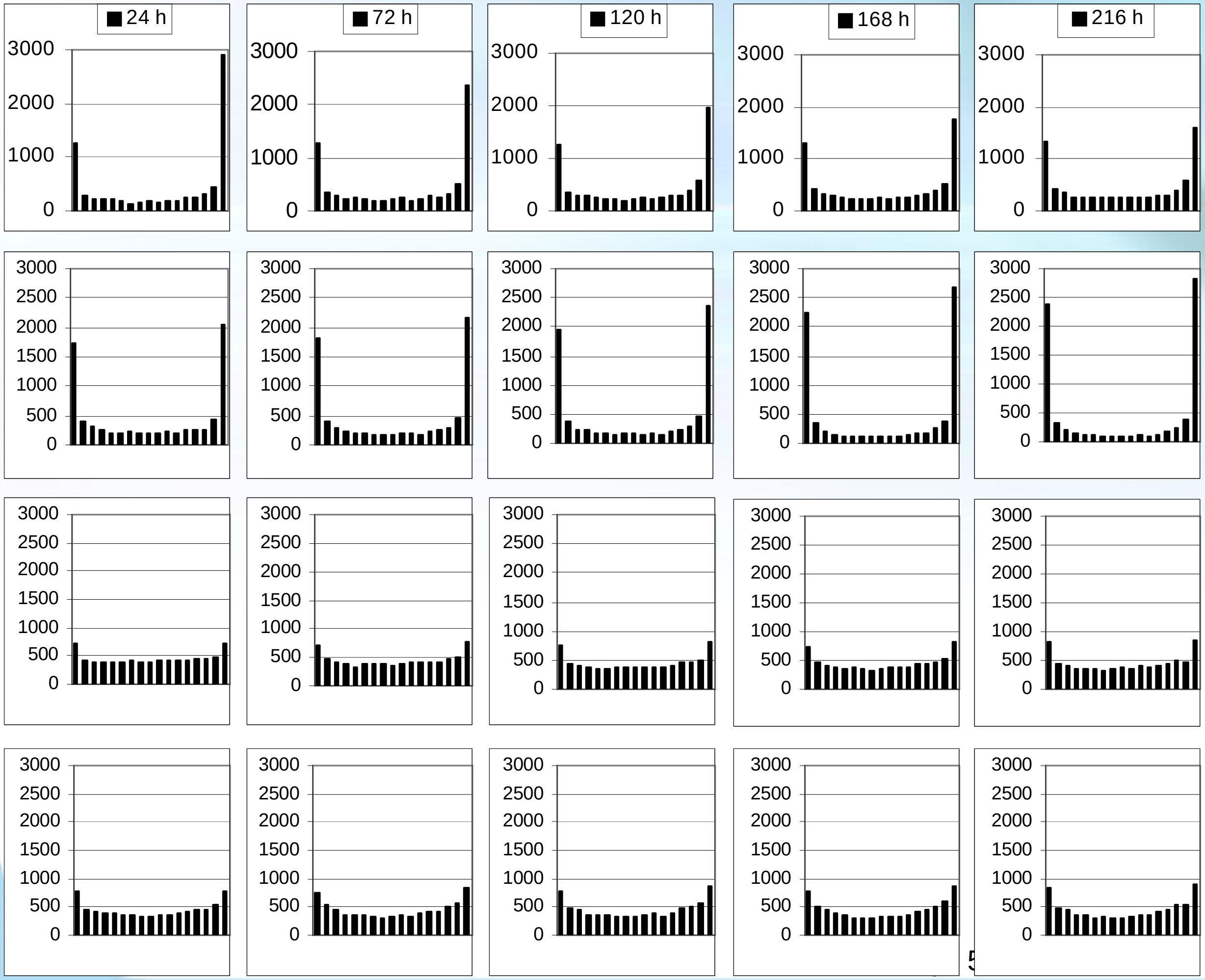
Outline

- Introduction – What is a perfect ensemble forecast?
- Overview of verification methods
- New (relatively unused methods)
- Issues:
  - False skill
  - Resolution vs ROC
  - Observation error

Survey of verification methods for ensembles

- Evaluate the ensemble distribution
  - Rank Histogram*
  - CRPS, CRPSS (Hersbach, 2000)
  - Minimum Spanning Tree (Smith, 2001; Wilks, 2004)
- Evaluate the ensemble distribution in the vicinity of the observation
  - Wilson et al, 1999
  - Ignorance score (Roulston and Smith, 2002)
- Evaluate probabilities from the ensemble distribution
  - Brier score (accuracy), reliability, resolution
  - Reliability (attributes) diagrams
  - ROC area (discrimination)*
  - BSS, RPSS (skill)

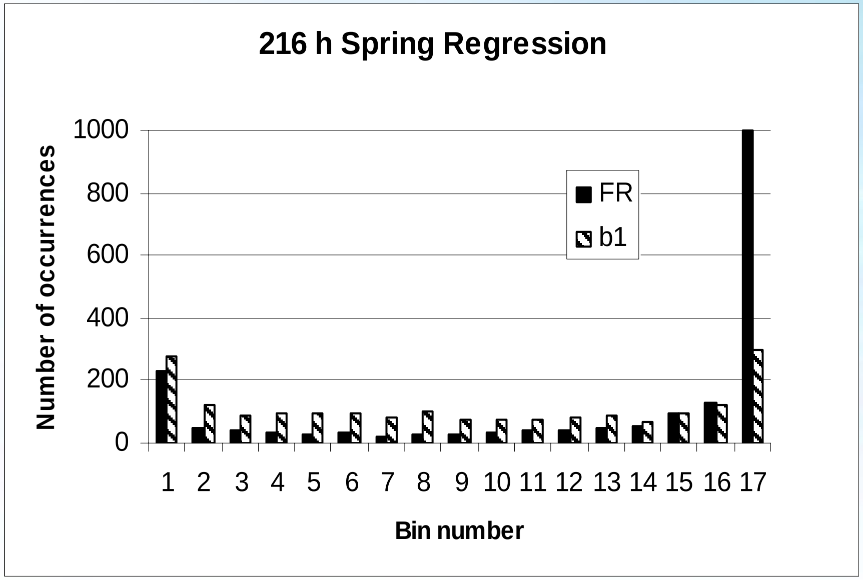
Quantification of “departure from flat”

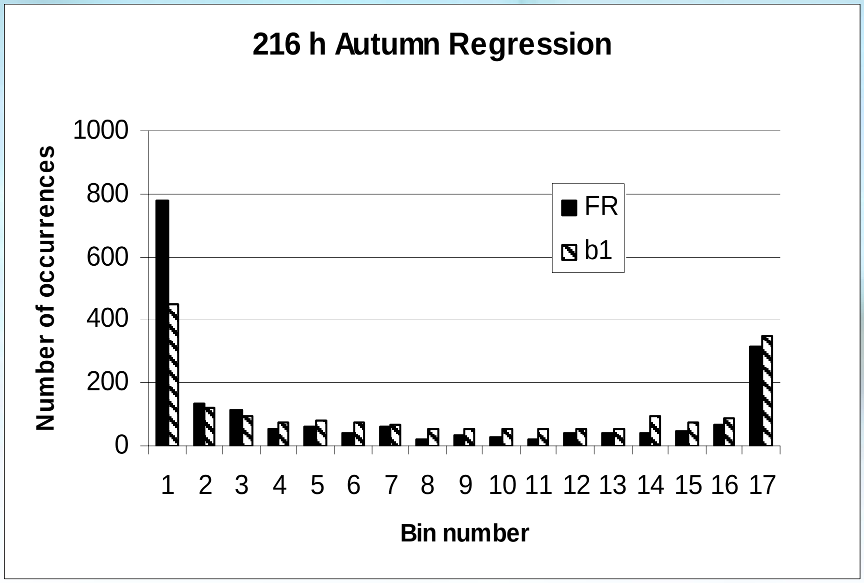


$$RMSD = \sqrt{\frac{1}{N+1}\sum_{k=1}^{N+1}\left(S_k - \frac{M}{N+1}\right)^2}$$

$$\sqrt{\frac{MN}{(N+1)^2}}$$

Rank Histogram

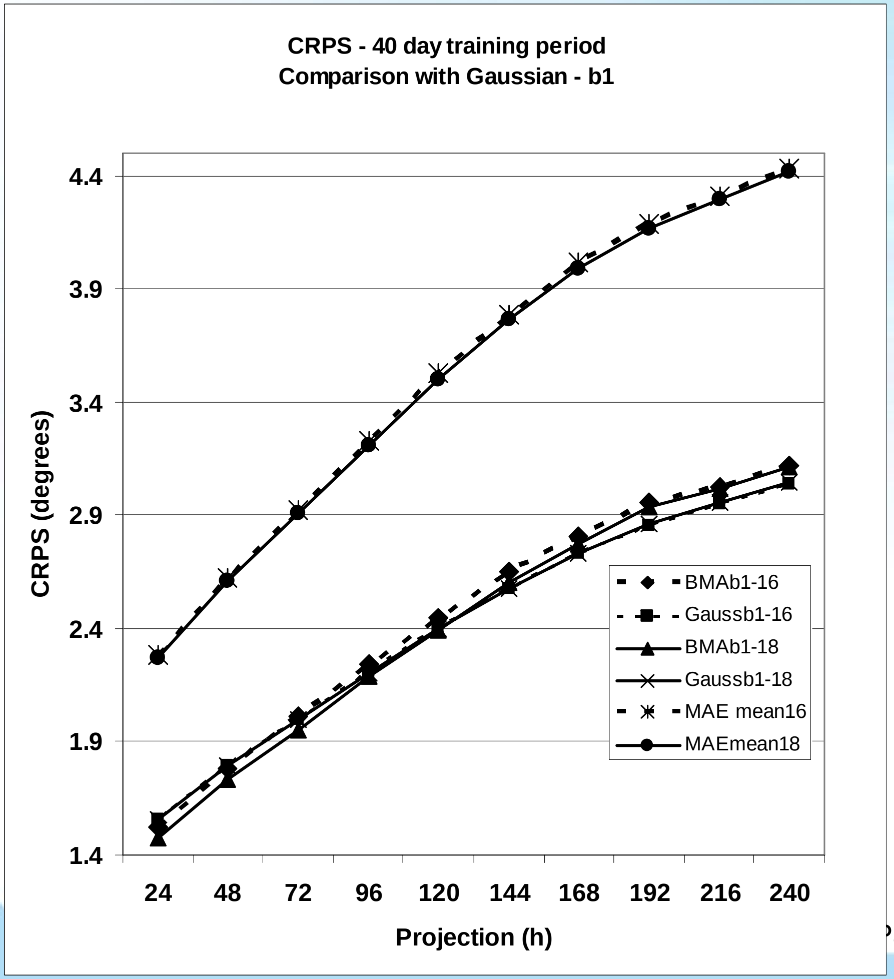




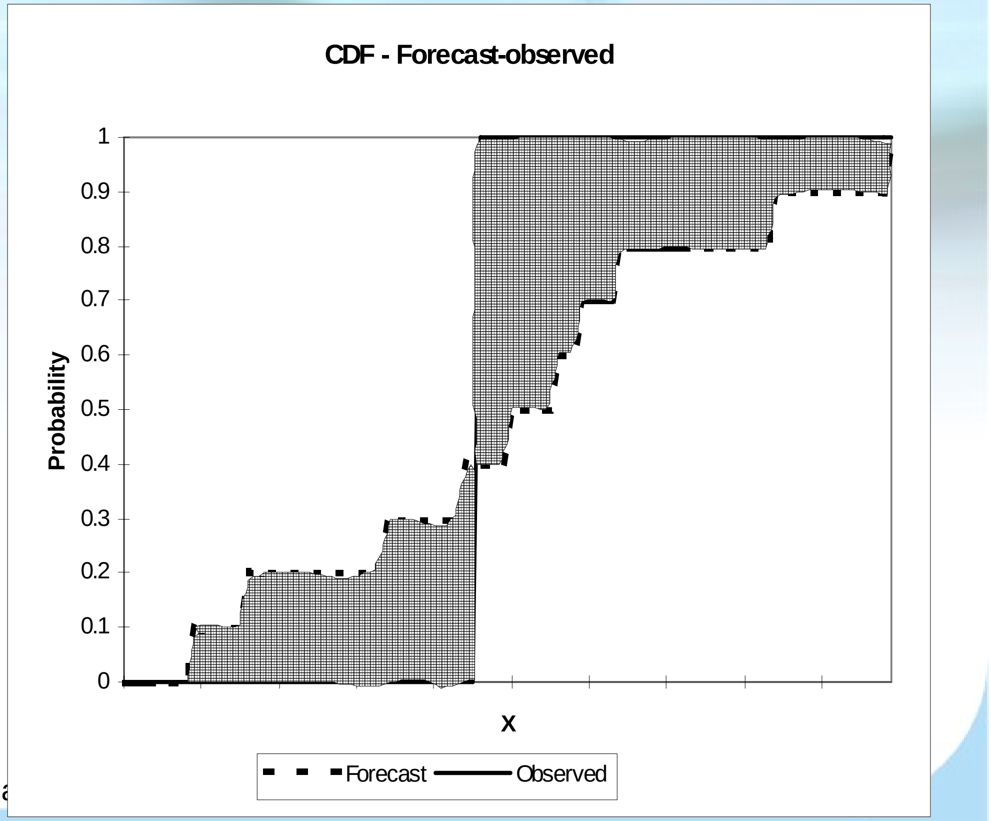
Cold bias in the spring, warm bias in the fall, cancels out when accumulated over the year

Pointed out by Hamill 2001

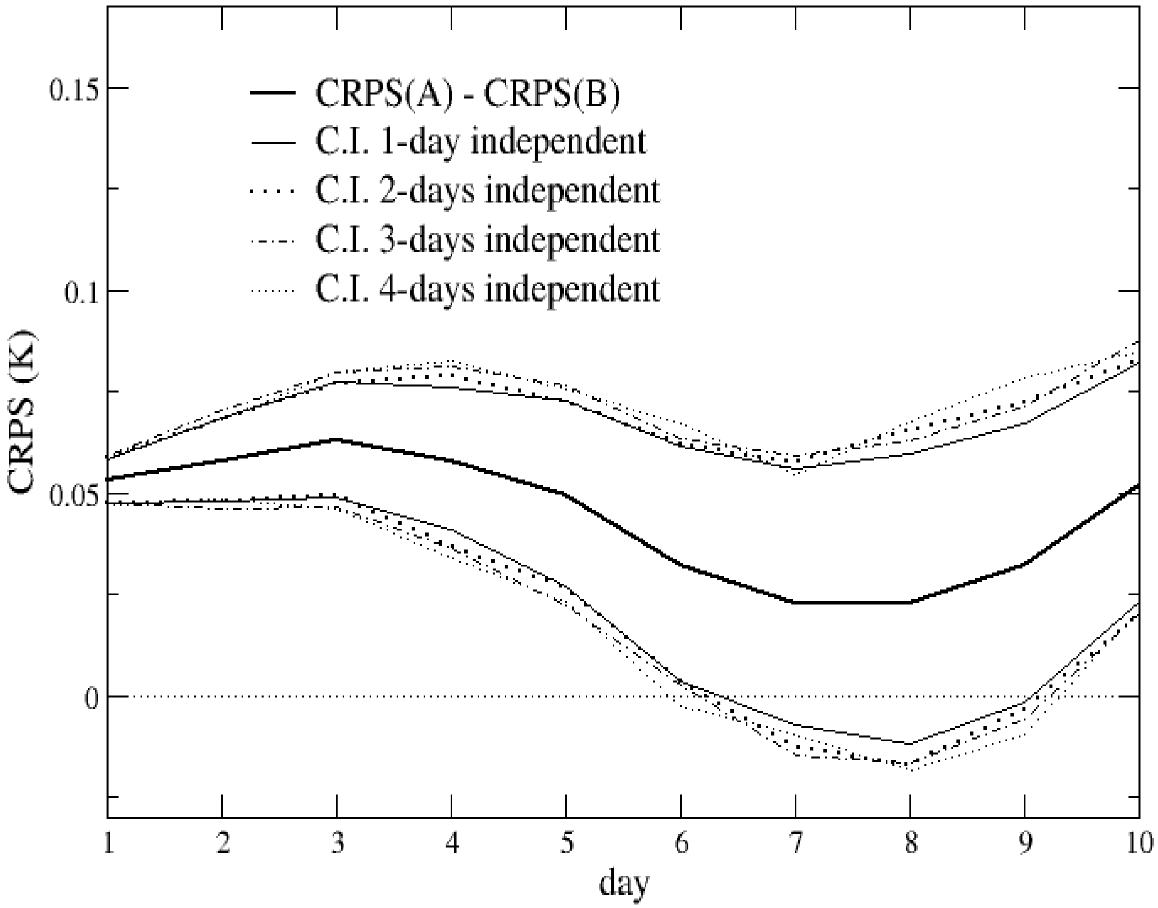
Continuous Rank Probability Score



$$CRPS(P, x_a) = \int_{-\infty}^{\infty} \left[P(x) - P_a(x)\right]^2 dx$$



CRPS difference – Temperature 850 mb

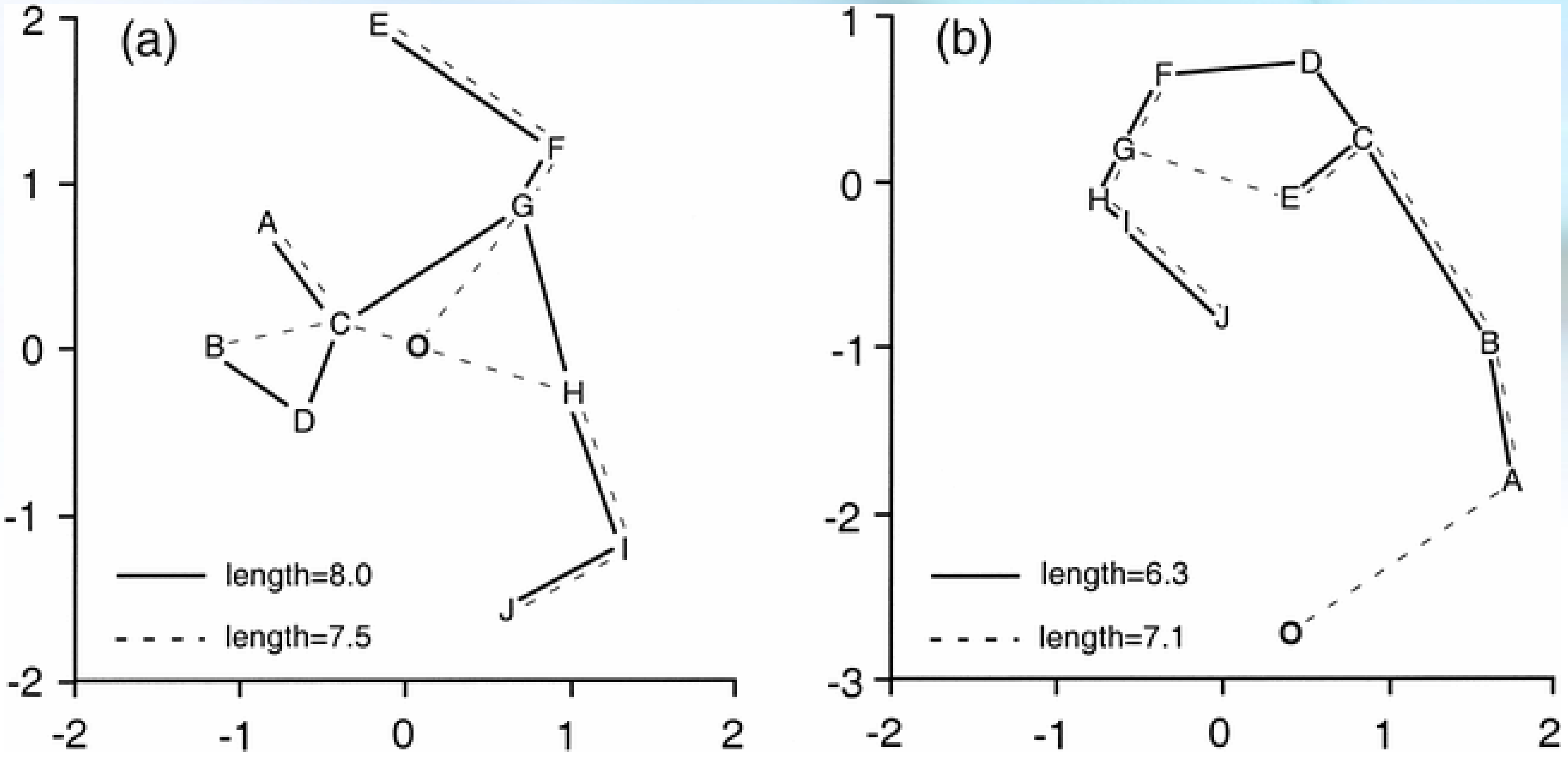


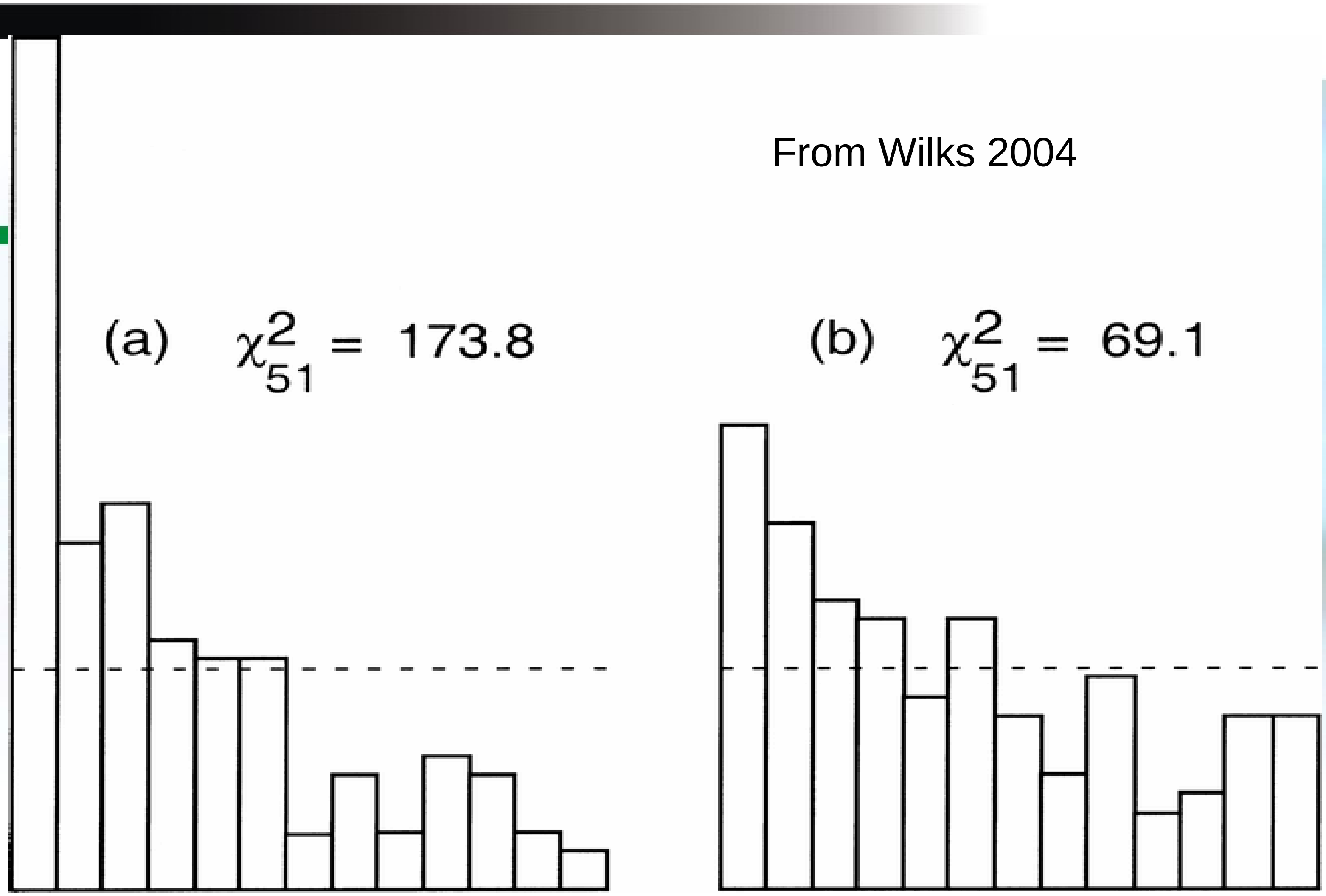
From Candille et al 2007 (to appear in MWR)

CRPS difference for 60 days sample, 17000 fcst-observations pairs

Use of block bootstrapped confidence intervals

Minimum Spanning Tree – MST (Wilks 04)





From Wilks 2004

Rank histograms for ECMWF forecasts, 149 forecasts, 15 dimensions

-5 UK stations

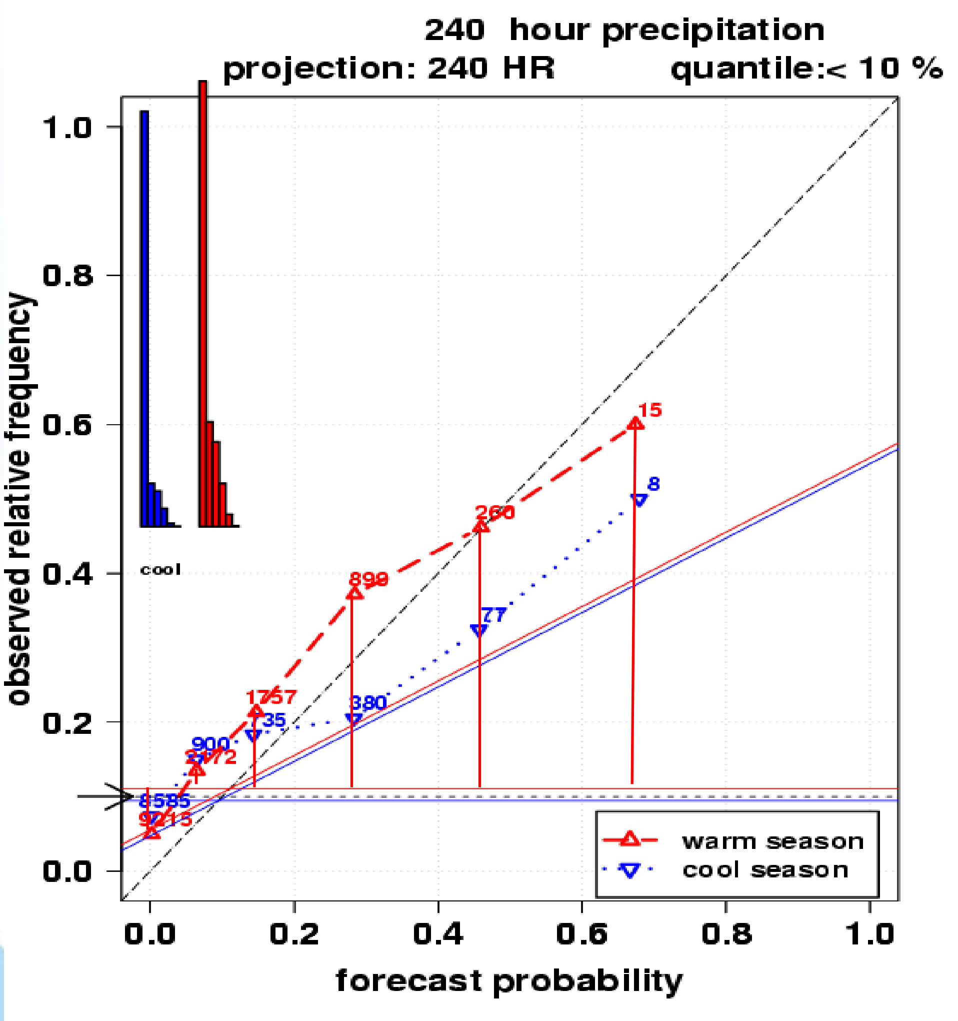
-3 variables – 2 m temperature, 10 m wind and cloud cover

-variables standardized (Mahalanobis method and by dividing by std (R))

Issues in Ensemble Verification

- 1. Resolution vs the ROC
  - – Discussed in the recent Thorpex discussion groups, apparently some confusion.
  - – Murphy’s framework:
    - Resolution can be defined as the variance of the conditional distribution of observations given the forecast probability
    - The ROC area relates to the conditional distribution of the forecasts given the observations – the separation of the two “likelihood distributions”

Resolution



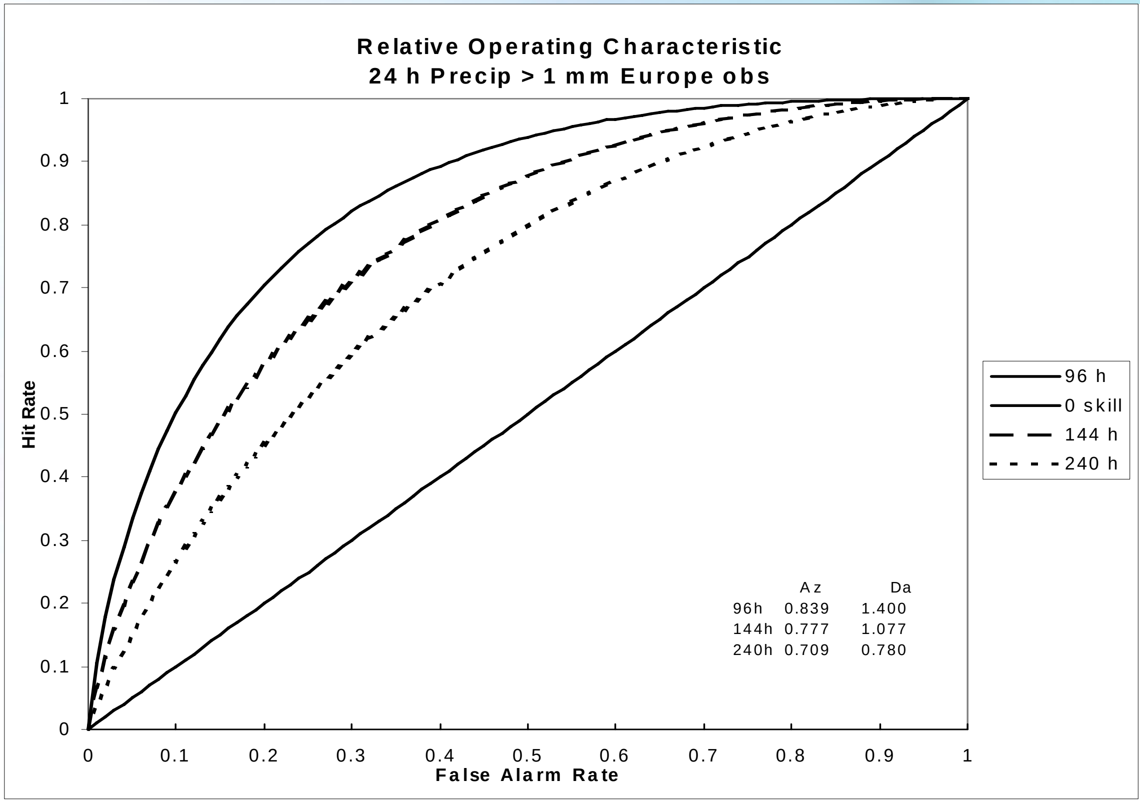
Resolution:

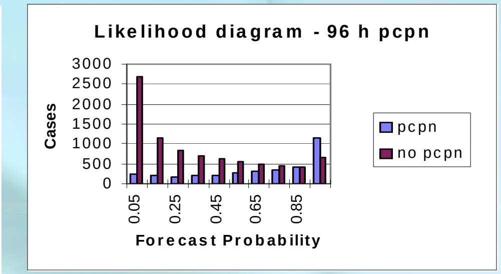
-The variance of the conditional observed frequencies about the climatological frequency, conditioned on the forecast

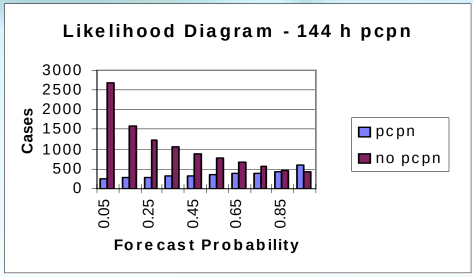
-A component of the Brier score

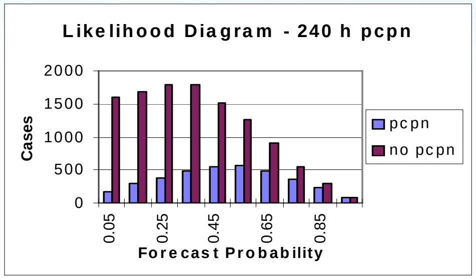
-Steeper slope than 45 degree line suggests over-resolved forecasts.

ROC example - 24 h POP (>1 mm)









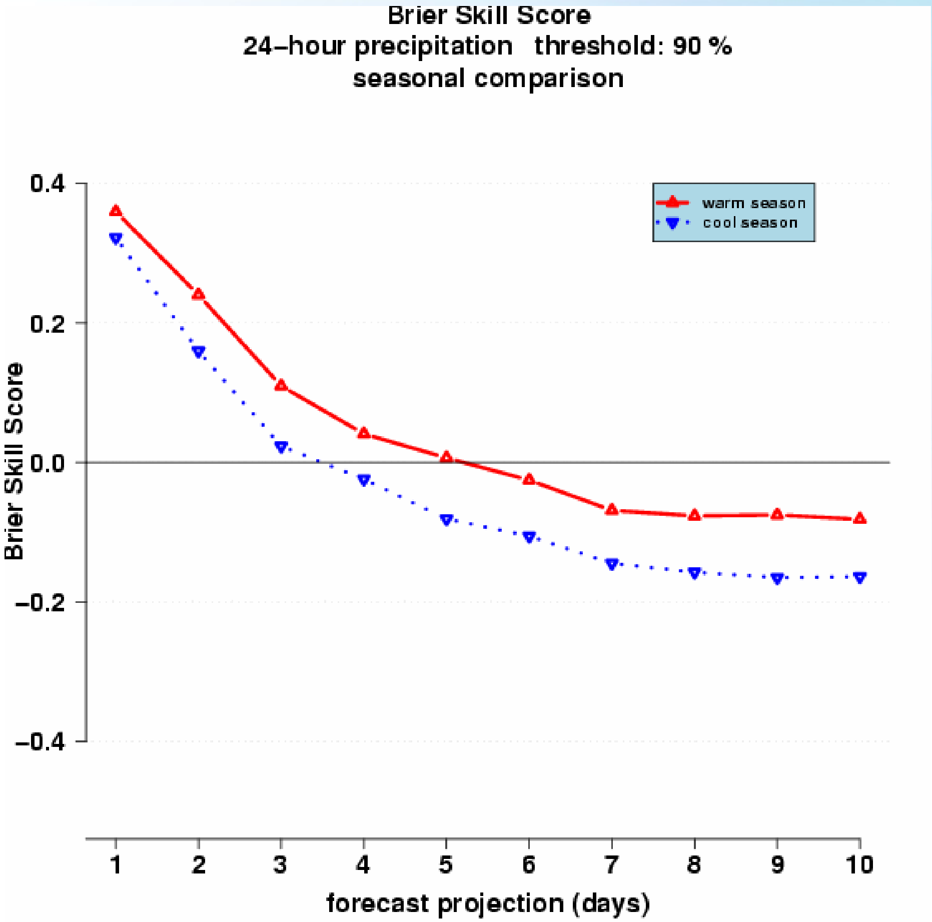
The Likelihood diagram shows the two conditional probability distributions

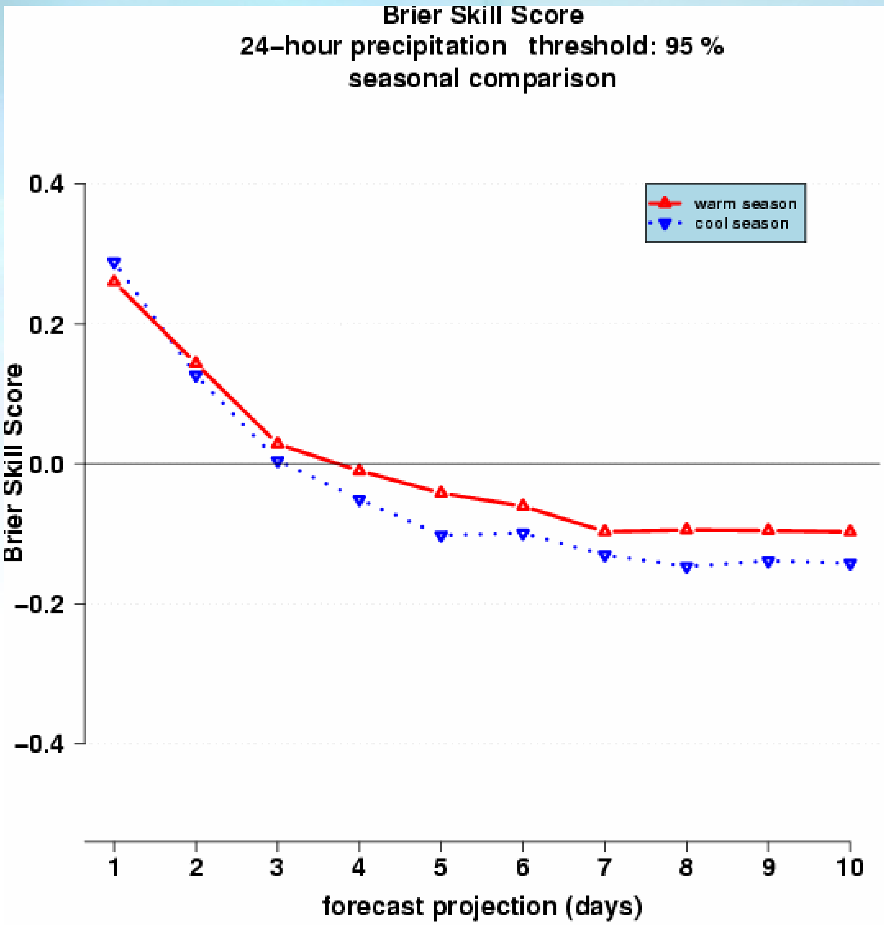
The distance can be computed directly and is given in terms of the std of the distribution for non-occurrences

Issue 2: False Skill

- Called Simpson’s paradox (Simpson, 1951)
- Hamill and Juras, 2007
- The tendency to include spatial and/or temporal variance in climatology in a scoring system.
- A problem for skill scores and the ROC, wherever there is an underlying climatology
- Remedy:
  - 1. Reference skill scores to LOCAL climatology, stratify as much as possible by season
  - 2. To keep sample sizes large enough, express variables as anomalies from long term climatology
- Example: Verification of extreme precipitation forecasts

BSS for 90% and 95% threshold





Issue 3: “Observation” error

- Recent papers on impact of observation errors on ensemble verification e.g. Saetra et al 2004
  - Suggests that maybe the underdispersion shown by rank histograms is due to not taking into account “observation errors”
- Relates to discussion of “representativeness error”

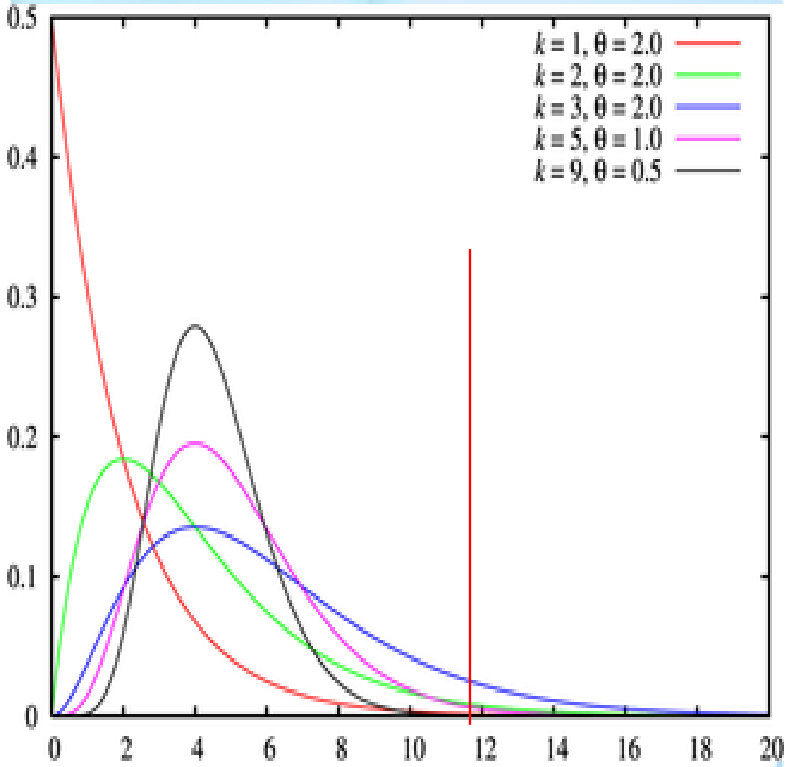
Conclusion

- Ensemble verification methodology is beginning to settle down, a few methods are finding general favour and are widely used
- The coming of ensembles has spawned renewed interest in probability verification methods, and there are many new papers out on the properties of scores, old and new.

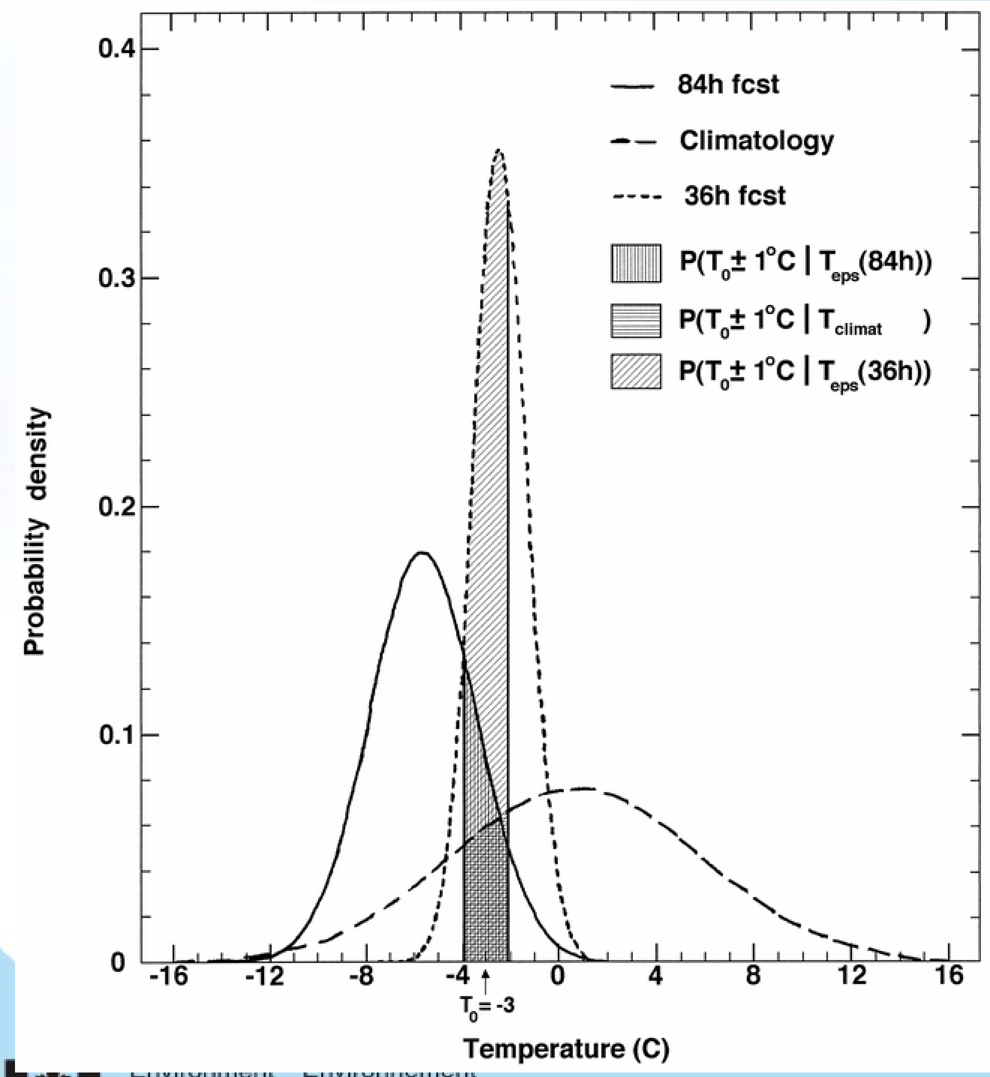
Thank you!

Data and method

- Data
  - 3.5 years of ensemble forecasts of precipitation from 36 Canadian stations, 24h accumulations, 0 to 10 days
  - Corresponding observations quality controlled without reference to models
  - Verification sample stratified into warm and cool seasons
  - Long-term precipitation climatology (~30 years) for all 36 stations as distribution
- Method
  - Using the long-term climatology, find 90th, 95th and 99th percentile thresholds for each station.
  - E.g. 90th percentile for Vancouver is 14.4 mm
  - Probability of exceedence of these thresholds as estimated from the ensemble forecast distribution (gamma distributions)



Probability score



-The probability assigned by the ensemble in the vicinity of the observation

-Maximized for sharp forecasts, correctly positioned

-can be used to evaluate one forecast

-not strictly proper

Ignorance Score (Roulston and Smith, 2002)

- From information theory, the number of bits needed to transmit the probability of the verifying category
- IGN = $-\log_2(f_i)$ where f_i is the probability assigned to the verifying category.
- Goes to infinity for 0 probability
- Heavily penalizes low probabilities
- Similar to probability score in that it considers the verification in the vicinity of the observation only