



[www.ec.gc.ca](http://www.ec.gc.ca)

Lawrence.wilson@ec.gc.ca

# *Verification of Ensemble Forecasts: A look to the future*

**Laurence J. Wilson**  
Environment Canada



Environment Canada  
Environnement Canada

Canada 

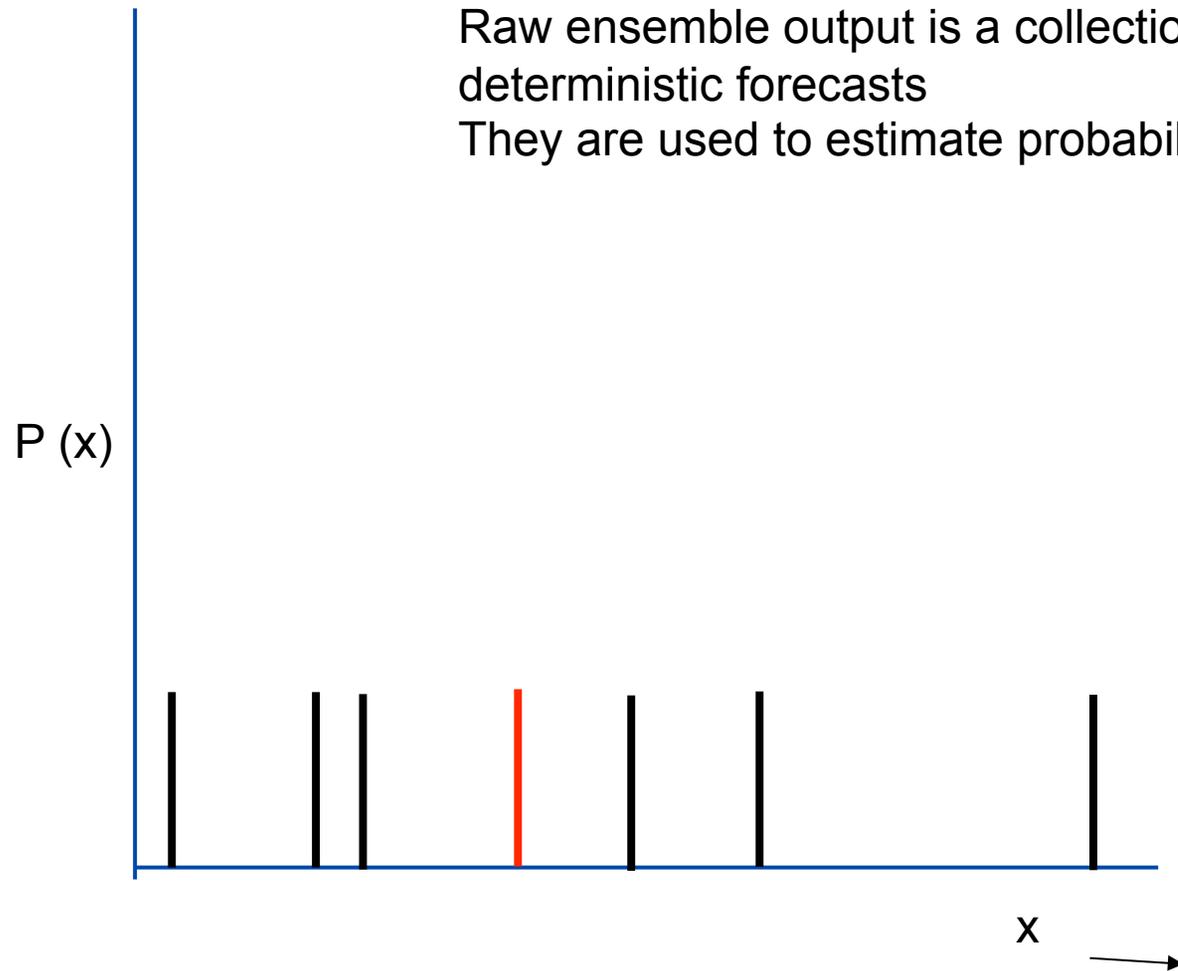
## *Outline*

- Introduction – The interpretation of ensemble forecasts into probabilities
- Where we are now – a very quick survey of proposed methods
- Factors influencing the future – issues
- Some examples of new methods
- Conclusions



# Interpreting an ensemble forecast

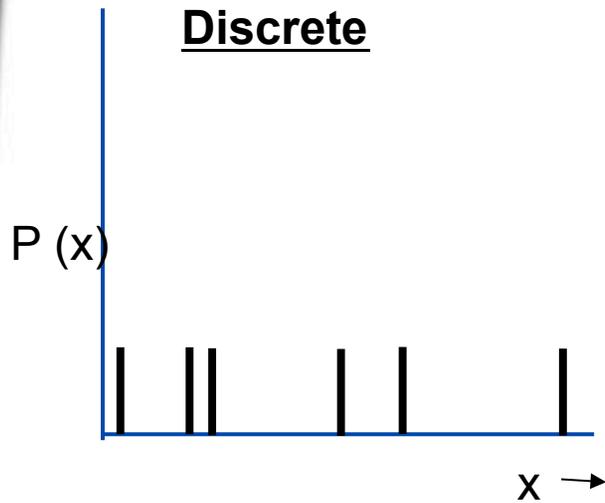
Raw ensemble output is a collection of deterministic forecasts  
They are used to estimate probabilities



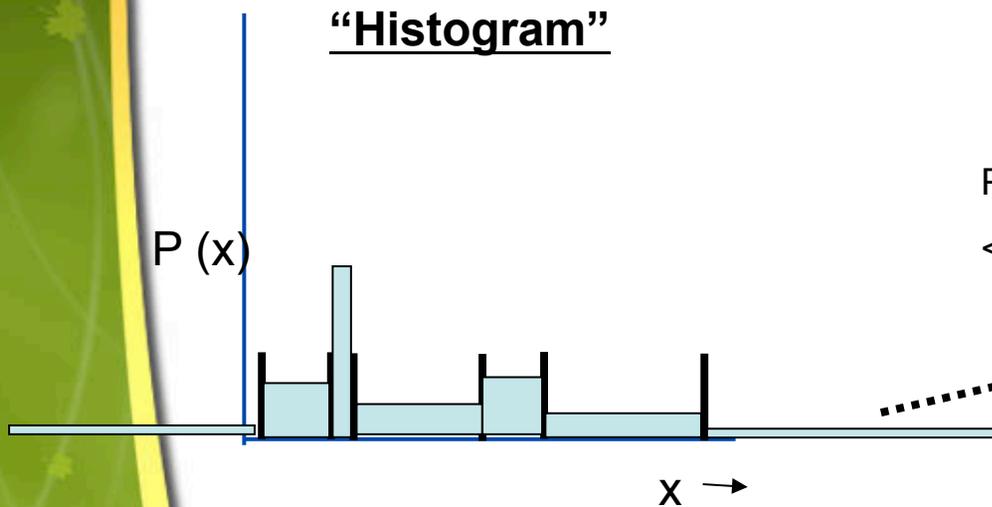
# PDF interpretation from ensembles

pdf

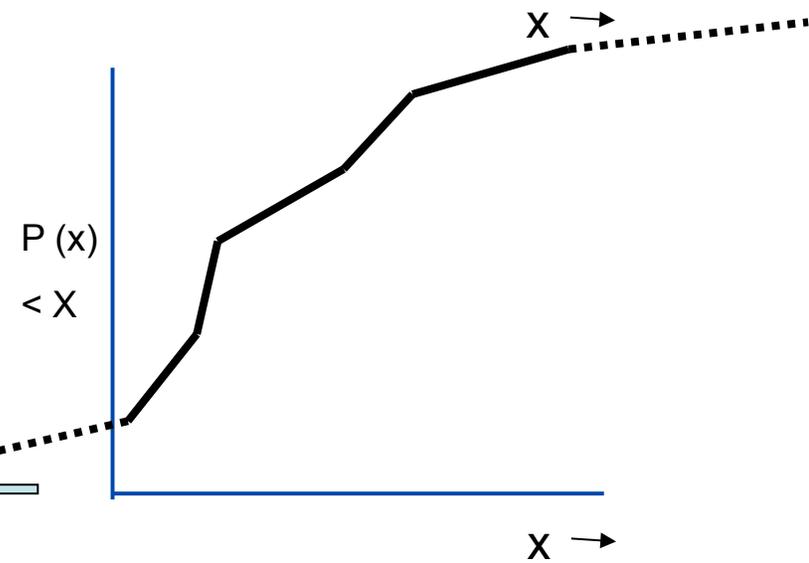
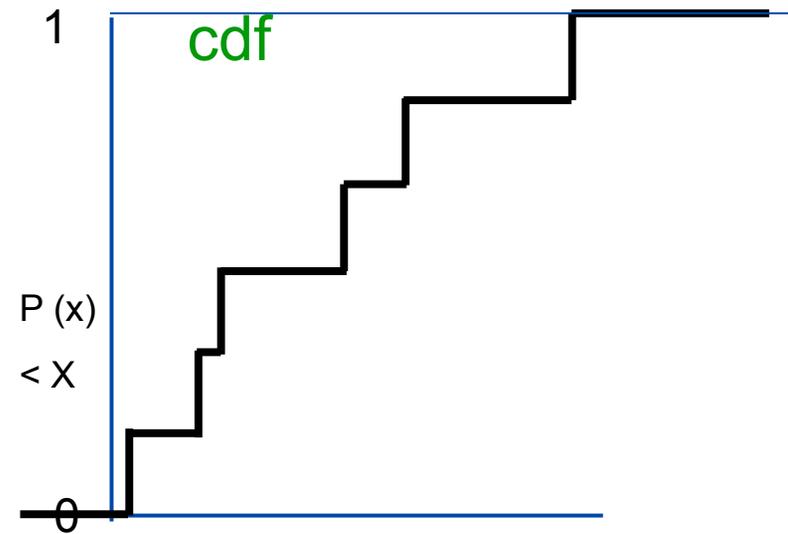
Discrete



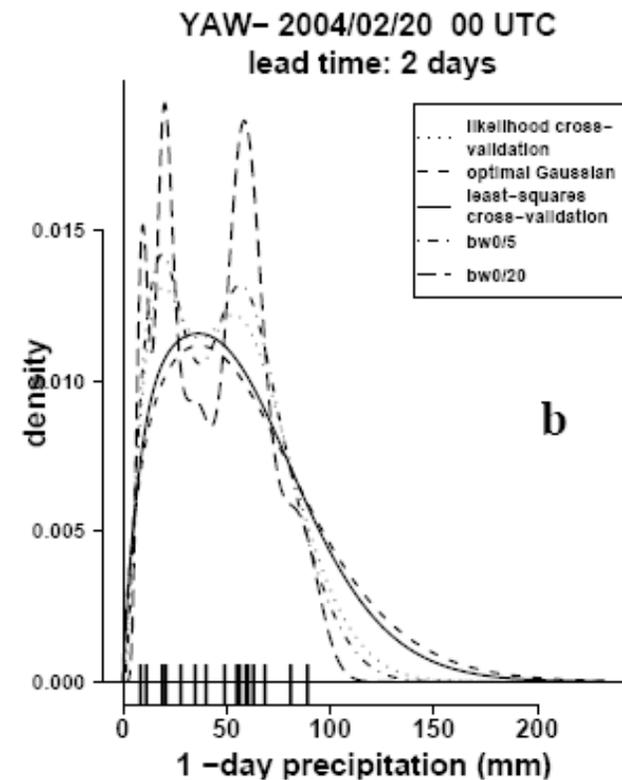
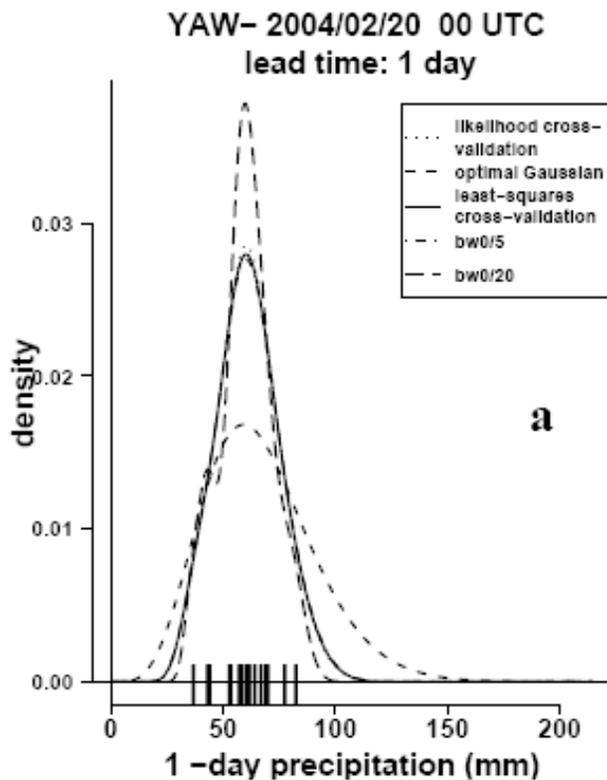
"Histogram"



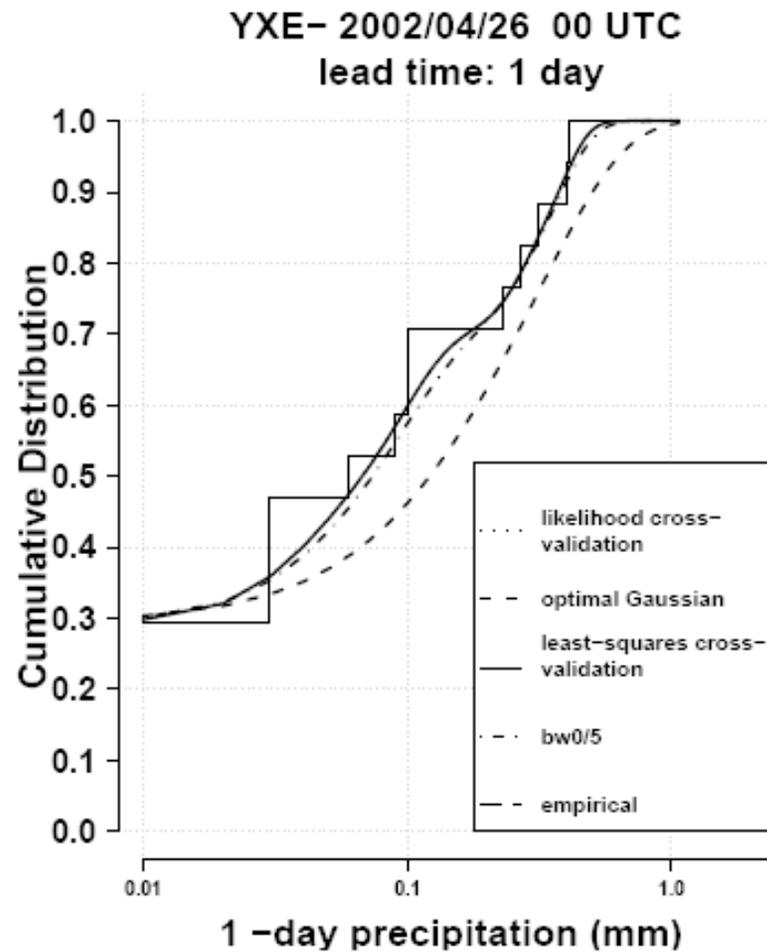
cdf



# Examples of distribution fitting (Peel and Wilson, 2008)



# Kernel Density fitting of ensemble cdfs (Peel and Wilson, 2008)

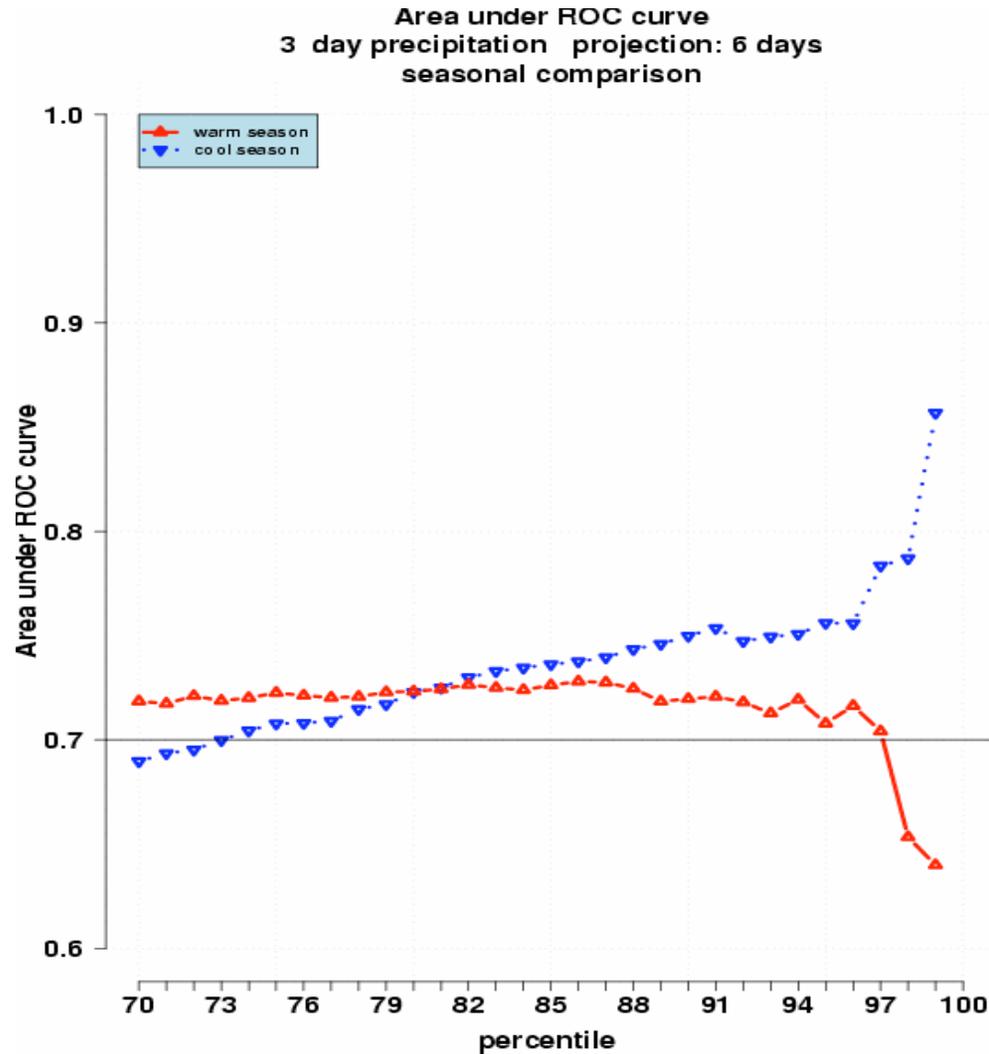


## *Why density fitting is needed*

- For extreme event prediction, to estimate centile thresholds.
- For some scoring rules e.g. ignorance
- Assists with the ROC computation.
- Kernel density estimation
  - Is non-parametric
  - Amount of smoothing determined by the bandwidth
  - Gaussian kernels fine for unbounded variables; gamma kernels for precip.
- Simple to compare ensembles with different numbers of members



# ROC area – 72h accumulation – as function of threshold



# *Estimating probabilities from ensembles*

- Probability, pdf and cdf estimates are **interpretations** of eps output
- Three kinds:
  - Discrete (empirical)
  - Histogram
  - Continuous (parametric or non-parametric) pdf

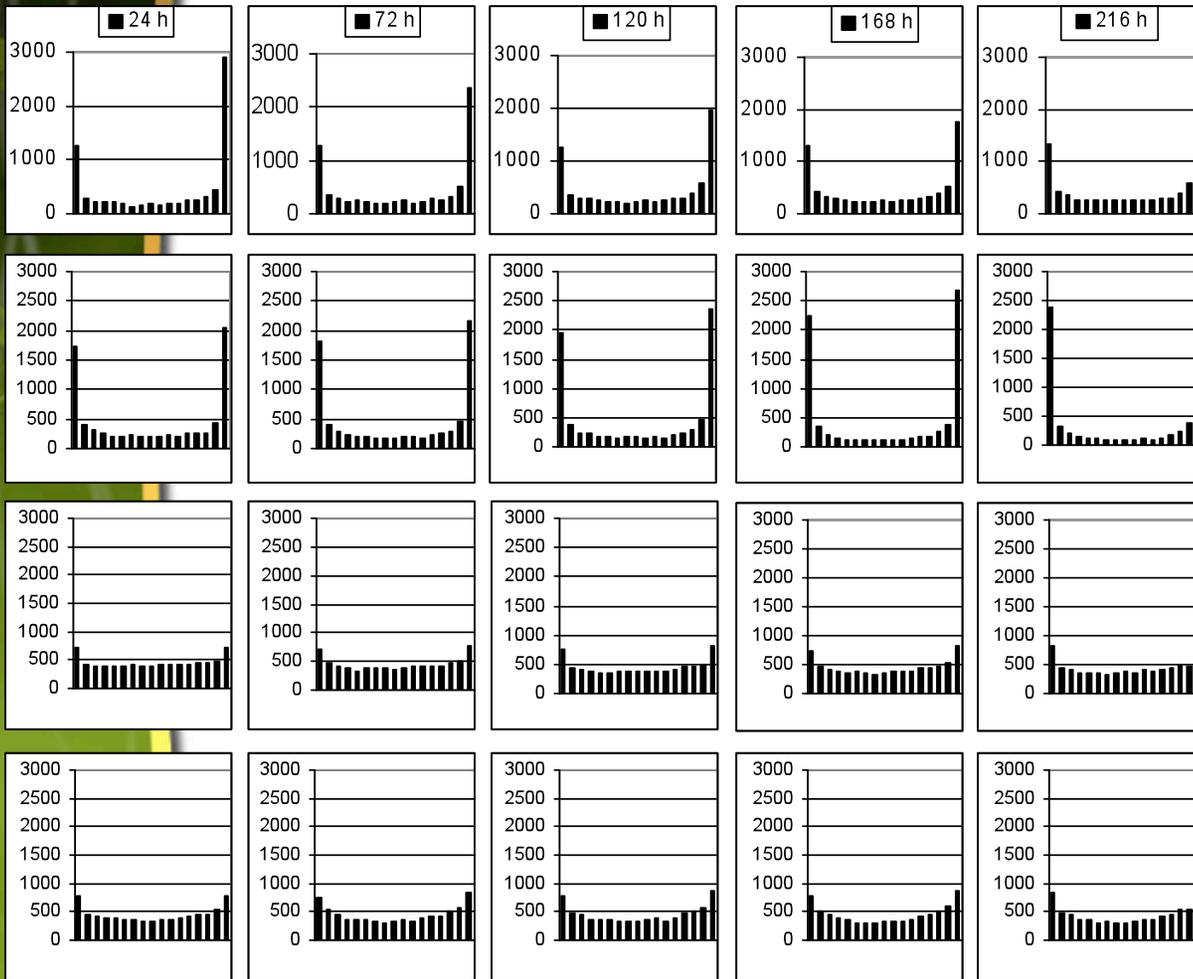


# *Survey of verification methods for ensembles*

- Evaluate the ensemble distribution
  - Rank Histogram
  - CRPS, CRPSS (Hersbach, 2000)
- 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”

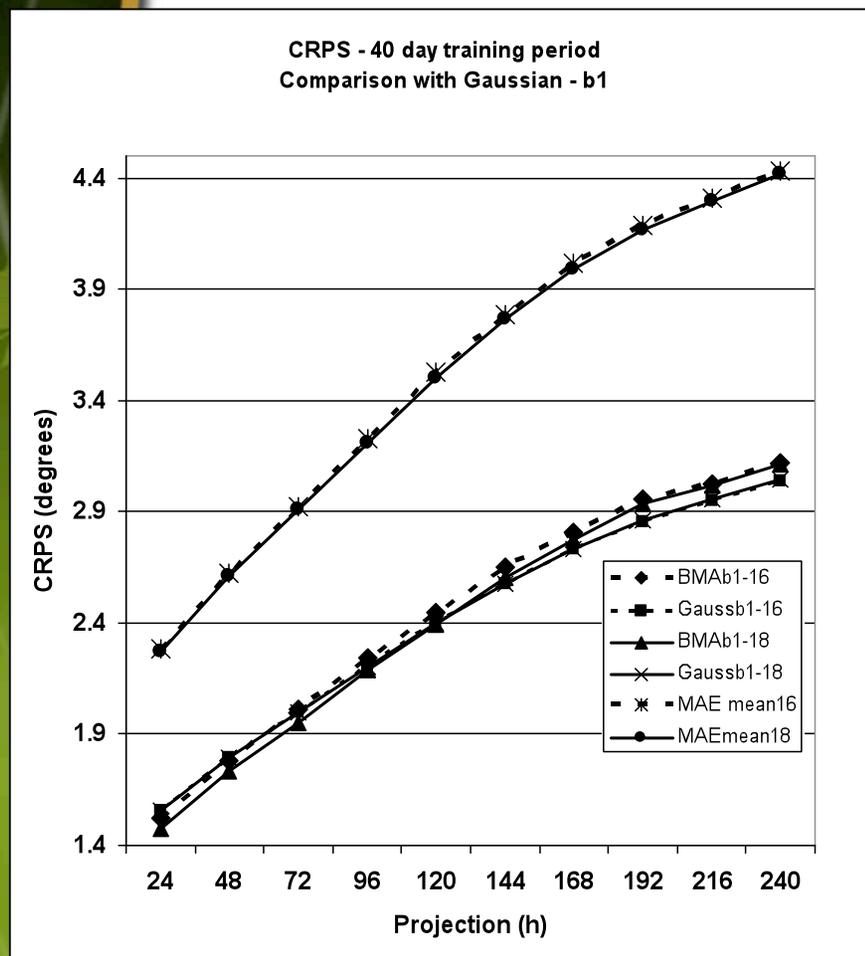


$$RMSD = \sqrt{\frac{1}{N+1} \sum_{k=1}^{N+1} \left( s_k - \frac{M}{N+1} \right)^2}$$

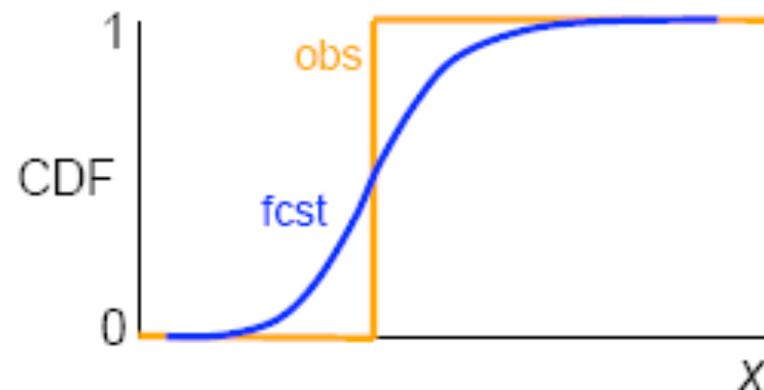
$$\sqrt{\frac{MN}{(N+1)^2}}$$



# Continuous Rank Probability Score



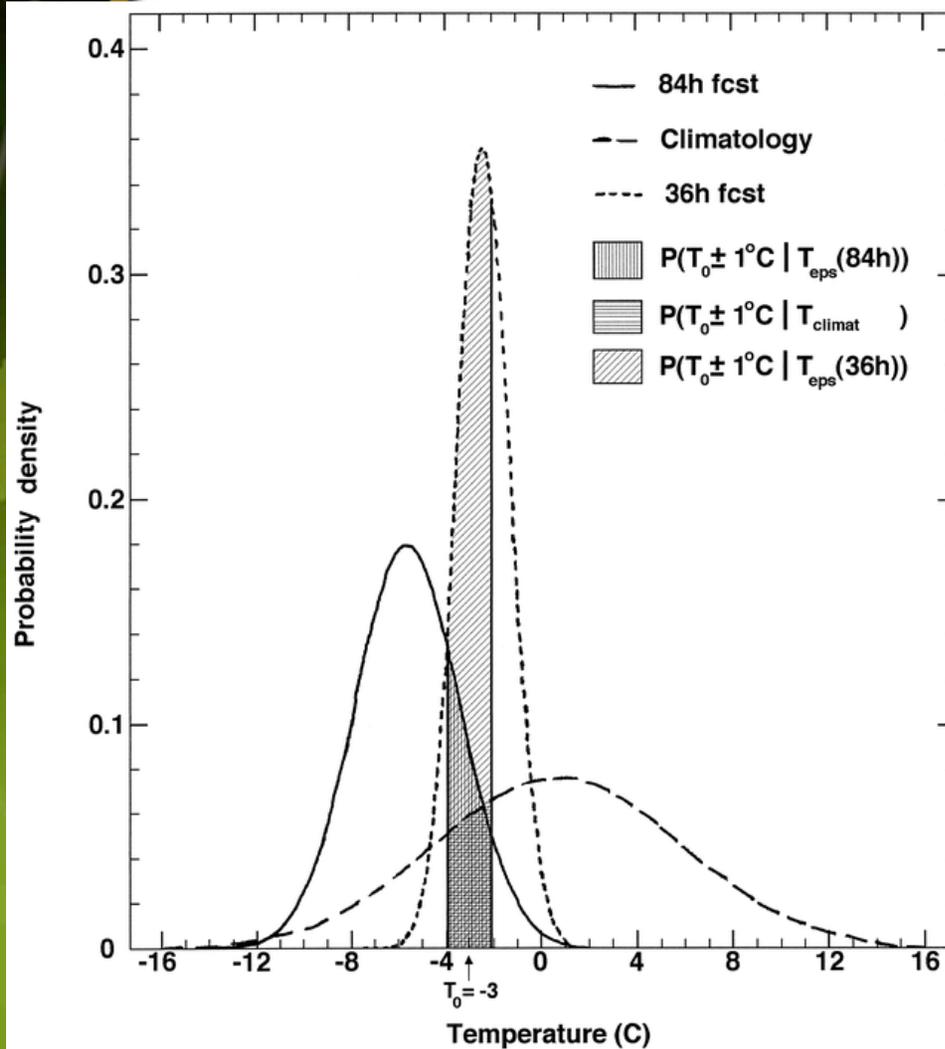
$$CRPS(P, x_a) = \int_{-\infty}^{\infty} [P(x) - P_a(x)]^2 dx$$



- difference between observation and forecast, expressed as cdfs
- defaults to MAE for deterministic fcst
- flexible, can accommodate uncertain obs



# 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



# Summary scores for event probability verification

- Measure different attributes of the forecasts
  - Brier score (accuracy)
  - Brier Skill score (skill) (climatology or persistence)
  - Reliability Tables (reliability, resolution, uncertainty)
  - Relative operating characteristic (ROC) (discrimination)
    - Issues of computation: should include binormal fit and S. Mason's method, both of which are or will be in R.
- **The last two represent a reasonably complete coverage of the characteristics of probability forecasts for dichotomous events**
  - Rank probability score (accuracy) and RPSS (skill) are corresponding measures for multiple categories



# *Factors affecting ensemble verification development*

- User-oriented verification
  - Other than modelers
  - Spatial methods
- “High Impact Weather”
  - Can ensembles give early warning of HIW events?
- Data volumes
  - Extra dimension
  - Sample sizes
- Globalization of operational forecasting
  - SWFDP; GIFS
  - NAEFS
  - TIGGE – multimodel ensembles



# *User-relevant verification: A basic model*

- Provide information that is relevant to a wide spectrum of users
  - Ex: Multiple (user-selectable) thresholds
- Utilize diagnostic techniques
  - Ex: Distributions of statistics rather than (or in addition to) summary scores
- Provide uncertainty information about verification measures (e.g., confidence intervals)
- Ideally – strong interaction with users
  - Understand applications of verification information
  - Requires engagement of social science community and understanding of communication and decision-making aspects
- At the very least: Know your target user, and what his/her needs are

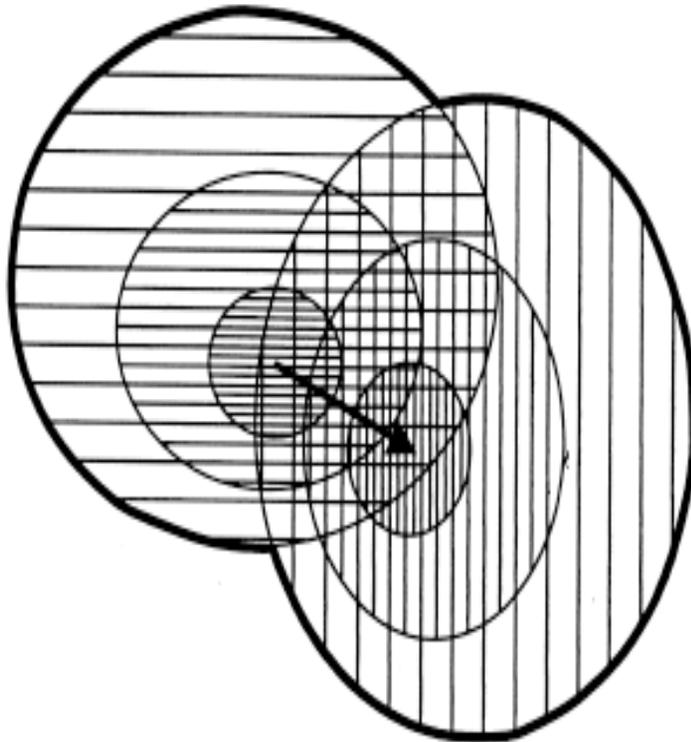


# *Some ideas on new verification methods for ensembles*

- Approach to verification of extremes
  - Verify quantiles of distribution, pooled over many locations, as far out to the tails of the distribution as the data will allow.
- Studies using spatial methods
  - Ebert – application of CRA technique to ensemble forecasts. So far, only ECMWF.
  - Application of Wilks minimum spanning tree – rank histogram for TC centers. (idea stage)



# *CRA method (Ebert and McBride 2000)*



Partition MSE for contiguous areas of e.g. precip into 3 components:

- displacement error
- volume error
- shape error

Quite widely used for deterministic verification

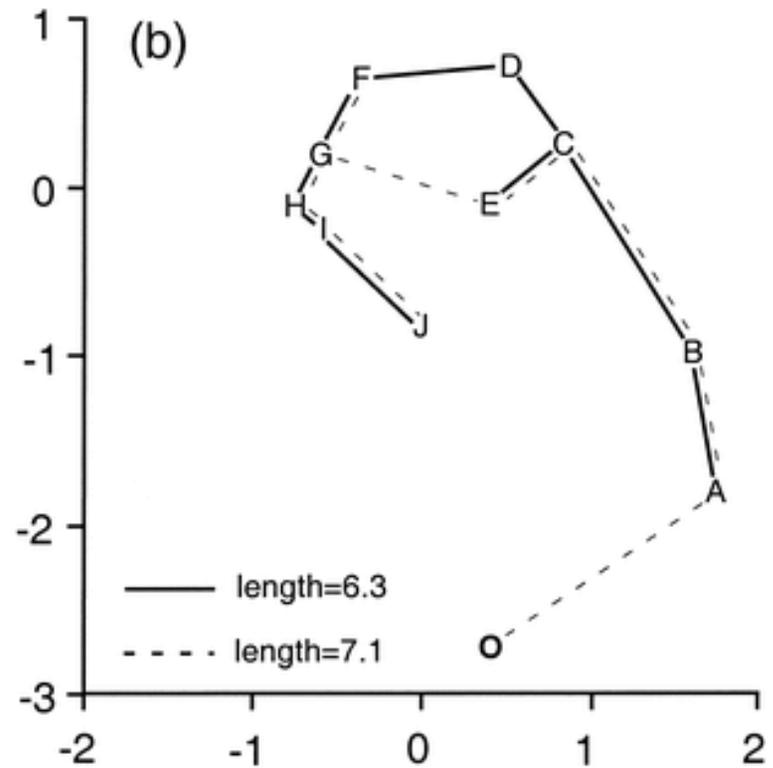
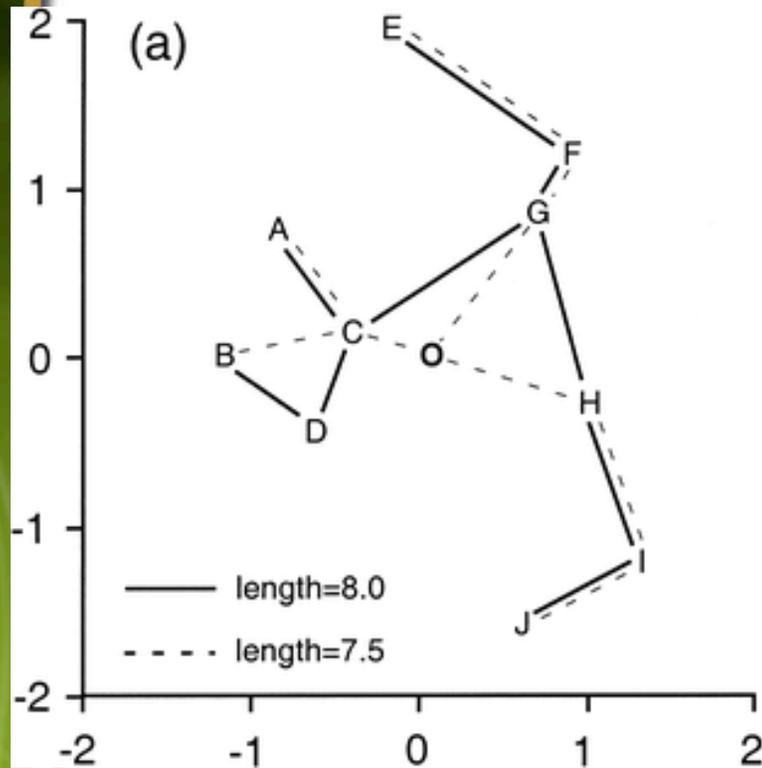


## *Spatial verification methods for ensembles – some thoughts*

- Verification of centers (hurricanes, low centers, precip maxima etc)
  - Ideal application of the minimum spanning tree (Smith 2001, Wilks 2004)
  - Evaluates whether the ensemble “contains” the observation spatially in 2-d
- Other ideas? Bill Gallus?
  - How to match spatial (deterministic obs) with field of probability forecasts?



# Minimum Spanning Tree – MST (Wilks 04)



# Summary

- Importance of taking user requirements into account in the design of a verification system
- System for ensembles should include the capability to fit distributions with a non-parametric method (especially for extreme “high impact weather” verification, also for ROC
- Spatial verification using MST; CRA can be extended to ensembles.
- If build a system only for summary verification, need use CRPS, CRPSS, along with standard probability scores.



www.ec.



# Thank you!



Environment  
Canada

Environnement  
Canada

September 2, 2009

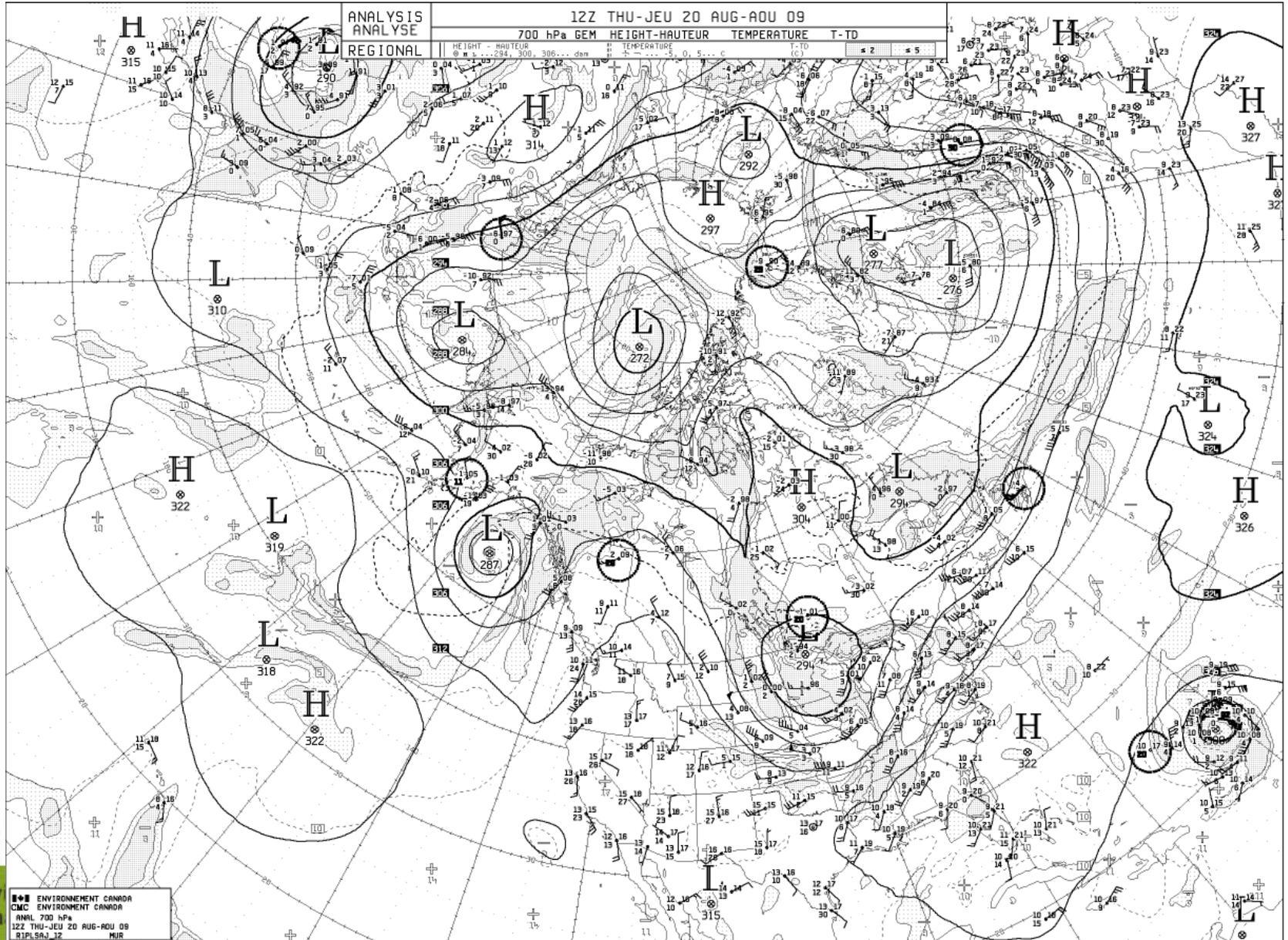
Canada  
Canada

# *Thoughts on the use of satellite data assimilation products for verification*

- Data assimilation systems are designed to blend models and data.
  - It is the moisture field that is of interest in model initialization, not clouds per se
  - Satellite data causes the rejection of in situ data because it is smoother than in situ data and thus more palatable to the model.
  - Use of the model as trial field gives the analysis the statistical “flavour” of the model being verified, and filters the obs dataset by comparison with a short range model prediction.
- Verification systems have rather different requirements, especially for users other than modelers
  - Independence of model and observations, both in qc and analysis phases.
  - Predictand to be verified isn’t necessarily q or RH, but total cloud usually.



# Example



# *Some suggestions for the use of satellite obs*

- What can we do to properly use satellite data for verification?
  - Can the rejection criteria and weights of satellite vs radiosondes be changed to be more meaningful for user-oriented verification?
  - Can visible and IR data be used together to give digital estimates of total cloud, in concert with sfc and radiosonde cloud obs?
    - For Total Cloud amount, which is where sat obs are best.
  - If expressed on a grid, then the grid length should be determined by the highest (spatial) resolution data available.
  - No model output should be used in the analysis process.
  - THIS requires an analysis system that is rather separate from model-based assimilation systems, but could use similar processing methods such as 4-d Var.



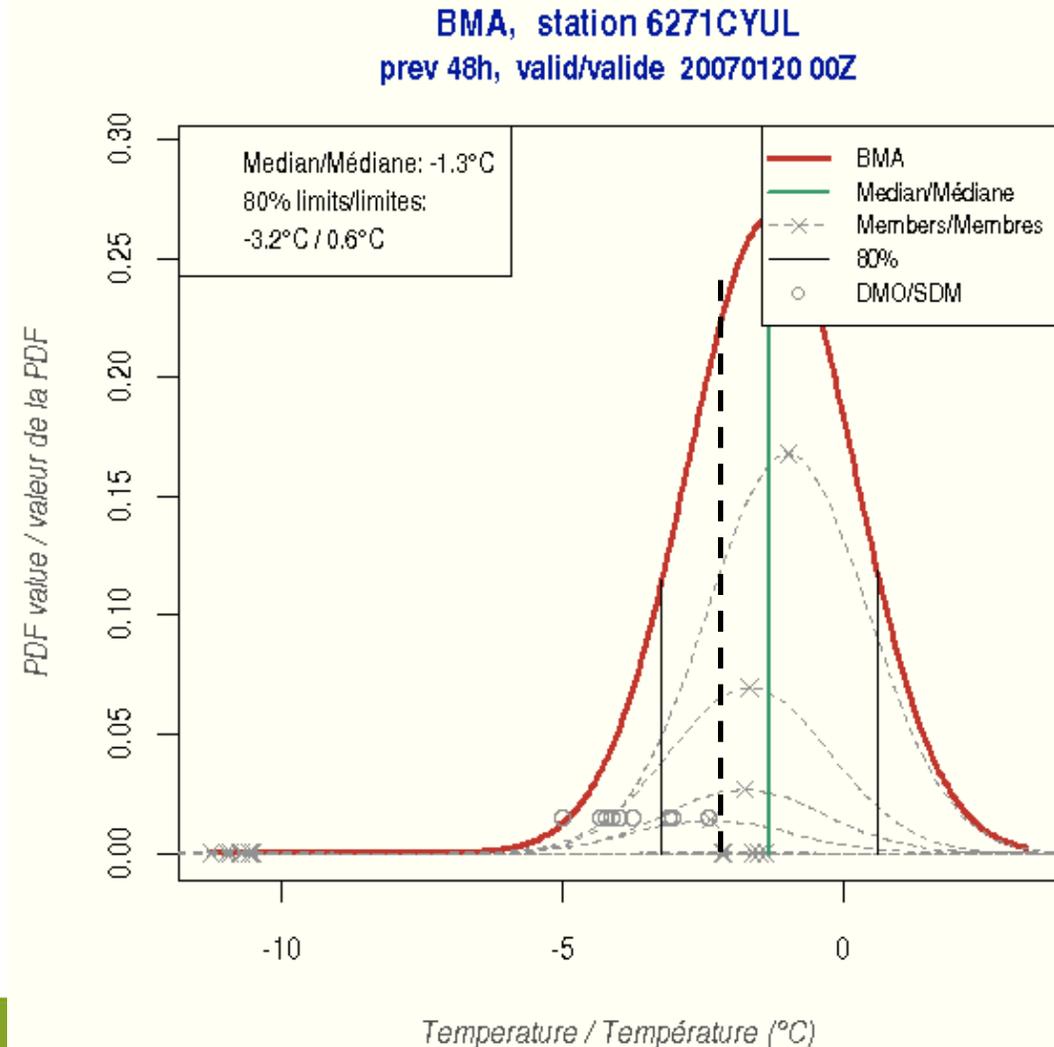
# Ensemble Verification

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  $f$ ”



# Properness study

