PROBABILISTIC FORECASTING:

WHY DO WE NEED IT?

(DON'T WE WANT TO KNOW EXACTLY THE FUTURE WEATHER?)

Zoltan Toth

Environmental Modeling Center
NOAA/NWS/NCEP

Ackn.: Yuejian Zhu and Olivier Talagrand [(1)]

[(1)] : Ecole Normale Superior and LMD, Paris, France

http://wwwt.emc.ncep.noaa.gov/gmb/ens/index.html

OUTLINE

- **WHY ARE WEATHER FORECASTS UNCERTAIN?**
  - Isn’t the atmosphere deterministic?
- **WHY DO USERS NEED TO KNOW ABOUT FORECAST UNCERTAINTY?**
  - They want to know, and not guess, about future weather?
- **TWO MAIN ATTRIBUTES OF FORECAST SYSTEMS**
- **MAIN TYPES OF FORECAST METHODS**
- **ADVANTAGES OF ENSEMBLE FORECASTING**

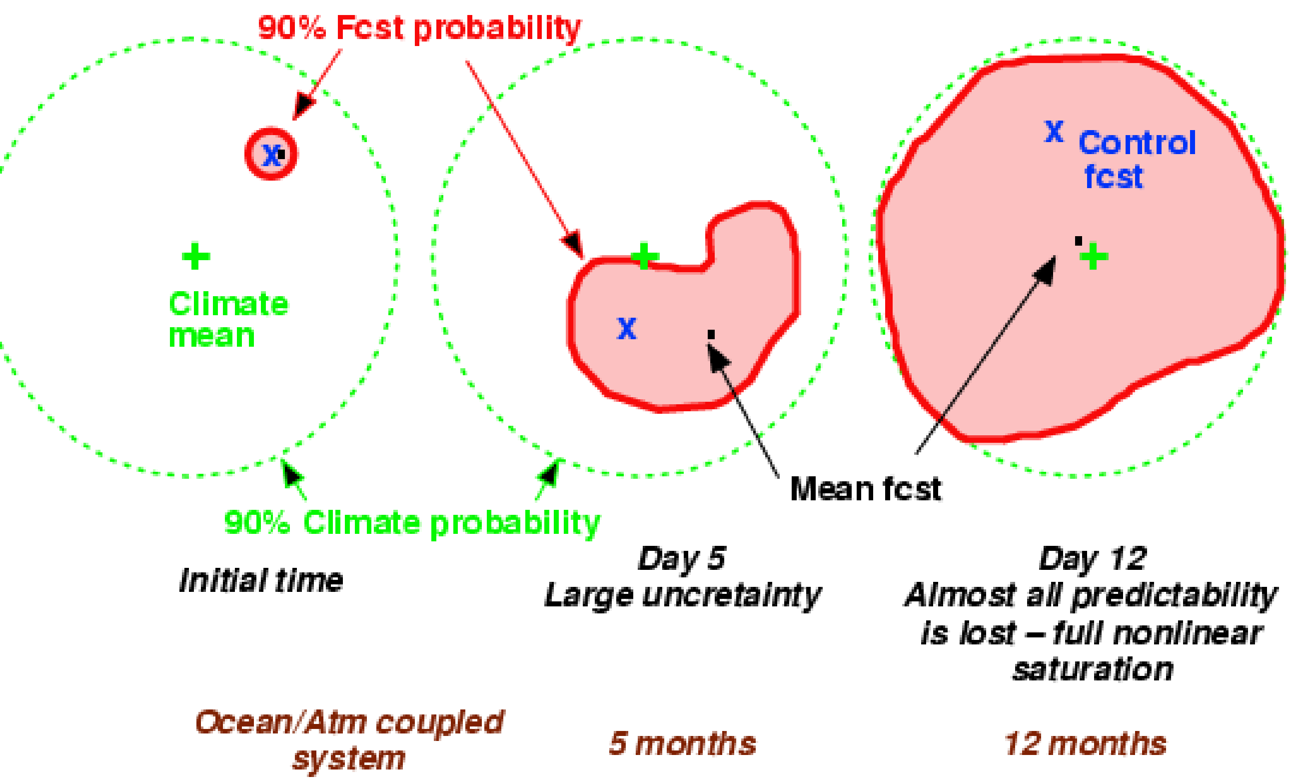
SCIENTIFIC BACKGROUND: WEATHER FORECASTS ARE UNCERTAIN

ORIGIN OF FORECAST UNCERTAINTY

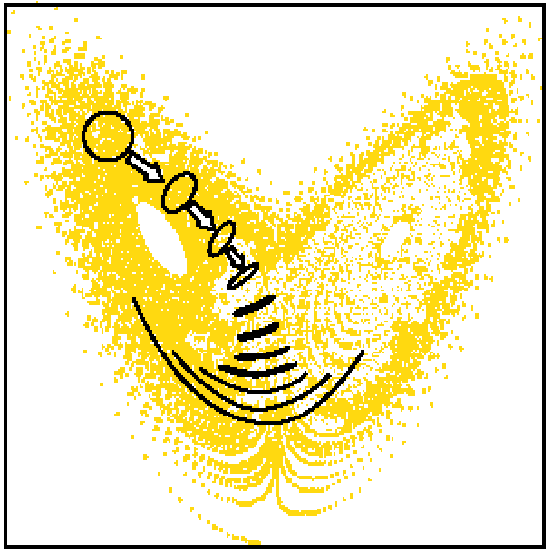
1) The atmosphere is a **deterministic system** *AND* has at least one direction in which **perturbations grow**

2) **Initial** state (and model) has **error** in it ==>

Chaotic system + Initial error =(Loss of) Predictability



Buizza 2002



USER REQUIREMENTS: PROBABILISTIC FORECAST INFORMATION IS CRITICAL

ECONOMIC VALUE OF FORECASTS

Given a particular forecast, a user either does or does not take action (eg, protects its crop against frost) *Mylne & Harrison, 1999*

OBSERVATION	FORECAST: *YES*	FORECAST: *NO*
YES	H(its) ***Mitigated Loss***	M(isses) ***Loss***
NO	F(alse alarms) ***Cost***	C(orrect rejections) ***No Cost***

$$\text{Mean Expense}_{fc} = hML + mL + fC$$

$$\text{Mean Expense}_{perf} = oML$$

$$ME_{cl} = \min[oL,\ oML + (1-o)C]$$

$$\text{Value} = \frac{ME_{cl} - ME_{fc}}{ME_{cl} - ME_{perf}}$$

o=climatological frequency

Optimum decision criterion for user action: P(weather event)=C/L

(Murphy 1977)

EVALUATION OF FORECAST SYSTEMS

Some statistics based on forecast system only

Other statistics based on comparison of forecast and observed systems =>

FORECAST SYSTEM ATTRIBUTES

- Abstract concepts (like length)
  - **Reliability and Resolution**
  - Both can be measured through different statistics
- Statistical properties
  - Interpreted for large set of forecasts (ie, describe behavior of forecast system), not for a single forecast
- For their definition
  - Assume that forecasts:
    - Can be of any format
    - Take a finite number of different "classes"
  - Consider empirical frequency distribution of
    - Verifying observations corresponding to large number of forecasts of same class => Observed Frequency Distribution (ofd)

STATISTICAL RELIABILITY

STATISTICAL CONSISTENCY OF FORECASTS WITH OBSERVATIONS

BACKGROUND:

- Consider particular forecast class – F_a
- Consider distribution of observations O_a that follow forecasts from F_a

DEFINITION:

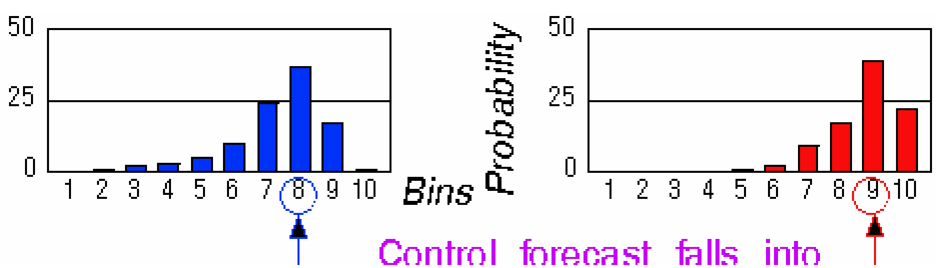
- If forecast F_a has the exact same form as O_a, for all forecast classes,
  the forecast system is statistically consistent with observations =>
  The forecast system is perfectly reliable

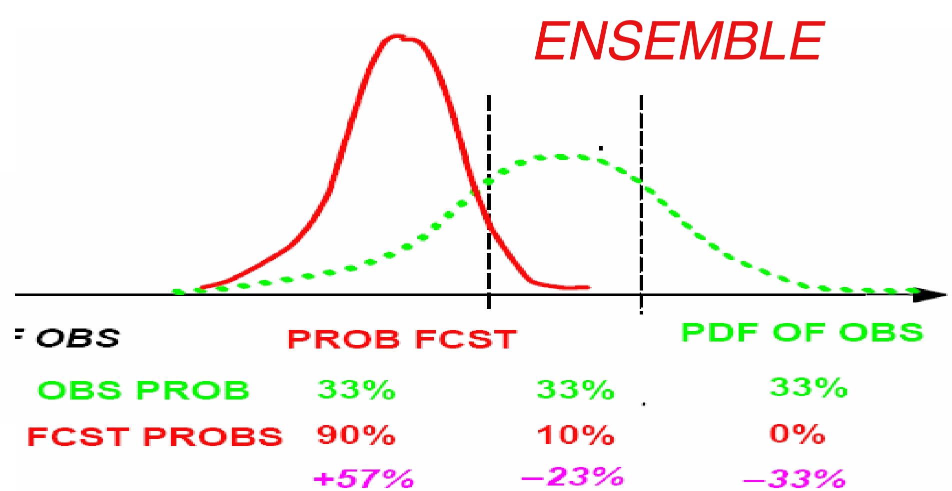
MEASURES OF RELIABILITY:

- Based on different ways of comparing F_a and O_a

EXAMPLES:

CONTROL FCST





STATISTICAL RESOLUTION

ABILITY TO DISTINGUISH, AHEAD OF TIME, AMONG DIFFERENT OUTCOMES

BACKGROUND:

- Assume observed events are classified into finite number of classes

DEFINITION:

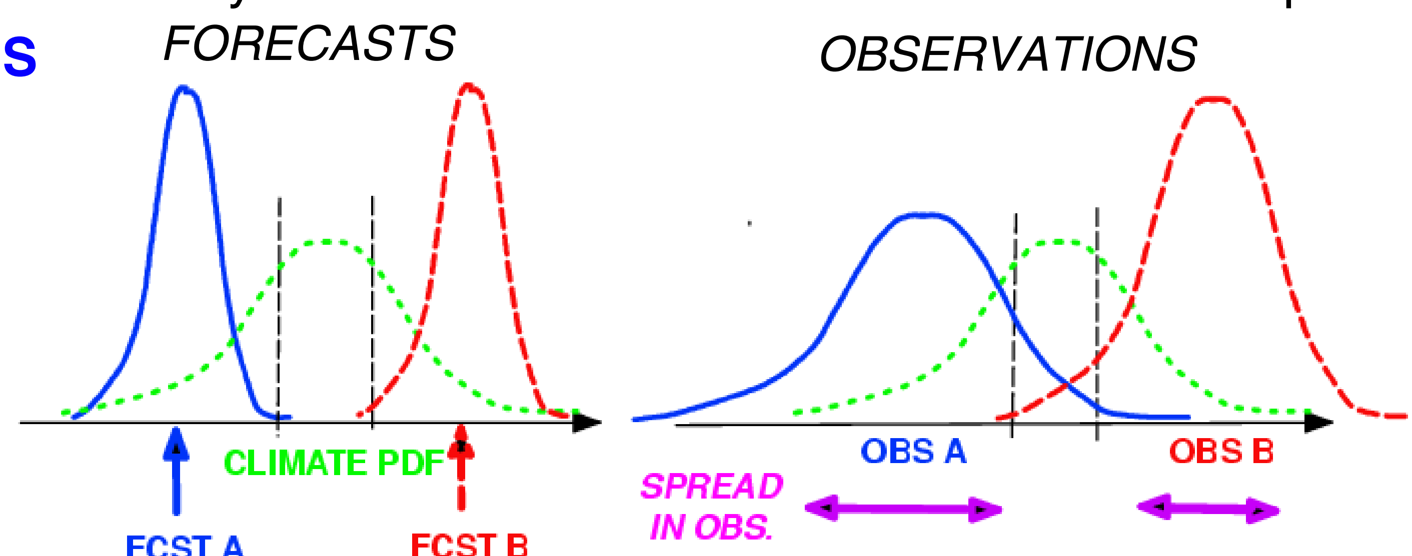
- If all observed classes are preceded by distinctly different forecasts, the forecasts "resolve" the problem =>

The forecast system has perfect resolution

MEASURES OF RELIABILITY:

- Based on degree of separation of distributions of observations that follow various forecast classes
- Measured by difference between ofd's & climate distribution
- Measures differ by how differences between distributions are quantified

EXAMPLES



CHARACTERISTICS OF FORECAST SYSTEM ATTRIBUTES

- **Reliability & resolution are general forecast attributes**
  - Valid for any forecast format (single, categorical, probabilistic, etc)
- **Reliability**
  - **Can be statistically imposed at one time level**
    - If both natural & forecast systems are stationary in time, and
    - If there is a large enough set of observed-forecast pairs
      - Replace forecast by corresponding observed frequency distribution
  - Not related to time evolution of forecast/observed systems
- **Resolution reflects inherent value of forecast system**
  - Can be improved only through more knowledge about time evolution
  - Statistical consistency at one time level (reliability) is irrelevant
- **Reliability & resolution are independent attributes**
  - Climate pdf fcst is perfectly reliable, yet has no resolution
  - Reversed rain /no-rain fcst can have perfect resolution and no reliability
- **Perfect reliability and perfect resolution = perfect fcst system**
  - "Deterministic" forecast system that is always correct
- **Utility of forecast systems**
  - **Need both reliability and resolution**
    - Especially if no observed/forecast pairs available (eg, extreme forecasts, etc)

FORECAST SYSTEMS

- **Empirical**
  - Based on record of observations =>
    - Possibly very good reliability
    - Will fail in “new” (not yet observed) situations (eg., climate trend, etc)
  - Resolution (forecast skill) depends on length of observations
    - Useful for now-casting, climate applications
    - Not practical for typical weather forecasting

- **Theoretical**
  - Based on general scientific principles
    - Incomplete/approximate knowledge =>
      - Prone to statistical inconsistency
  - Run-of-the-mill cases *can be statistically calibrated* to insure reliability
  - For rare/extreme event fcsts, *statistical consistency must be improved*
  - Predictability limited by
    - Gaps in knowledge about system
    - Errors in initial state of system

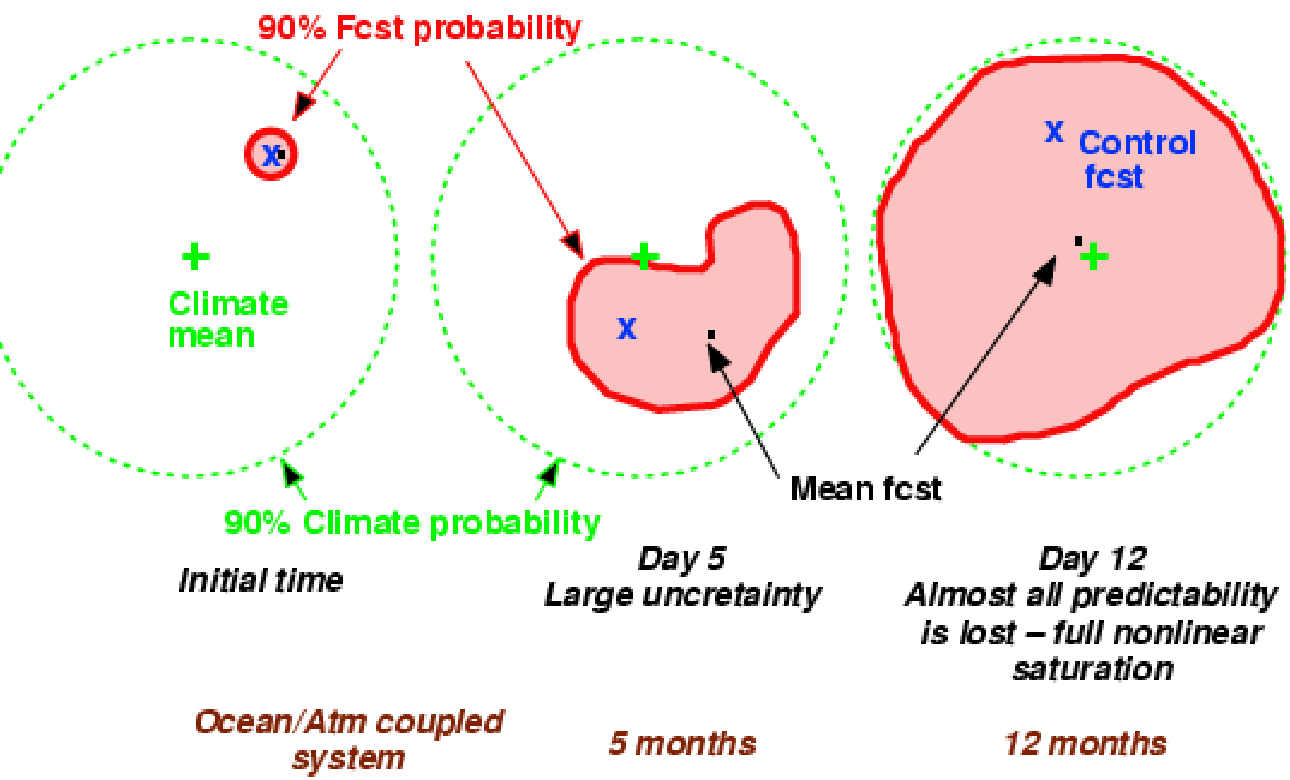
SCIENTIFIC BACKGROUND: WEATHER FORECASTS ARE UNCERTAIN

ORIGIN OF FORECAST UNCERTAINTY

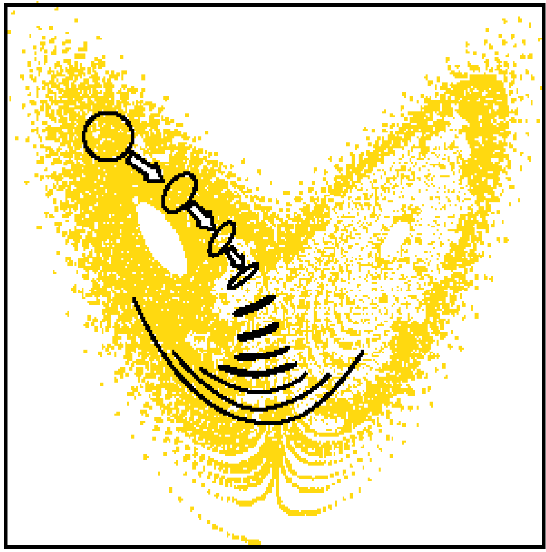
1) The atmosphere is a **deterministic system** *AND* has at least one direction in which **perturbations grow**

2) **Initial** state (and model) has **error** in it ==>

Chaotic system + Initial error =(Loss of) Predictability



Buizza 2002

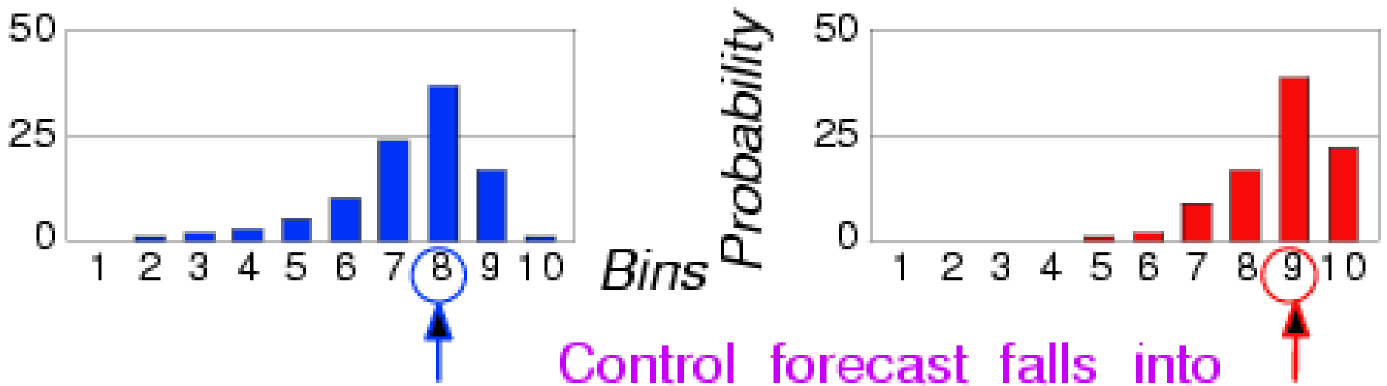


FORECASTING IN A CHAOTIC ENVIRONMENT –

PROBABILISTIC FORECASTING BASED A ON SINGLE FORECAST –

One integration with an NWP model, **combined with past verification statistics**

DETERMINISTIC APPROACH - PROBABILISTIC FORMAT



•Does not contain all forecast information

•Not best estimate for future evolution of system

•***UNCERTAINTY CAPTURED IN TIME AVERAGE SENSE -***

•**NO ESTIMATE OF CASE DEPENDENT VARIATIONS IN FCST UNCERTAINTY**

FORECASTING IN A CHAOTIC ENVIRONMENT - 2

DETERMINISTIC APPROACH - PROBABILISTIC FORMAT

PROBABILISTIC FORECASTING -

Based on Liuville Equations

Continuity equation for probabilities, given dynamical eqs. of motion

- Initialize with probability distribution function (pdf) at analysis time
- Dynamical forecast of pdf based on conservation of probability values
- **Prohibitively expensive** -
    - Very high dimensional problem (state space x probability space)
    - Separate integration for each lead time
    - Closure problems when simplified solution sought

FORECASTING IN A CHAOTIC ENVIRONMENT - 3

DETERMINISTIC APPROACH - PROBABILISTIC FORMAT

MONTE CARLO APPROACH – *ENSEMBLE FORECASTING*

- ***IDEA*: Sample sources of forecast error**
  - Generate initial ensemble perturbations
  - Represent model related uncertainty

- ***PRACTICE*: Run multiple NWP model integrations**
  - Advantage of perfect parallelization
  - Use lower spatial resolution if short on resources

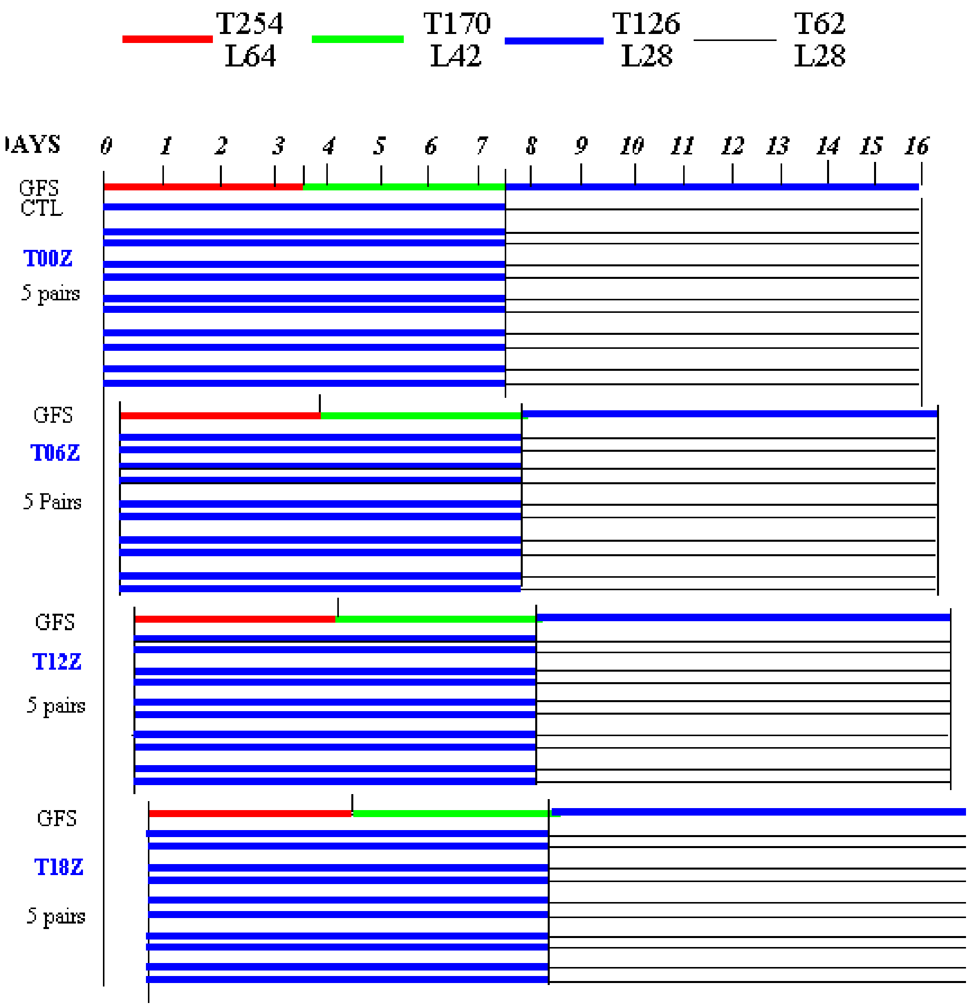
- ***USAGE*: Construct forecast pdf based on finite sample**
  - Ready to be used in real world applications
  - Verification of forecasts
  - Statistical post-processing (remove bias in 1^{st}, 2^{nd}, higher moments)

CAPTURES FLOW DEPENDENT VARIATIONS

IN FORECAST UNCERTAINTY

NCEP GLOBAL ENSEMBLE FORECAST SYSTEM

MARCH 2004 CONFIGURATION



MOTIVATION FOR ENSEMBLE FORECASTING

- ***FORECASTS ARE NOT PERFECT* - IMPLICATIONS FOR:**
  - **USERS:**
    - Need to know how often / by how much forecasts fail
    - Economically optimal behavior depends on
      - Forecast error characteristics
      - User specific application
        - » Cost of weather related adaptive action
        - » Expected loss if no action taken
      - EXAMPLE: Protect or not your crop against possible frost

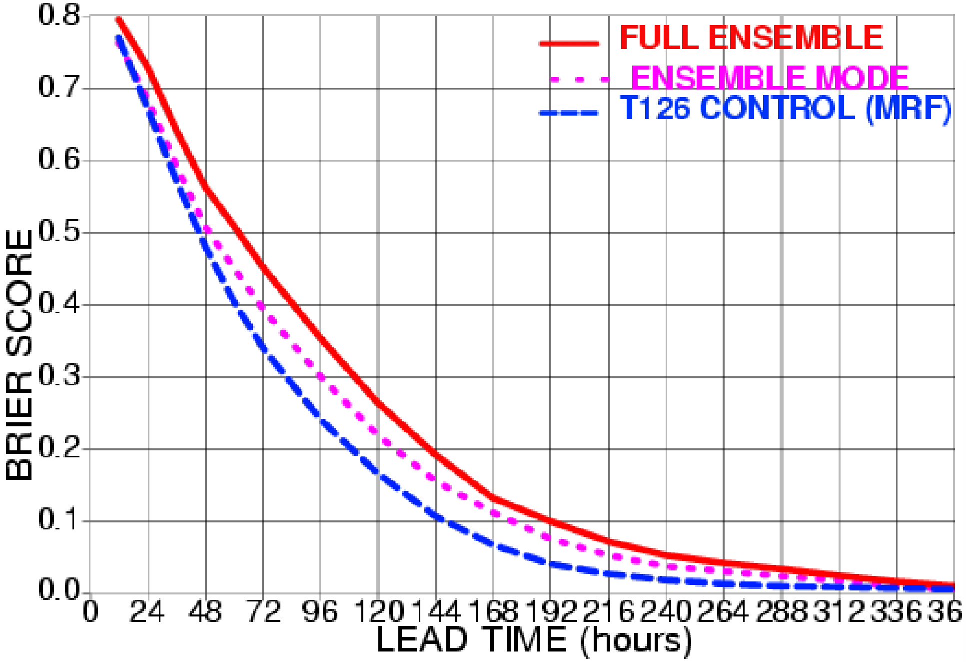
    Cost = 10k, Potential Loss = 100k => Will protect if P(frost) > Cost/Loss=0.1

    - ***NEED FOR PROBABILISTIC FORECAST INFORMATION***

  - **DEVELOPERS:**
    - Need to improve performance - *Reduce error in estimate of first moment*
      - Traditional NWP activities (I.e., model, data assimilation development)
    - Need to account for uncertainty - *Estimate higher moments*
      - New aspect – How to do this?
    - Forecast is incomplete without information on forecast uncertainty
    - ***NEED TO USE PROBABILISTIC FORECAST FORMAT***

BRIER SKILL SCORE

COMBINED MEASURE OF RELIABILITY AND RESOLUTION



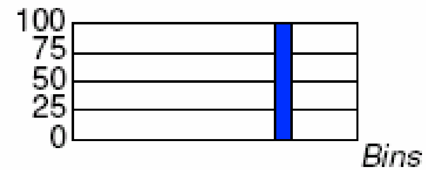
*Brier Skill Score for the **NH extratropics**, for March–May 1997. Forecasts are made for 10 climatologically equally likely bins; results shown here are the average for the two extreme bins. The bin where the control or ensemble mode falls is assigned a probability corresponding to the observed frequency of the verifying analysis falling into the same bin (P), while the remaining 9 bins are assigned (1–P)/9 (assuming perfect reliability). Note that depending on the value of the mode (1≤M≤10), the corresponding observed frequency for the ensemble (but not for the control) varies widely.*

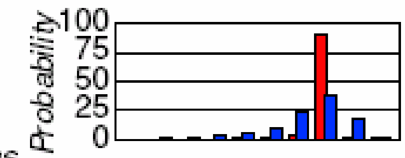
HOW TO COMPARE CONTROL VS ENSEMBLE?

1) **ACCURACY:** Compare control with ensemble mean

2) **PROBABILITIES:** In framework of 10 climate bins

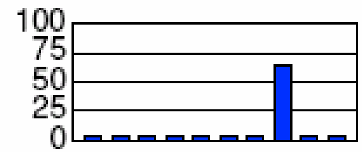
CONTROL	ENSEMBLE
Yes or No fcst for an event	Full probability distribution

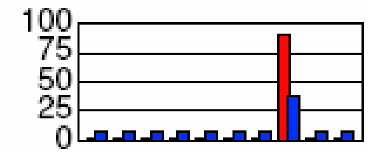




"UPGRADE" control:
Based on past verification, can be calibrated (like ens.)

"DOWNGRADE" ensemble:
Take probability at mode (P_m), distribute ($1-P_m$) over 9 other b





For control: Use average reliability when fcst falls/doesn't fall in a climate bin (FIXED VALUE)

For ensemble: Use average reliability for bin with most ensemble members (depends on how many fcsts fell in bin), distribute remaining probabilities equally among rest of bins

EQUAL FOOTING, FAIR COMPARISON

RESOLUTION OF ENSEMBLE BASED PROB. FCSTS

QUESTION:

What are the typical **variations in foreseeable forecast uncertainty?**

What variations in predictability can the ensemble resolve?

METHOD:

Ensemble mode value to distinguish high/low predictability cases

Stratify cases according to ensemble mode value –

Use 10–15% of cases when ensemble is highest/loewest

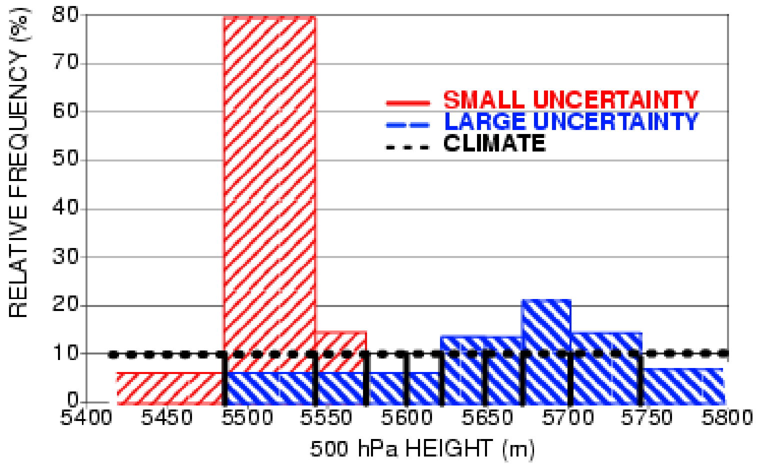
DATA:

NCEP **500 hPa NH extratropical ensemble fcsts** for March–May 1997

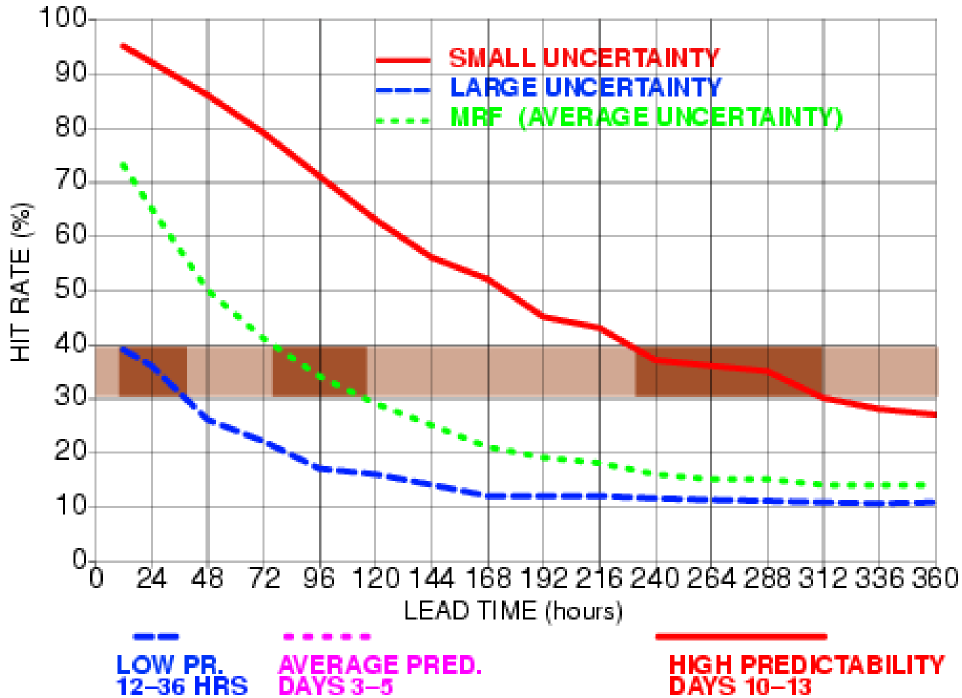
14 perturbed fcsts and high resolution control

VERIFICATION:

Hit rate for ensemble mode and hires control fcst



SEPARATING HIGH VS. LOW UNCERTAINTY FCSTS



*THE **UNCERTAINTY OF FCSTS** CAN BE **QUANTIFIED IN ADVANCE***

HIT RATES FOR 1–DAY FCSTS

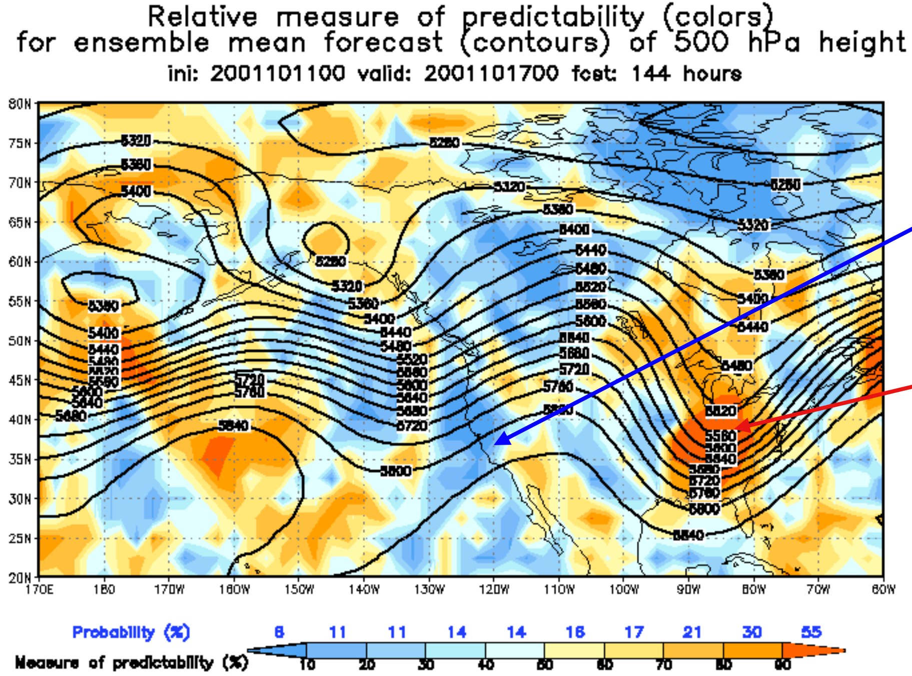
*CAN BE **AS LOW AS 36%**, OR **AS HIGH AS 92%***

10–15% OF THE TIME A 12–DAY FCST CAN BE AS GOOD, OR A 1–DAY FCST CAN BE AS POOR AS AN AVERAGE 4–DAY FCAST

1–2% OF ALL DAYS THE 12–DAY FCST CAN BE MADE WITH MORE CONFIDENCE THAN THE 1–DAY FCST

AVERAGE HIT RATE FOR EXTENDED–RANGE FCSTS IS LOW –
VALUE IS IN KNOWING WHEN FCST IS RELIABLE

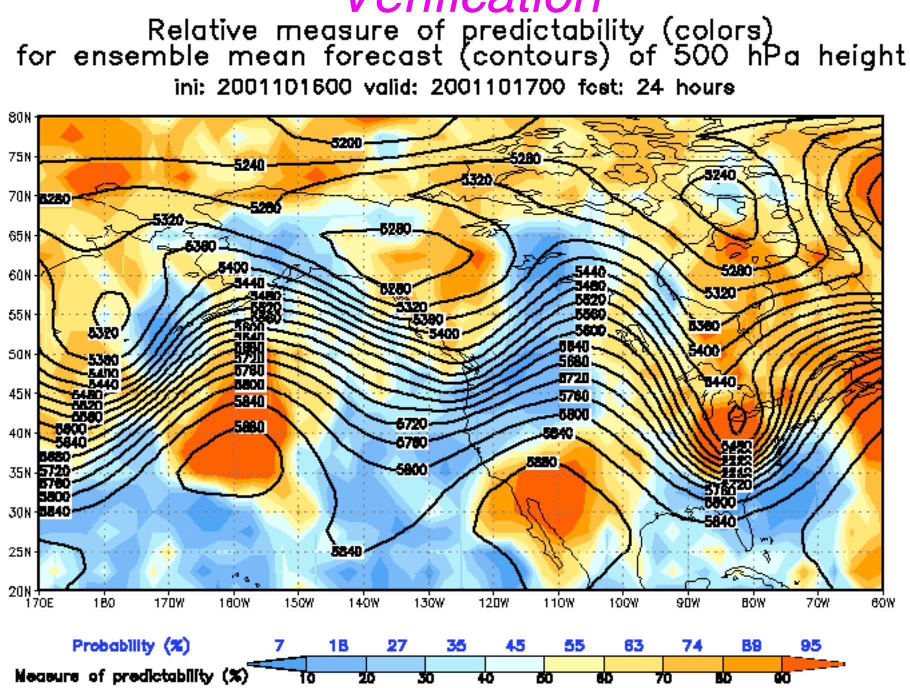
144 hr forecast

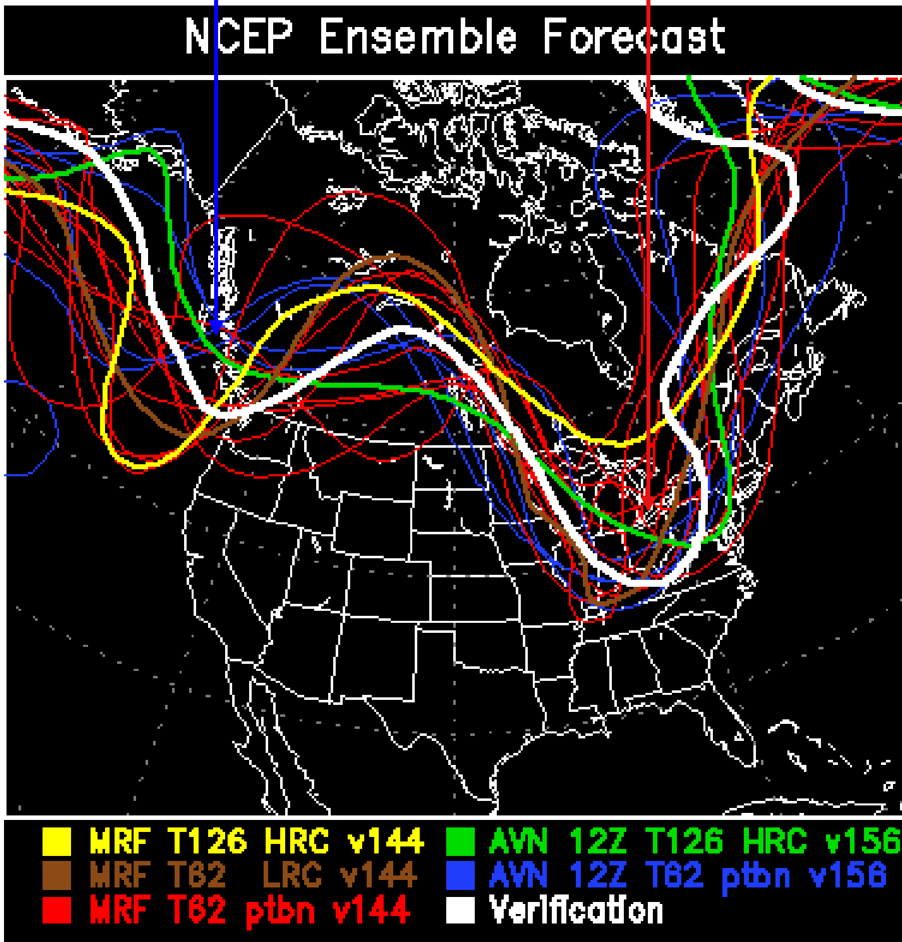


Poorly predictable large scale wave
Eastern Pacific – Western US

Highly predictable small scale wave
Eastern US

Verification





OUTLINE / SUMMARY

- **WHY DO WE NEED PROBABILISTIC FORECASTS?**
  - **Isn't the atmosphere deterministic?** **YES, but it's also CHAOTIC**

FORECASTER'S PERSPECTIVE	*USER'S PERSPECTIVE*
Ensemble techniques	**Probabilistic description**

- **WHAT ARE THE MAIN ATTRIBUTES OF FORECAST SYSTEMS?**
  - ***RELIABILITY*** Stat. consistency with distribution of corresponding observations
  - ***RESOLUTION*** Different events are preceded by different forecasts

- **WHAT ARE THE MAIN TYPES OF FORECAST METHODS?**
  - ***EMPIRICAL*** Good reliability, limited resolution (problems in "new" situations)
  - ***THEORETICAL*** Potentially high resolution, prone to inconsistency

- **ENSEMBLE METHODS**
  - **Only practical way of capturing fluctuations in forecast uncertainty** due to
    - Case dependent dynamics acting on errors in
      - Initial conditions
      - Forecast methods

REFERENCES

Stanski, H. R., L. J. Wilson, and W. R. Burrows, 1989: Survey of common verification methods in meteorology. WMO World Weather Watch Technical Reprot No. 8, WMO/TD. No. 358.

Talagrand, O., R. Vautard, and B. Strauss, 1998: Evaluation of probabilistic prediction systems. Proceedings of ECMWF Workshop on Predictability 20–22 October 1997, ECMWF, Shinfield Park, Reading, RG2 9AX, UK. 1–25.

Murphy, A. W., 1977: The value of climatological, categorical and probabilistic forecasts in the cost–loss ratio situation. Mon. Wea. Rev., 105, 803–816.

Mylne,K.R., 1999 The use of forecast value calculations for optimal decision making using probability forecasts. 17th Conf on Weather Analysis and Forecasting, 13–17 September 1999, Denver, Colorado, pp235–239.

Richardson, DS, 2000, "The application of cost–loss models to forecast verification", Proceedings of 7th ECMWF Workshop on Meteorological Systems. ECMWF, Shinfield Park, Reading, RG2 9AX, UK.

Toth, Z., and Kalnay, E., 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.*, 125, in print.

Toth, Z., E. Kalnay, S. Tracton, R. Wobus, and J. Irwin,1997: A synoptic evaluation of the NCEP ensemble. *Weather and Forecasting*, 12, 140–153.

Toth, Z., Y. Zhu, T. Marchok, . Tracton, and E. Kalnay, 1998: Verification of the NCEP global ensemble forecasts. Preprints of the 12th Conference on Numerical Weather Prediction, 11–16 January 1998, Phoenix, Arizona, 286–289.

Zhu,, Y., G. Iyengar, Z. Toth, M. S. Tracton, and T. Marchok, 1996: Objective evaluation of the NCEP global ensemble forecasting system. Preprints of the 15th AMS Conference on Weather Analysis and Forecasting, 19–23 August 1996, Norfolk, Virginia, p. J79–J82.

Zhu, Y., and Z. Toth, 1999: Calibration of Probabilistic Quantitative Precipitation Forecasts. Preprints of the 16th AMS Conference on Weather Analysis and Forecasting, 13–17 September 1999, Denver, CO, 214–215.

Wobus, R., Z. Toth, and Y. Zhu, 1999: An evaluation of probabilistic forecasts based on the NCEP global ensemble. Preprints of the 16th AMS Conference on Weather Analysis and Forecasting, 13–17 September 1999, Denver, CO, 212–213.

Zhu, Y., Toth, Z., E. Kalnay, and S. Tracton, 1998: Probabilistic quantitative precipitation forecasts based on the NCEP global ensemble. Preprints of the 12th Conference on Numerical Weather Prediction, 11–16 January 1998, Phoenix, Arizona, J8–J11.

Tracton, M. S. and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting*, 8, 379–398.

Wilks, D. S., 1995: Statistical methods in the atmospheric sciences. Academic Press. 467 pp.

Toth, Z., Y. Zhu, and T. Marchok, 2001: The ability of ensembles to distinguish between forecasts with small and large uncertainty. *Weather and Forecasting*, 16, 436–477.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble based weather forecasts. *Bull. Amer. Meteorol. Soc.*, 83, 73–83.

Toth, Z., O. Talagrand, and Y. Zhu, 2005: The Attributes of Forecast Systems: A Framework for the Evaluation and Calibration of Weather Forecasts. In: Predictability Seminars, 9-13 September 2002, Ed.: T. Palmer, ECMWF, in press.

Toth, Z., O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. In: Environmental Forecast Verification: A practitioner's guide in atmospheric science. Ed.: I. T. Jolliffe and D. B. Stephenson. Wiley, p. 137-164.

BACKGROUND

FORECAST PERFORMANCE MEASURES

COMMON CHARACTERISTIC: Function of both forecast and observed values

MEASURES OF RELIABILITY:

DESCRIPTION:

Statistically compares ***any sample of forecasts with sample of corresponding observations***

GOAL:

To assess similarity of samples (e.g., whether 1st and 2nd moments match)

EXAMPLES:

Reliability component of
- Brier Score
- Ranked Probability Score

Analysis Rank Histogram

Spread vs. Ens. Mean error

Etc.

MEASURES OF RESOLUTION:

DESCRIPTION:

Compares the ***distribution of observations that follows different classes of forecasts with the climate distribution***

GOAL:

To assess how well the observations are separated when grouped by different classes of preceding fcsts

EXAMPLES:

Resolution component of
- Brier Score
- Ranked Probability Score

Information content

Relative Operational Characteristics

Relative Economic Value

Etc.

COMBINED (REL+RES) MEASURES: Brier, Ranked Probab. Scores, rmse, PAC, etc

EXAMPLE – PROBABILISTIC FORECASTS

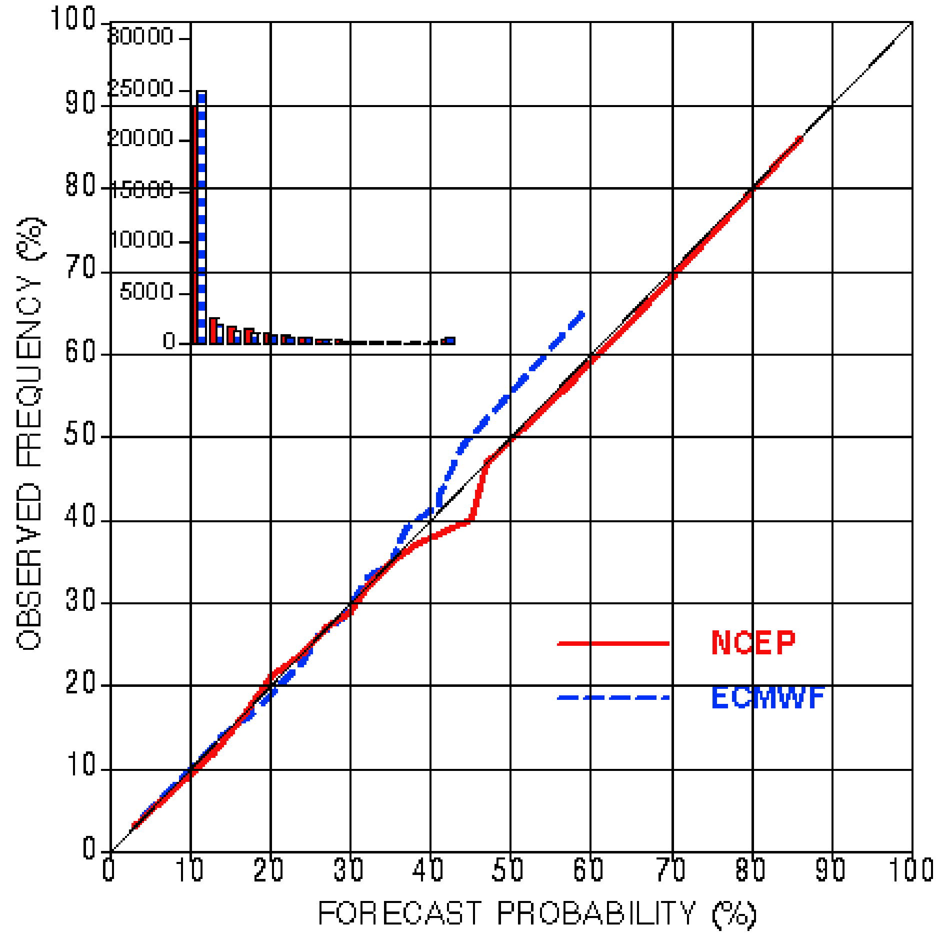
RELIABILITY:

Forecast probabilities for given event match observed frequencies of that event (with given prob. fcst)

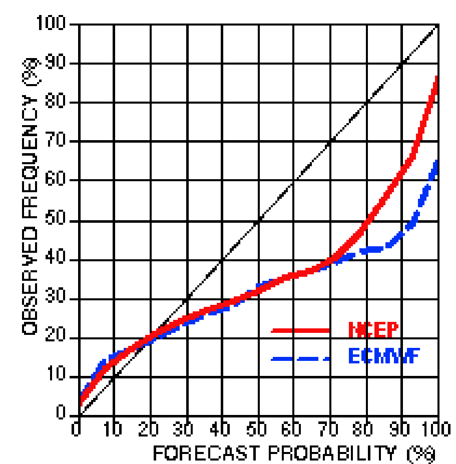
RESOLUTION:

Many forecasts fall into classes corresponding to high or low observed frequency of given event

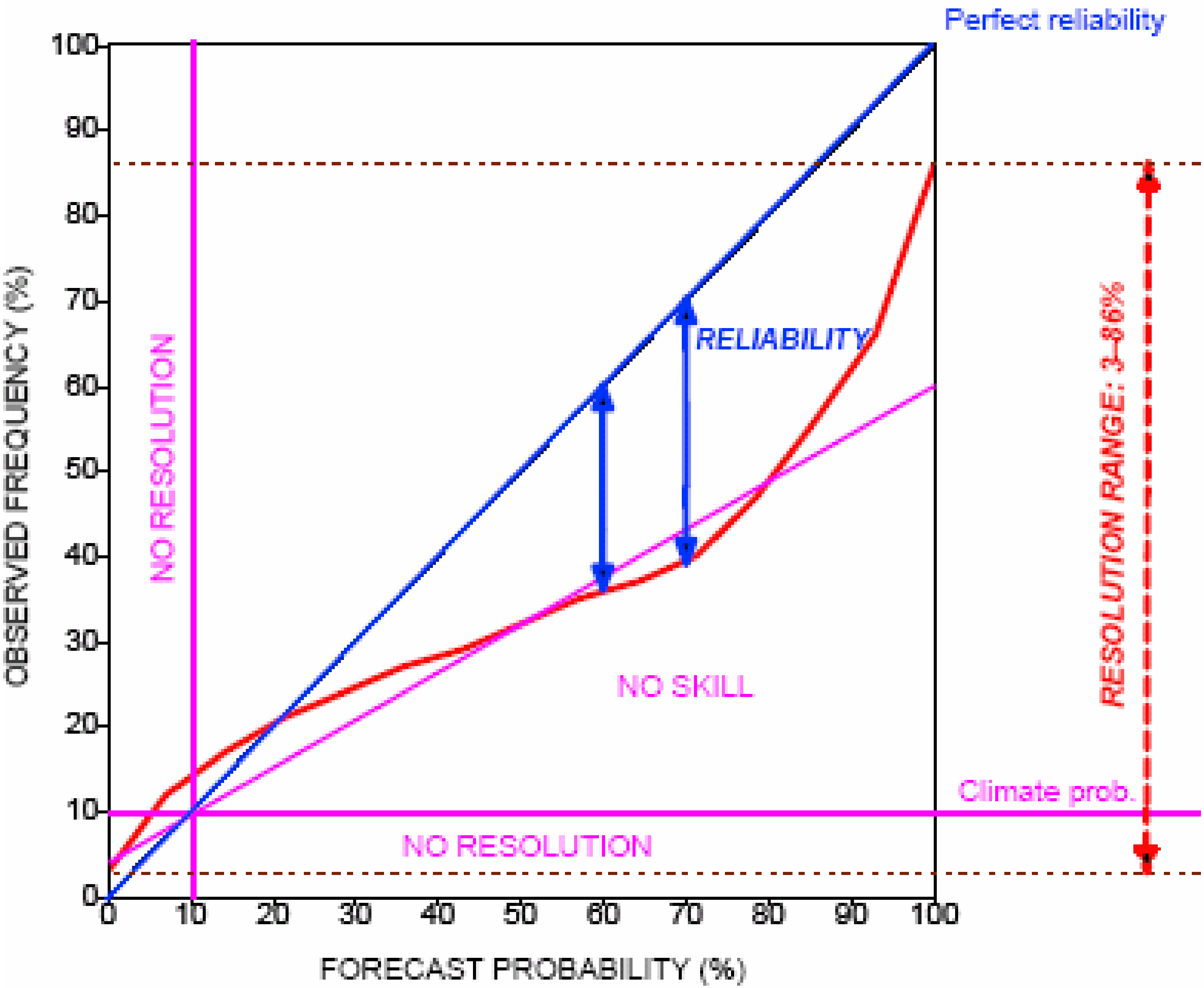
(Occurrence and non-occurrence of event is *well resolved* by fcst system)



Reliability diagram for 3-day lead time ensembles for January 1996. Forecast probabilities are based on observed frequencies associated with the same number of ensemble members falling in a particular bin during December 1-20, 1995. The diagram for uncalibrated forecasts is shown on the right.



RELIABILITY / ATTRIBUTES DIAGRAM



PROBABILISTIC FORECAST PERFORMANCE MEASURES

TO ASSESS TWO MAIN ATTRIBUTES OF PROBABILISTIC FORECASTS:

RELIABILITY AND RESOLUTION

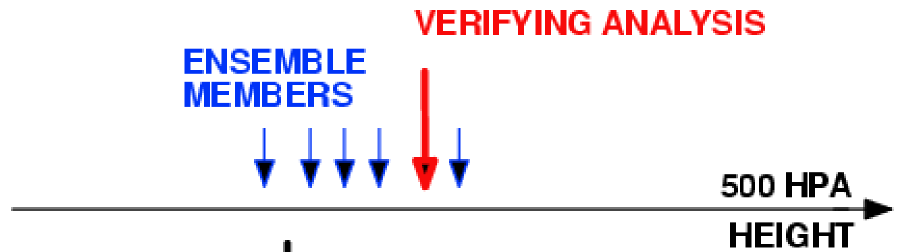
Univariate measures: Statistics accumulated point by point in space

Multivariate measures: Spatial covariance is considered

EXAMPLE:

BRIER SKILL SCORE (BSS)

COMBINED MEASURE OF RELIABILITY AND RESOLUTION



OBSERVATION d_i	0	1
FCST PROB p_i	20%	80%

Total of n pairs of cases

N_k cases with p_k probability

$$\bar{d}_k = \frac{1}{N_k} \sum_{i \in N_k} d_i$$

$$BS(p, d) = \frac{1}{n} \left[\sum_{i=1}^{n} (p_i - d_i)^2 \right]$$

$$BS = \underbrace{\frac{1}{n} \left[\sum_{k=1}^{K} N_k (p_k - \bar{d}_k)^2 \right]}_{\text{Reliability}} - \underbrace{\frac{1}{n} \left[\sum_{k=1}^{K} N_k (\bar{d}_k - \bar{d})^2 \right]}_{\text{Resolution}} + \underbrace{\bar{d}(1 - \bar{d})}_{\text{Uncertainty}}$$

$$BSS = 1 - \frac{BS\,(\text{forecast})}{BS\,(\text{climatology})}$$

BRIER SKILL SCORE (BSS)

COMBINED MEASURE OF RELIABILITY AND RESOLUTION

METHOD:

Compares pdf against analysis

- Resolution (random error)
- Reliability (systematic error)

EVALUATION

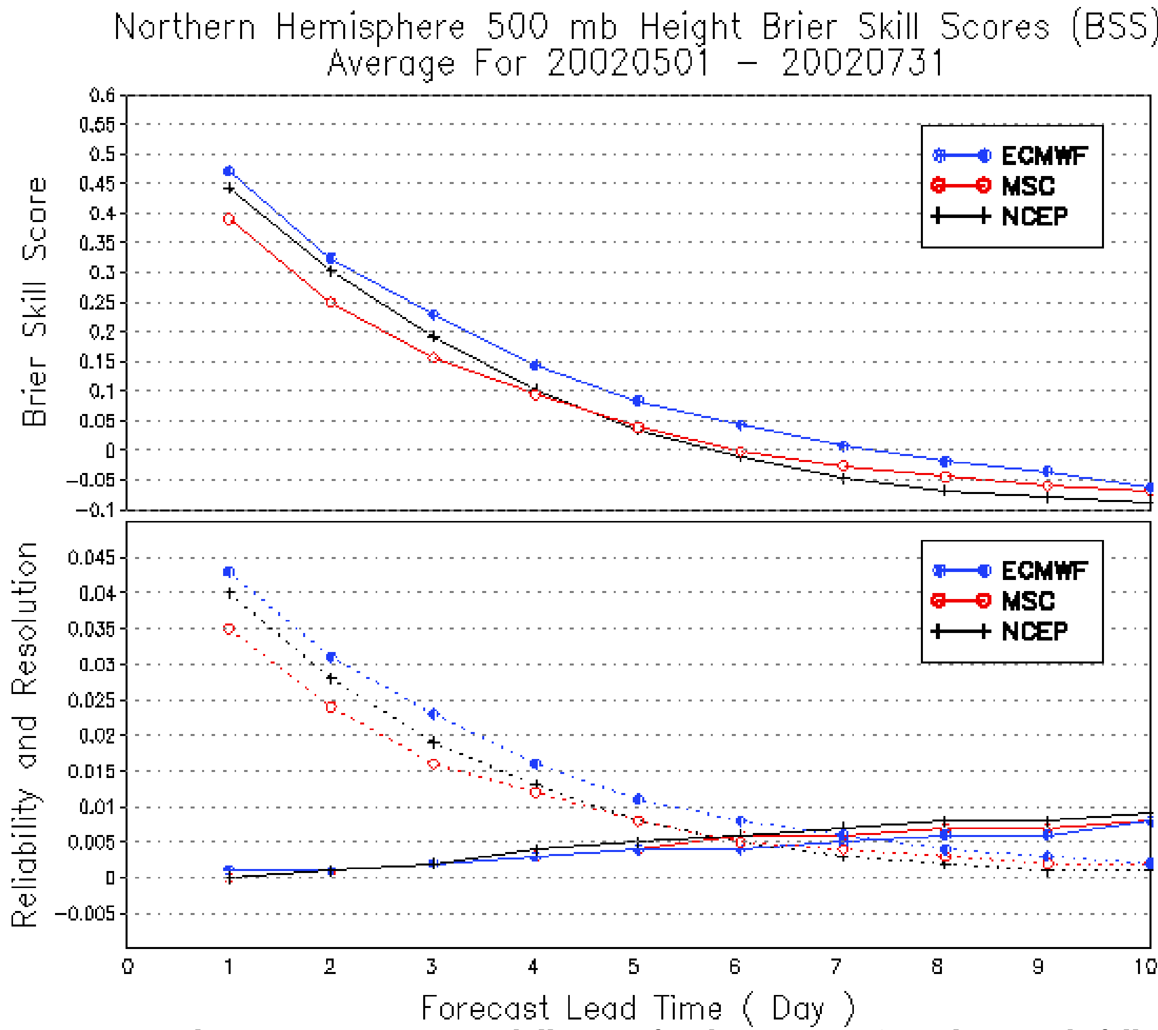
BSS	Higher better
Resolution	Higher better
Reliability	Lower better

RESULTS

Resolution dominates initially

Reliability becomes important later

- **ECMWF** best throughout
  - Good analysis/model?
- **NCEP** good days 1-2
  - Good initial perturbations?
  - No model perturb. hurts later?
- **CANADIAN** good days 8-10
  - Model diversity helps?



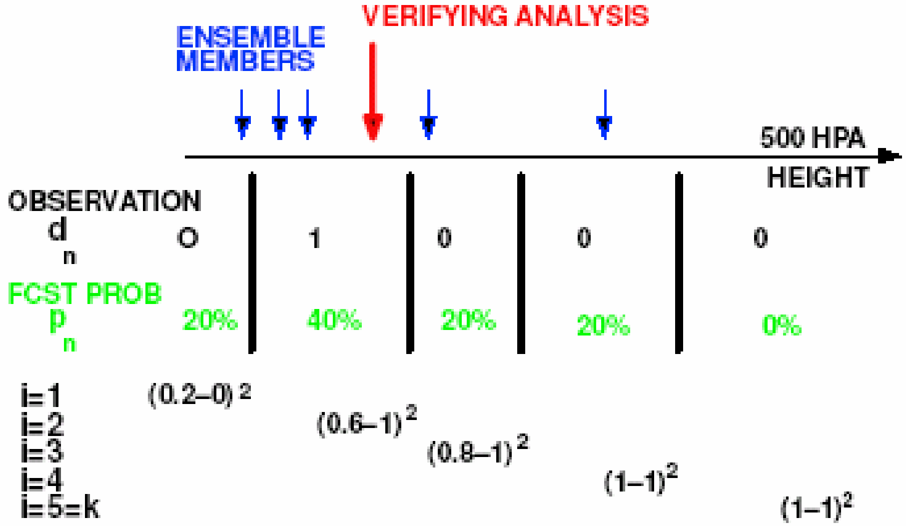
May-June-July 2002 average Brier skill score for the EC-EPS (grey lines with full circles), the MSC-EPS (black lines with open circles) and the NCEP-EPS (black lines with crosses). Bottom: resolution (dotted) and reliability(solid) contributions to the Brier skill score. Values refer to the 500 hPa geopotential height over the northern hemisphere latitudinal band 20º-80ºN, and have been computed considering 10 equally-climatologically-likely intervals (from Buizza, Houtekamer, Toth et al, 2004)

RANKED PROBABILITY SCORE

COMBINED MEASURE OF RELIABILITY AND RESOLUTION

For verifying multicategory probability forecasts in case categories can be ranked or ordered (like temperature)

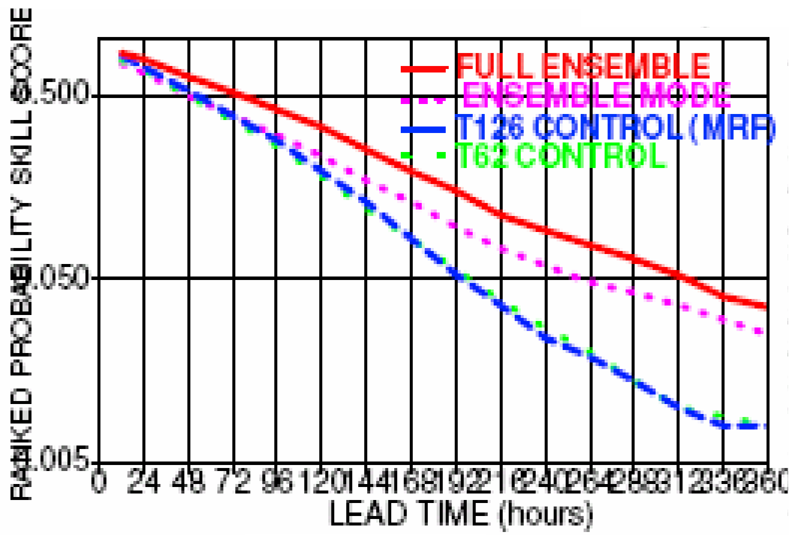
Generalization of Brier score (used for 2–category prob.)



k = number of categories

$$RPS(p,d) = \frac{1}{k-1}\left[\sum_{i=1}^{k}\left(\sum_{n=1}^{i} P_n - \sum_{n=1}^{i} d_n\right)^2\right]$$

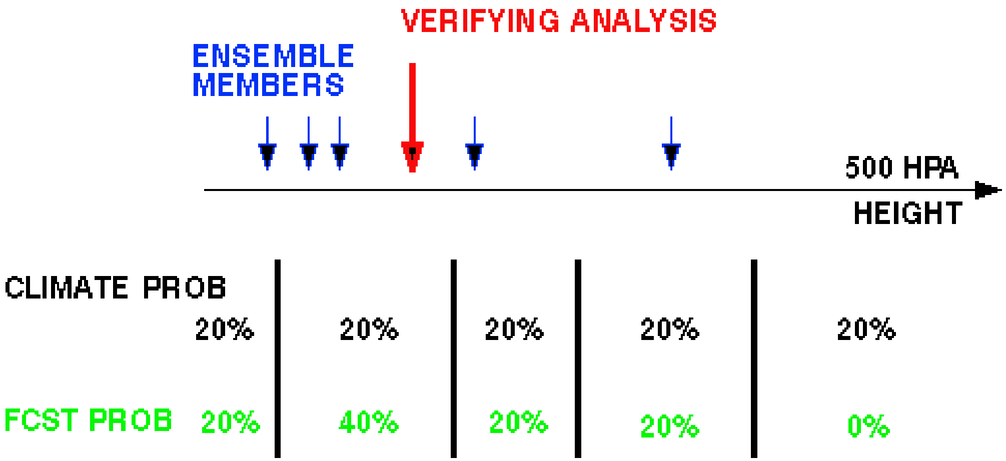
$$\text{RPS Skill Score (RPSS)} = 1 - \frac{\text{RPS (forecast)}}{\text{RPS (climatology)}}$$

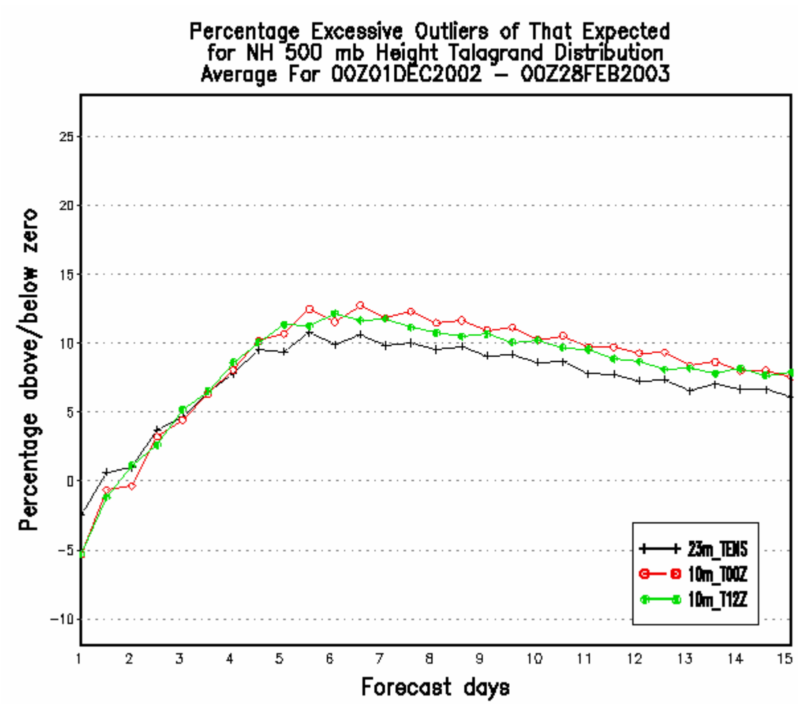


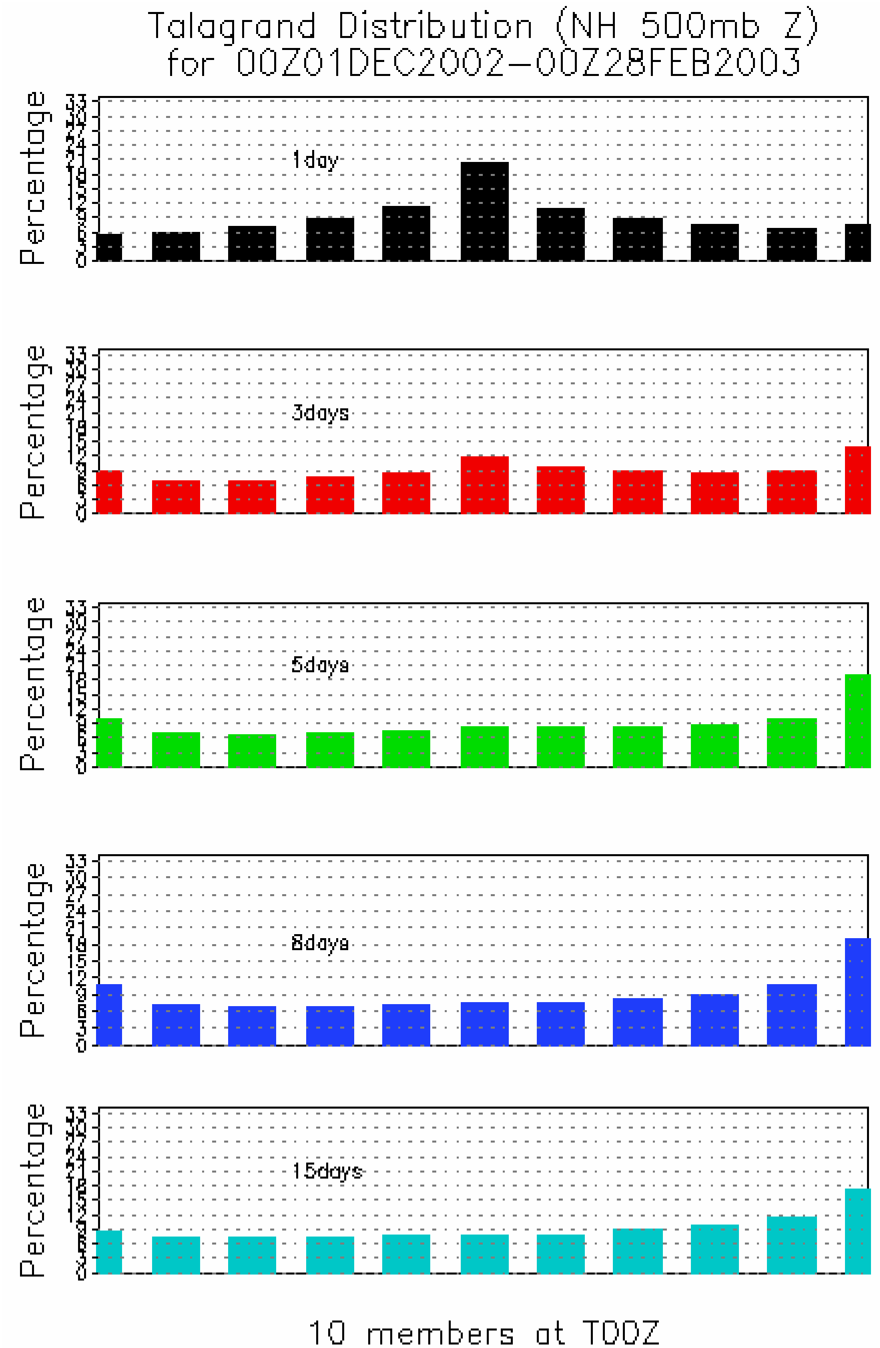
Ranked probability skill score for a T62 and T126 control and a 10–member ensemble forecast for the 500 hPa height, NH extratropics, March–May 1997. Forecast probabilities are made for 10 climatologically equally likely bins and are based on verification statistics from previous month (calibrated forecasts). Control forecasts have two probabilities depending on whether the forecast is in or not in a bin whereas the ensemble probabilities vary depending on how many ensemble members fall in a bin.

ANALYSIS RANK HISTOGRAM (TALAGRAND DIAGRAM)

MEASURE OF RELIABILITY





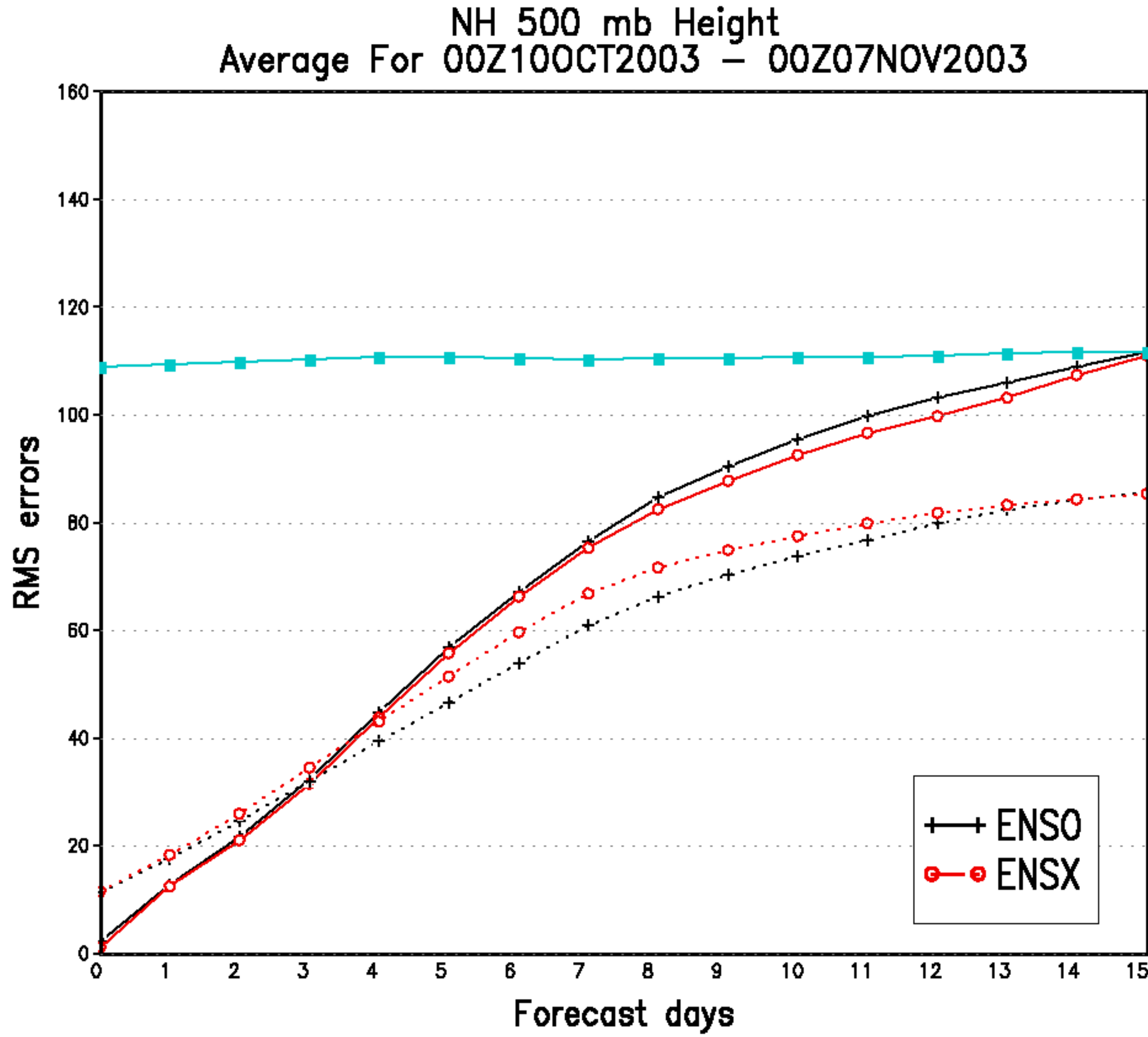


ENSEMBLE MEAN ERROR VS. ENSEMBLE SPREAD

MEASURE OF RELIABILITY

Statistical consistency between the ensemble and the verifying analysis means that the verifying analysis should be statistically indistinguishable from the ensemble members =>

Ensemble mean error (distance between ens. mean and analysis) should be equal to ensemble spread (distance between ensemble mean and ensemble members)



In case of a *statistically consistent ensemble*, ens. spread = ens. mean error, and they are both a MEASURE OF RESOLUTION. In the presence of bias, both rms error and PAC will be a combined measure of reliability and resolution

INFORMATION CONTENT

MEASURE OF RESOLUTION

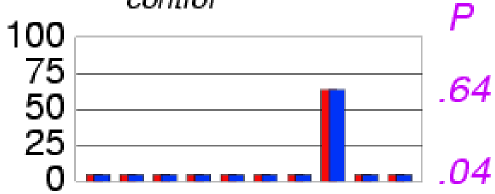
Use 10 climatologically equally likely bins to define events

Entropy $= P\log_2 P_i$

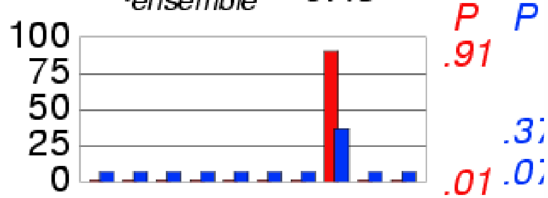
Information in one forecast $= I = 1 - \sum_{i=1}^{10} P_i \log_{10} P_i$

Average info in n ***independent fcsts*** $= I_{ave} = \frac{1}{n} \sum_{i=1}^{n} I_i$

$I_{control} = 0.37$



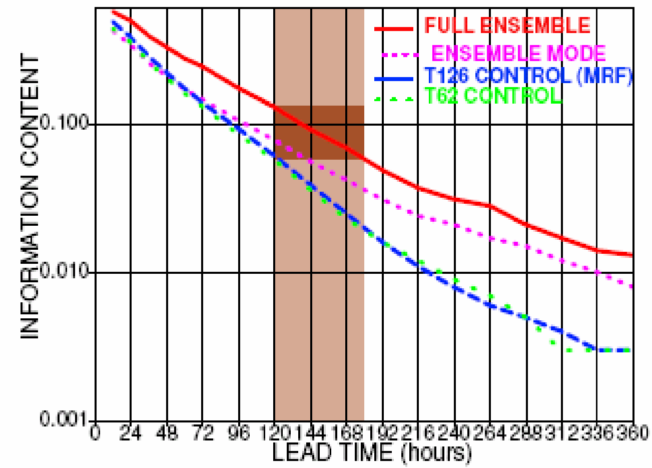
$I_{ensemble} = 0.45$



Categorical control fcst can use only a fixed set of probabilities based on average reliability

Ensemble can differentiate between well and less predictable situations

Information content of probabilistic forecasts based on the full ensemble distribution (red continuous line), the mode (most frequent value) of a 10–member ensemble (purple dotted), and the T62 (greed short dash) and T126 (blue long dash) control forecasts for the ***NH extratropics****, for March–May 1997. Forecasts are made for 10 climatologically equally likely bins. The bin where the control or ensemble mode falls is assigned a probability corresponding to the observed frequency of the verifying analysis falling into the same bin (P), while the remaining 9 bins are assigned (1–P)/9 (assuming perfect reliability that is close to be satisfied when using calibrated forecasts). Probabilities for the full ensemble are based on the number of ensemble members falling into the various bins. Note that the ensemble–based forecast probabilities can vary widely from case to case, depending on how the ensemble members spread while they are fixed for the control forecasts. The advance knowledge of the case dependent reliability of the forecasts transletes into substantial gains in terms of the information content the forecasts carry.*



ON AVERAGE A 7.5–DAY FULLY PROBABILISTIC FORECAST OR A 6–DAY CATEGORICAL FORECAST ASSOCIATED WITH CASE DEPENDENT RELIABILITY ESTIMATES HAS AS MUCH INFORMATION CONTENT AS A 5–DAY CATEGORICAL FORECAST

A 7.5–DAY FULLY PROBABILISTIC FORECAST HAS MORE

We assume that forecasts are perfectly reliable (forecast probabilities match observed frequencies)

For control: Use average reliability when fcst falls/ doesn't fall in a climate bin (fixed value)

For ensemble: Use average reliability for bin with most ensemble members (depends on how many fcsts fell in bin), distribute remaining probabilities equally among rest of bins

RELATIVE OPERATING CHARACTERISTICS

MEASURE OF RESOLUTION

Application of signal detection theory for measuring discrimination between two alternative outcomes

Worded, categorical and probab. forecasts can be compared

Stratification according to observations – reliability NOT measured

– Missed events not considered directly

		FORECAST	
		YES	*NO*
OBSERVATION	YES	H(its)	M(isses)
	NO	F(alse alarms)	C(orrect rejections)

$$\textit{Hit Rate (HR)} = \frac{H}{H+M} \qquad \textit{False Alarm Rate (FAR)} = \frac{F}{F+C}$$

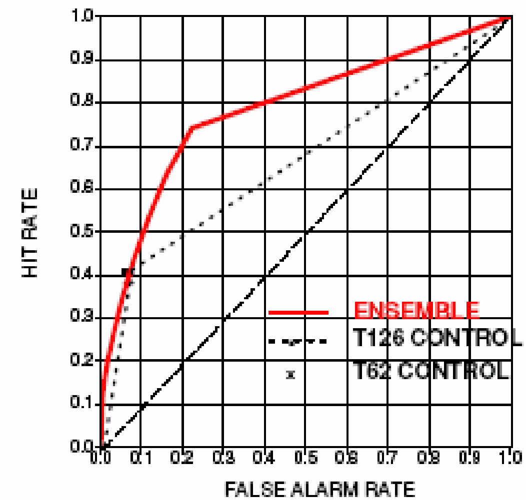
Use 10 climatologically equally likely bins to define events

Categorical forecast: If control falls in a given climate bin, forecast is YES and NO otherwise

Ensemble forecast: Probabilities converted to a categorical fcst given the probability exceeds a certain threshold. Eg., all 30% or higher probabilities count as YES. Using different threshold probabilities yield an HR/FA diagram.

Measures: 1) Area between HR–FAR curve and diagonal

2) How different forecast probabilities are given different observations



ROC (Relative Operating Characteristics) curve for a 10–member T62 ensemble of forecasts and for T126 and T62 control forecasts for the 500 hPa height, NH extratropics, March–May 1997. The closer a curve is to the upper left hand corner, the more ability the forecasting system has in delineating between cases when a certain event (in this case, the occurence of one of 10 climatologically equally likely bins) did or did not occur.

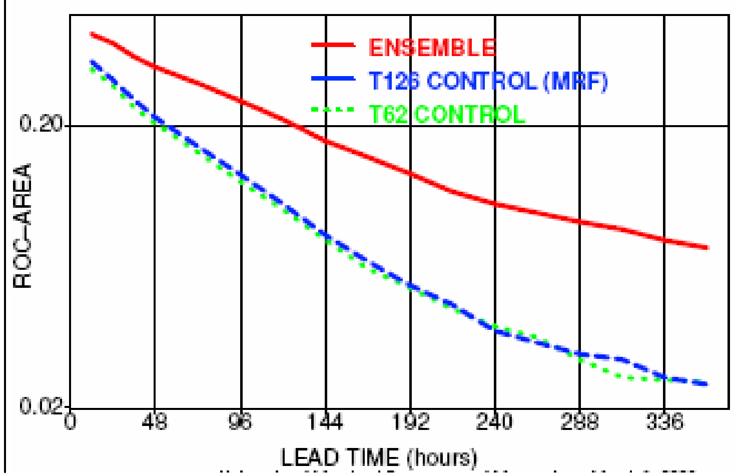


Fig. 6. ROC (Relative Operating Characteristics) area for T126 and T62 control and 10–member T62 ensemble forecasts for the 500 hPa height, NH extratropics, March–May 1997.

ECONOMIC VALUE OF FORECASTS

MEASURE OF RESOLUTION

Given a particular forecast, a user either does or does not take action (eg, protects its crop against frost) *Mylne & Harrison, 1999*

		FORECAST YES	FORECAST NO
OBSERVATION	YES	H(its) ***Mitigated Loss***	M(isses) ***Loss***
OBSERVATION	NO	F(alse alarms) ***Cost***	C(orrect rejections) ***No Cost***

$$Mean\ Expense_{fc} = hML + mL + fC$$

$$Mean\ Expense_{perf} = oML$$

$$ME_{cl} = \min[oL,\ oML + (1-o)C]$$

$$Value = \frac{ME_{cl} - ME_{fc}}{ME_{cl} - ME_{perf}}$$

o=climatological frequency

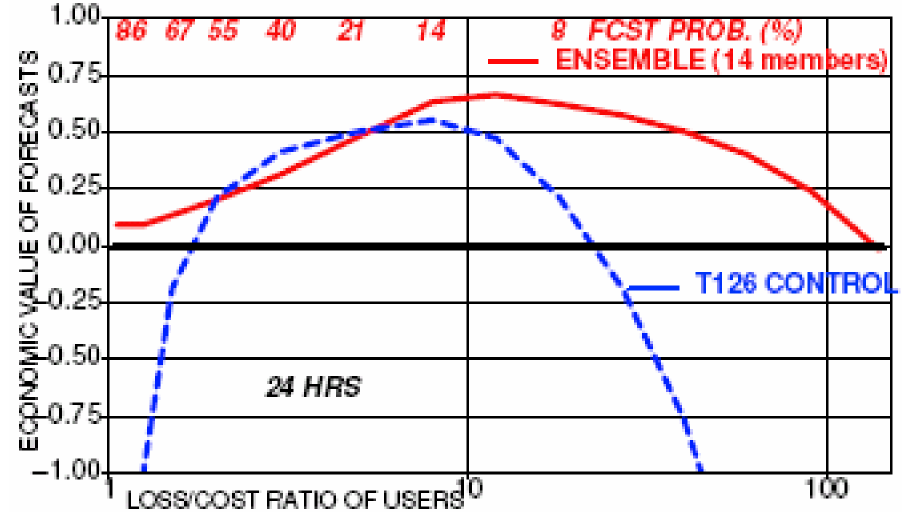
Use 10 climatologically equally likely bins to define events

Hi-res control forecast: If MRF control falls in a given climate bin, forecast is YES and NO otherwise

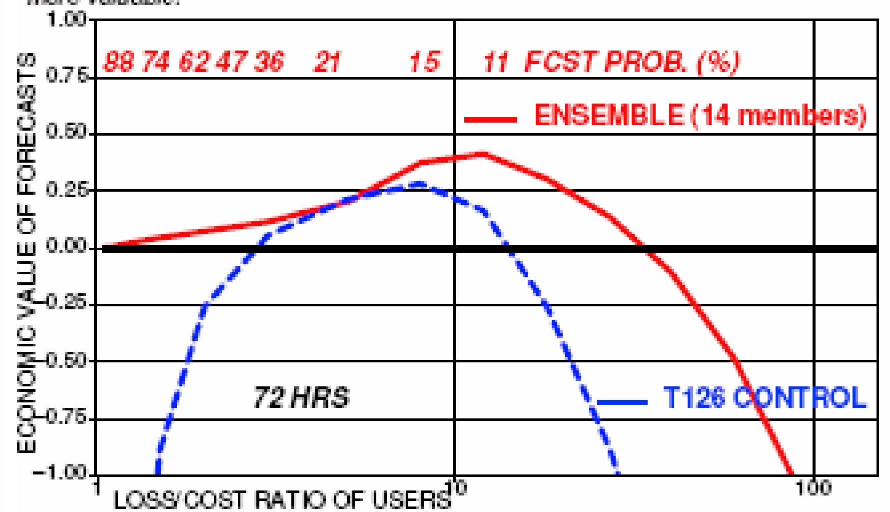
Lo-res ensemble forecast: Probabilities converted to a categorical fcst given the probability exceeds a certain threshold. Eg., all 30% or higher probabilities count as YES. Among different threshold probabilities one can select the one that results in largest economic value.

Results: For majority of users ensemble is more useful

Question: Is it because MRF is dichotomous, while ensemble provides full probability distribution?



Economic value of 24-hour MRF T126 control, and 14-member T62 ensemble forecasts in predicting events defined in terms of 10 climatologically equally likely bins for 500 hPa height over the NH extratropics, for April-June 1999, for users characterized by different loss/cost ratios (horizontal axis, logarithmic scale). A value of 1.0 stands for using perfect forecasts while values below zero indicate that climatological forecasts are more valuable.



Economic value of 72-hour MRF T126 control, and 14-member T62 ensemble forecasts in predicting events defined in terms of 10 climatologically equally likely bins for 500 hPa height over the NH extratropics, for April-June 1999, for users characterized by different loss/cost ratios (horizontal axis, logarithmic scale). A value of 1.0 stands for using perfect forecasts while values below zero indicate that climatological forecasts are more valuable.

PERTURBATION VS. ERROR CORRELATION ANALYSIS (PECA)

MULTIVATIATE COMBINED MEASURE OF RELIABILITY & RESOLUTION

METHOD: Compute correlation between ens perturbtns and error in control fcst for

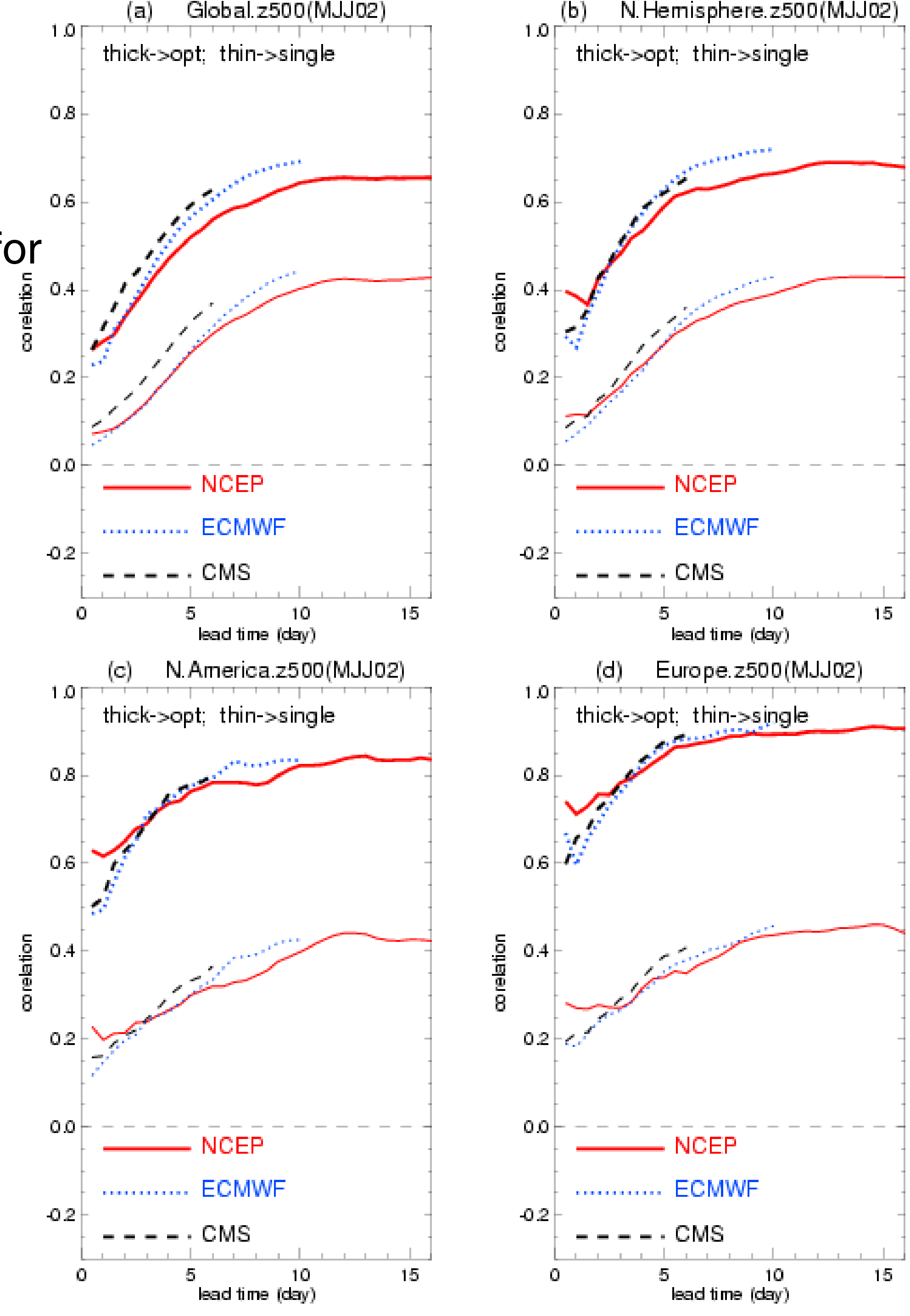
- Individual members
- Optimal combination of members
- Each ensemble
- Various areas, all lead time

EVALUATION: Large correlation indicates ens captures error in control forecast

- Caveat – errors defined by analysis

RESULTS:

- **Canadian** best on large scales
  - Benefit of model diversity?
- **ECMWF** gains most from combinations
  - Benefit of orthogonalization?
- **NCEP** best on small scale, short term
  - Benefit of breeding (best estimate initial error)?
- PECA increases with lead time
  - Lyapunov convergence
  - Nonlilnear saturation
- Higher values on small scales



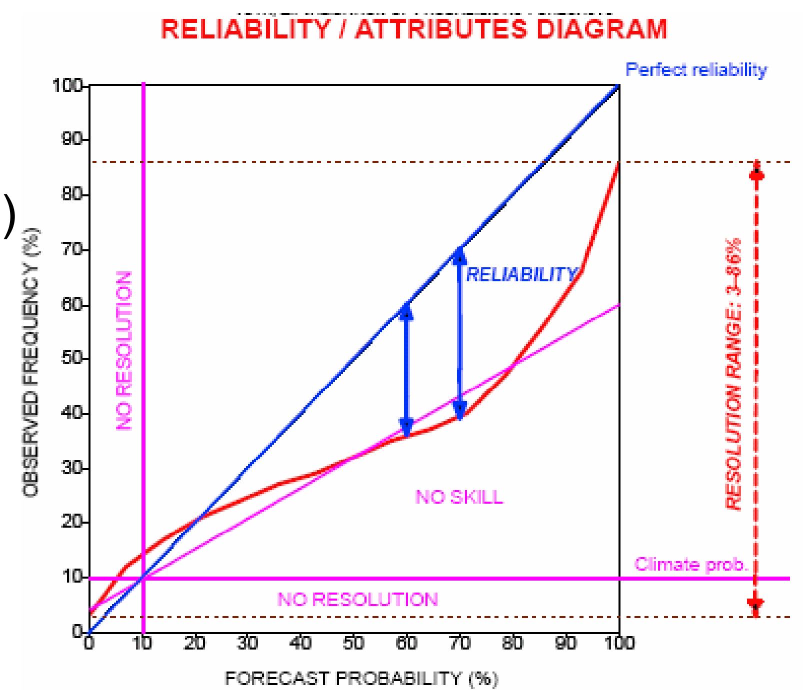
WHAT WE NEED FOR POSTPROCESSING TO WORK?

- **LARGE SET OF FCST – OBS PAIRS**
  - Consistency defined over large sample – need same for post-processing
  - Larger the sample, more detailed corrections can be made

- **BOTH FCST AND REAL SYSTEMS MUST BE STATIONARY IN TIME**
  - Otherwise can make things worse
  - Subjective forecasts difficult to calibrate

HOW WE MEASURE STATISTICAL INCONSISTENCY?

- **MEASURES OF STATIST. RELIABILITY**
  - Time mean error
  - Analysis rank histogram (Talagrand diagram)
  - Reliability component of Brier etc scores
  - Reliability diagram



SOURCES OF STATISTICAL INCONSISTENCY

- **TOO FEW FORECAST MEMBERS**
  - Single forecast – inconsistent by definition, unless perfect
      - MOS fcst hedged toward climatology as fcst skill is lost
  - Small ensemble – sampling error due to limited ensemble size

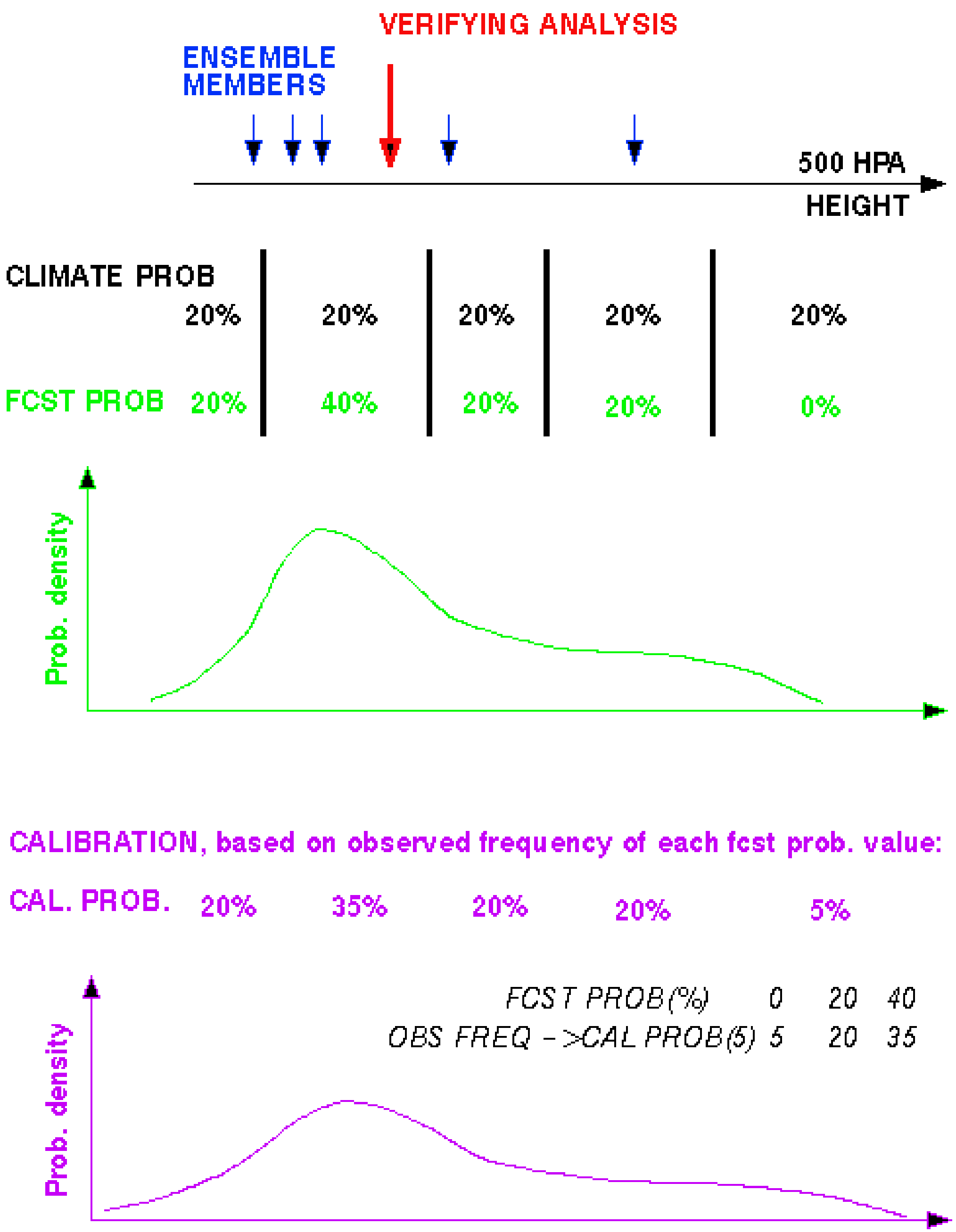
    (Houtekamer 1994?)
- **MODEL ERROR (BIAS)**
  - Deficiencies due to various problems in NWP models
    - Effect is exacerbated with increasing lead time
- **SYSTEMATIC ERRORS (BIAS) IN ANALYSIS**
  - Induced by observations
    - Effect dies out with increasing lead time
  - Model related
    - Bias manifests itself even in initial conditions
- **ENSEMBLE FORMATION (INPROPER SPREAD)**
  - Not appropriate initial spread
  - Lack of representation of model related uncertainty in ensemble
    - I. E., use of simplified model that is not able to account for model related uncertainty

HOW TO IMPROVE STATISTICAL CONSISTENCY?

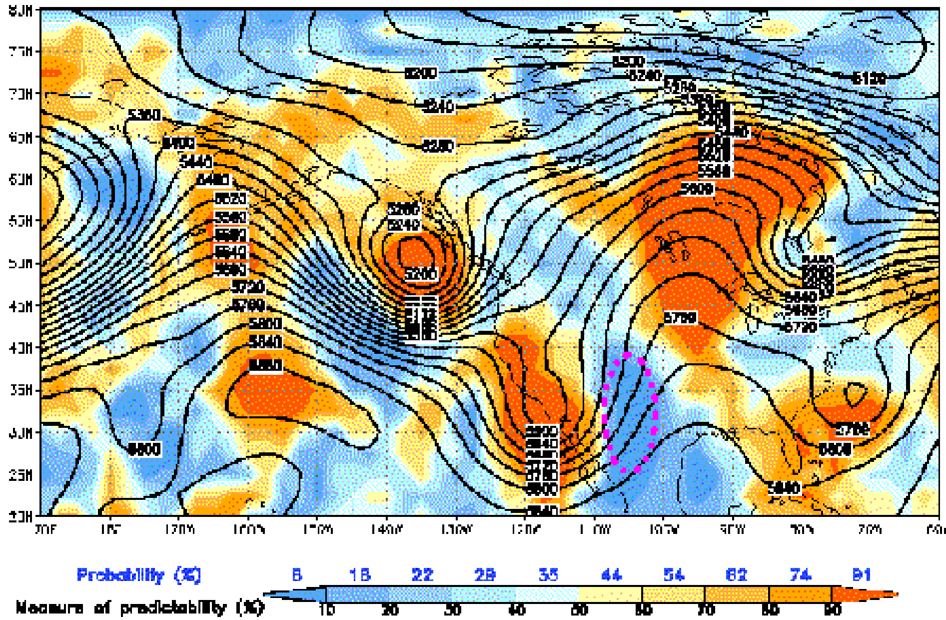
- **MITIGATE SOURCES OF INCONSISTENCY**
  - TOO FEW MEMBERS
    - Run large ensemble
  - MODEL ERRORS
    - Make models more realistic
  - INSUFFICIENT ENSEMBLE SPREAD
    - Enhance models so they can represent model related forecast uncertainty
  - OTHERWISE =>

- **STATISTICALLY ADJUST FCST TO REDUCE INCONSISTENCY**
  - Unpreferred way of doing it
  - What we learn can feed back into development to mitigate problem at sources
  - Can have LARGE impact on (inexperienced) users

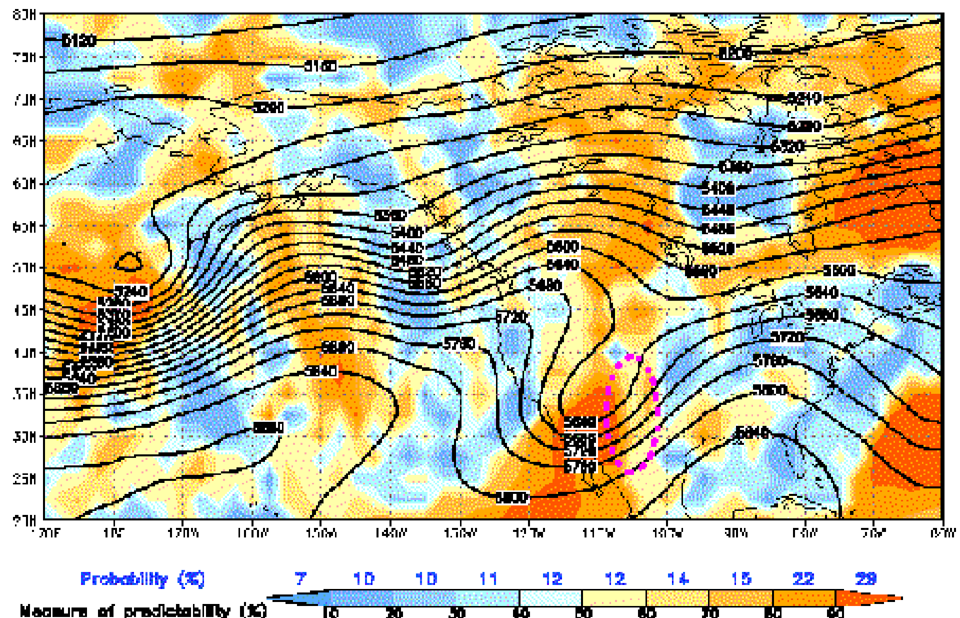
ENSEMBLE BASED PROBABILISTIC FORECASTS AND THEIR VERIFICATION



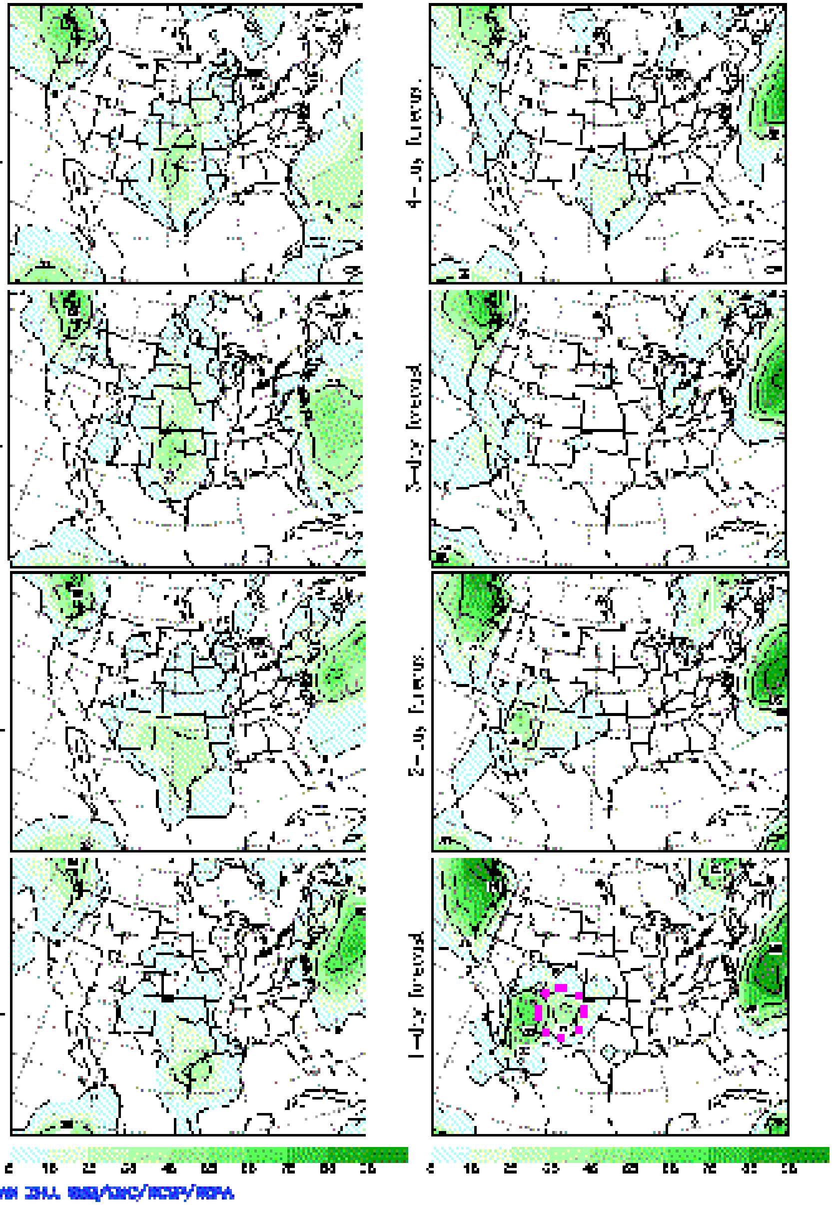
Relative measure of predictability (colors)
for ensemble mean forecast (contours) of 500 hPa height
ini: 2000102700 valid: 2000102800 fcst: 24 hours



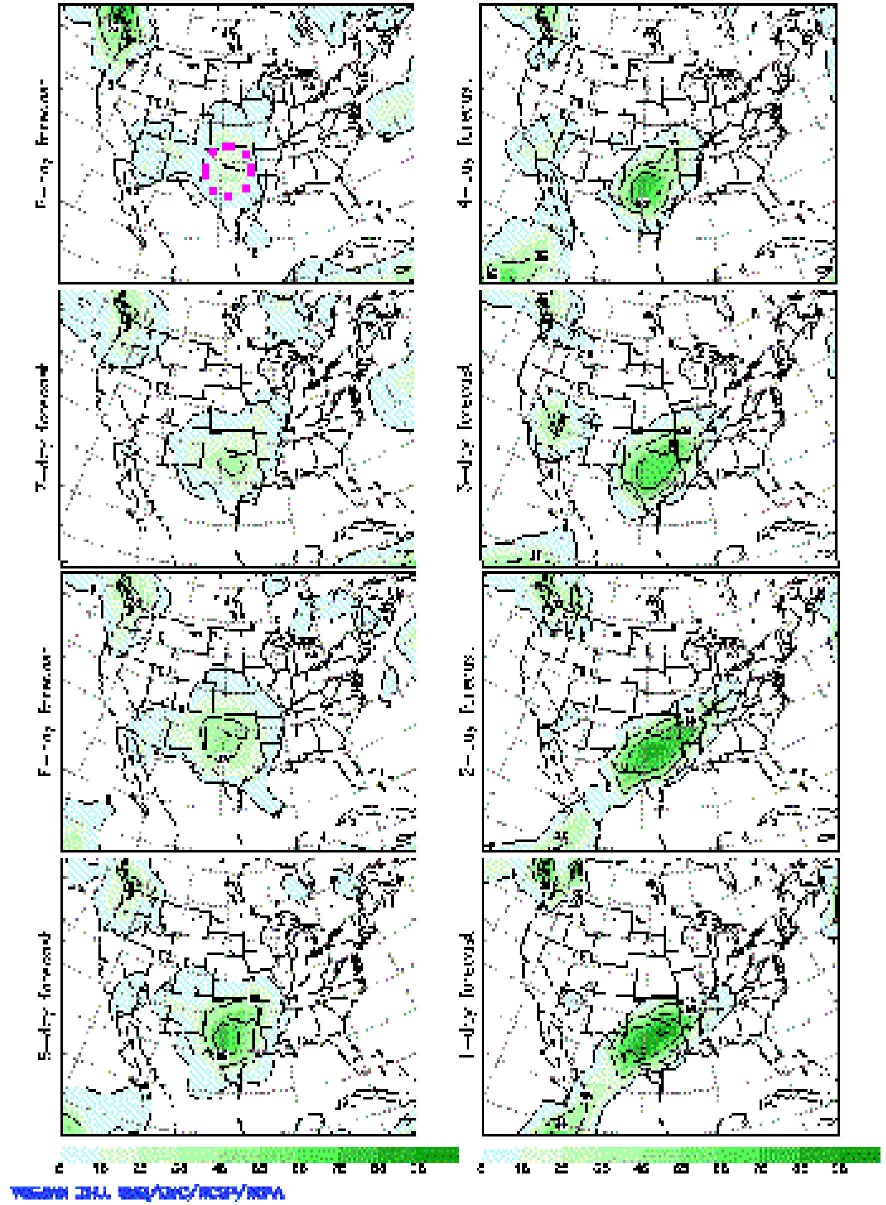
Relative measure of predictability (colors)
for ensemble mean forecast (contours) of 500 hPa height
ini: 2000102700 valid: 2000110400 fcst: 192 hours



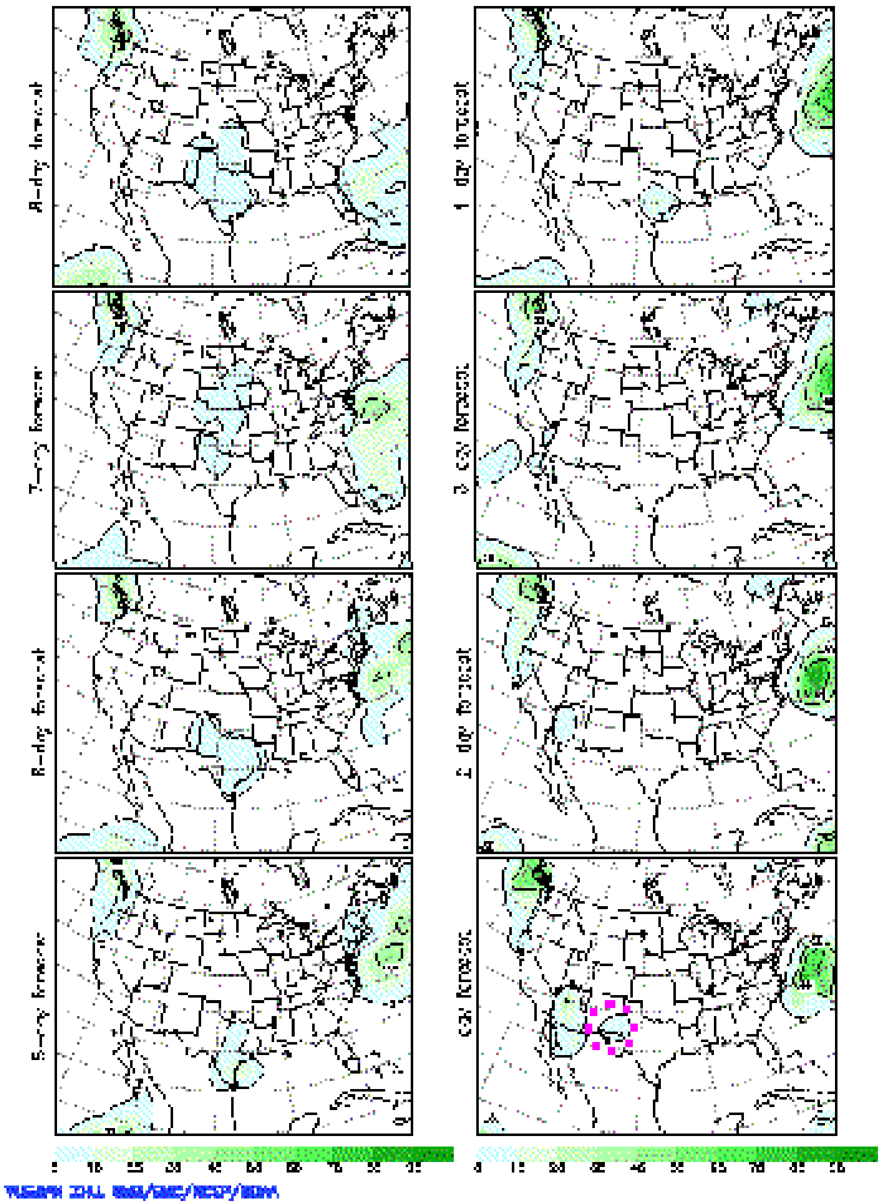
Ens Prob of Precip Amount Exceeding 0.5 inch (12.7 mm/day)
Valid Period: 2000102712–2000102812



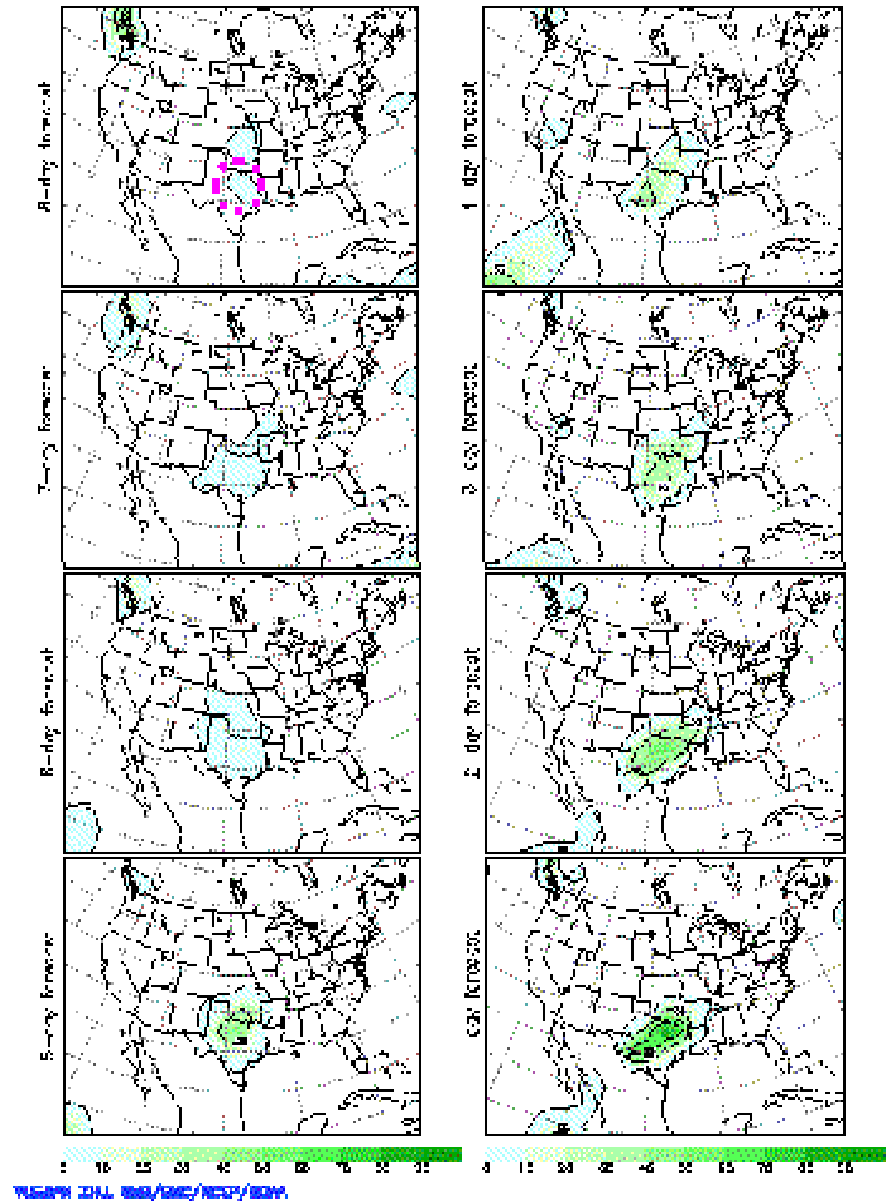
Ens Prob of Precip Amount Exceeding 0.5 inch (12.7 mm/day)
Valid Period: 2000110312–2000110412



Ens Prob of Precip Amount Exceeding 1.0 inch (25.4 mm/day)
Valid Period: 2000102712–2000102812



Ens Prob of Precip Amount Exceeding 1.0 inch (25.4 mm/day
Valid Period: 2000110312–2000110412



OUTLINE / SUMMARY

- **WHY DO WE NEED PROBABILISTIC FORECASTS?**
  - **Isn't the atmosphere deterministic?** **YES, but it's also CHAOTIC**

  FORECASTER'S PERSPECTIVE ***USER'S PERSPECTIVE***

  Ensemble techniques **Probabilistic description**

- **HOW CAN WE MAKE PROBABILISTIC FORECASTS?**

  STATISTICAL METHODS

  SINGLE DYNAMICAL FORECAST + VERIFICATION STATISTICS

  ENSEMBLE FORECASTS

- **WHAT ARE THE MAIN ATTRIBUTES OF FORECASTS?**
  - ***RELIABILITY*** Stat. consistency with distribution of corresponding observations
  - ***RESOLUTION*** Different events are preceded by different forecasts

- **HOW CAN PROBABILSTIC FORECAST PERFORMANCE BE MEASURED?**

  Various measures of reliability and resolution

- **STATISTICAL POSTPROCESSING**

  Based on verification statistics – reduce statistical inconsistencies