

Ensemble Based Probabilistic Forecast Verification

Yuejian Zhu and Zoltan Toth

Presents for DTC Verification Workshop
August 26-28 2009, NCAR, CO

Outlines

1. Why Do We Need Verification?
2. What Do We Like?
3. Deterministic Forecast
4. Probabilistic Forecast
5. Resolution and Reliability
6. Category Forecast
7. Economic Value
8. Hurricane Track (and Strike Probability)
9. References

1. Why do we need verification?

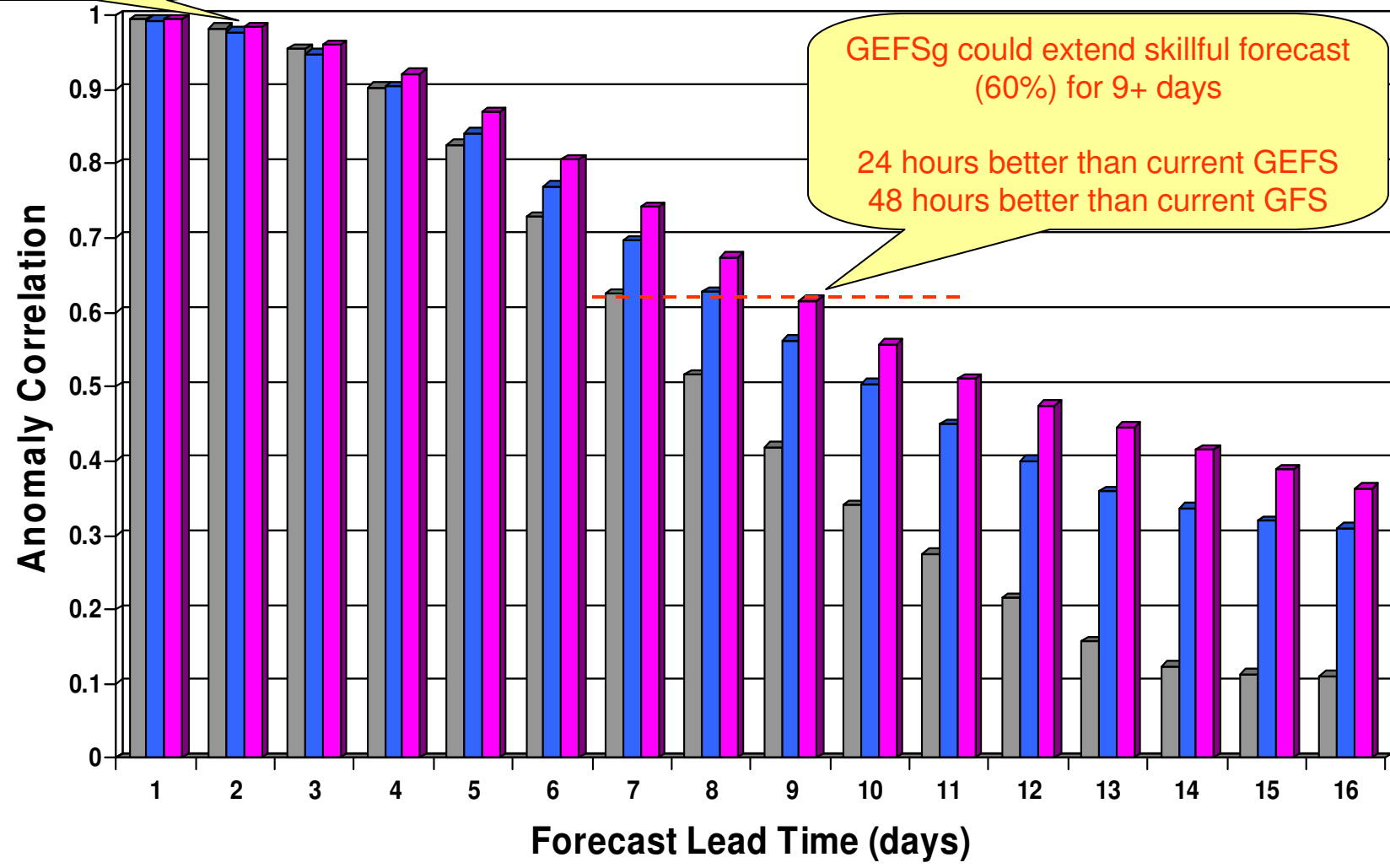
- For model developers
 - Retrospective experiments for
 - Data assimilation system
 - Model dynamics development
 - Model physics development
 - Comparison for
 - Different algorithms, systems
 - Different processes, models
- For forecasters/users
 - Confidence
 - To help them to understand the forecast
 - Systematic errors or bias
 - To make more accurate forecast
- For performance review
 - Short/long term plan

NH Anomaly Correlation for 500hPa Height

Period: August 1st – September 30th 2007

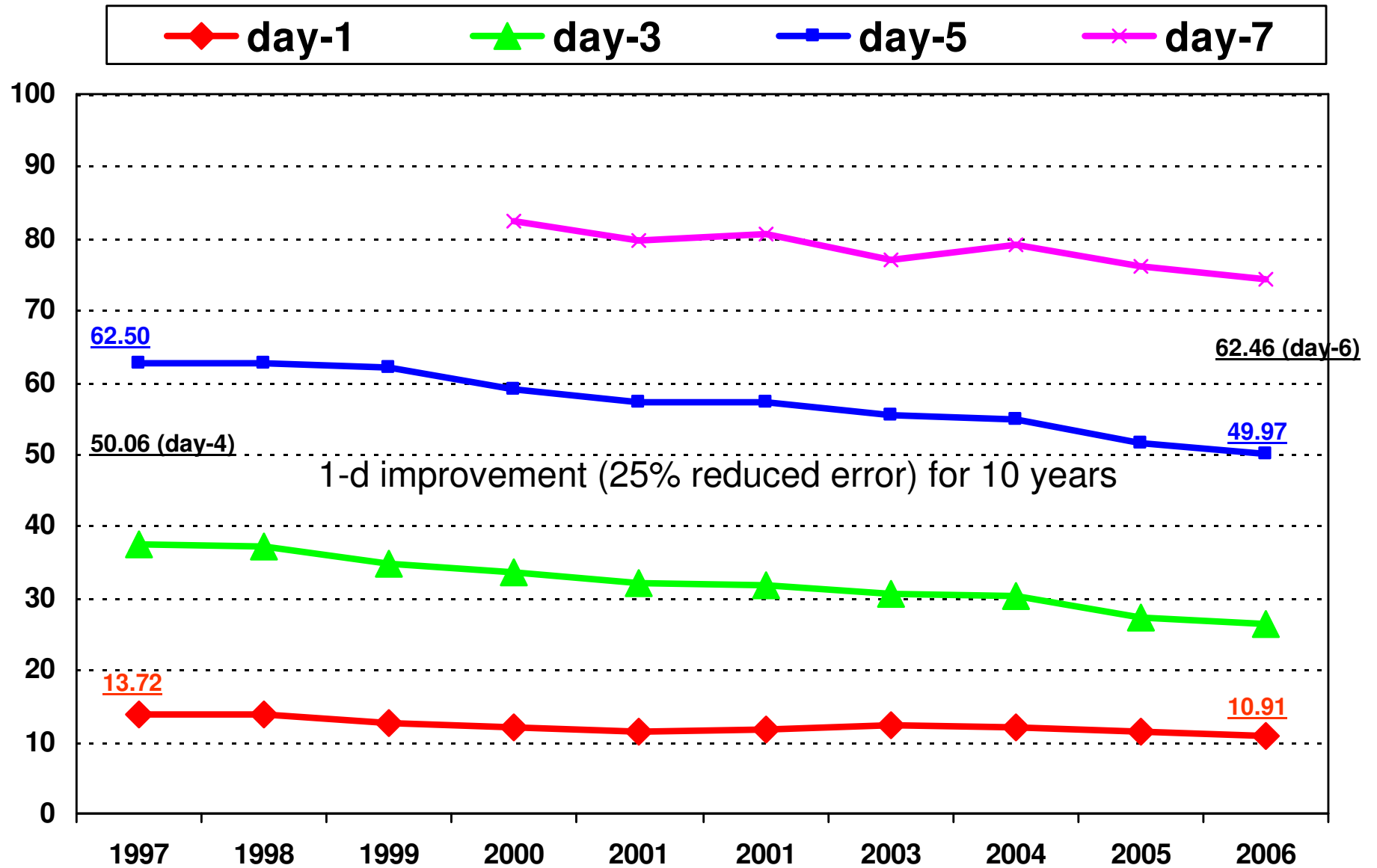
GEFSg is better than GFS at 48 hours

■ GFS ■ GEFS ■ GEFSg

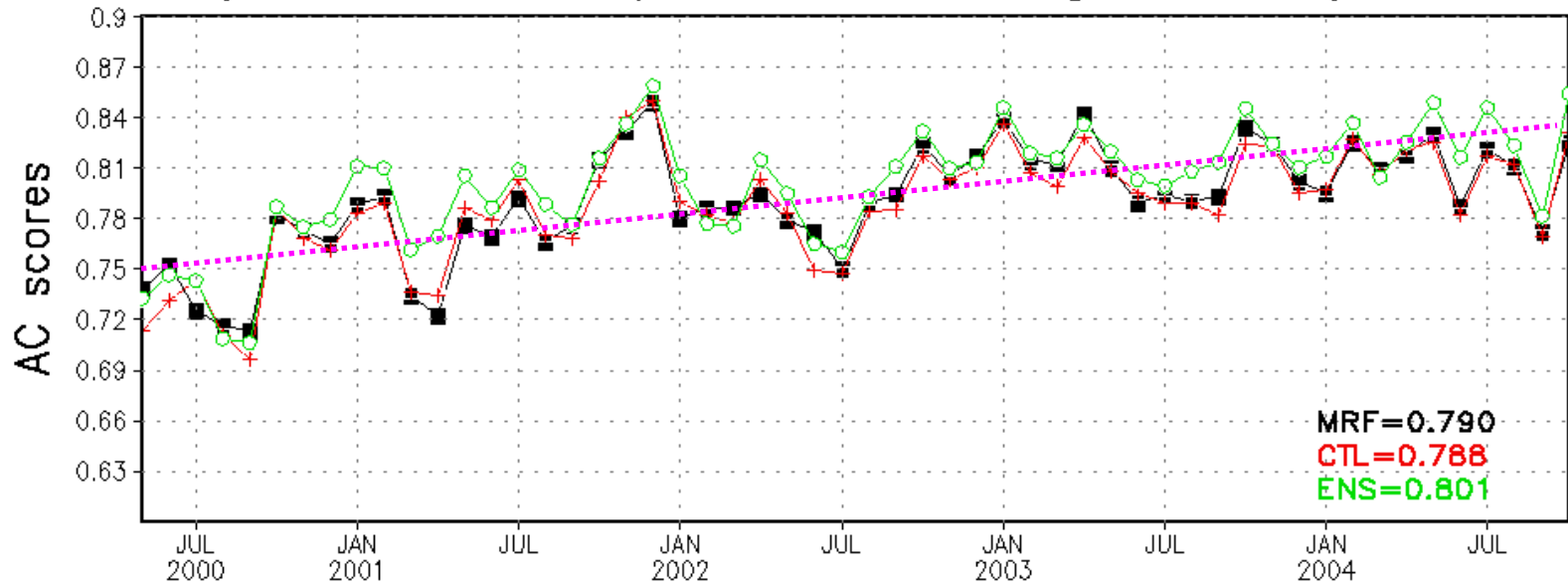


GEFSg could extend skillful forecast (60%) for 9+ days
24 hours better than current GEFS
48 hours better than current GFS

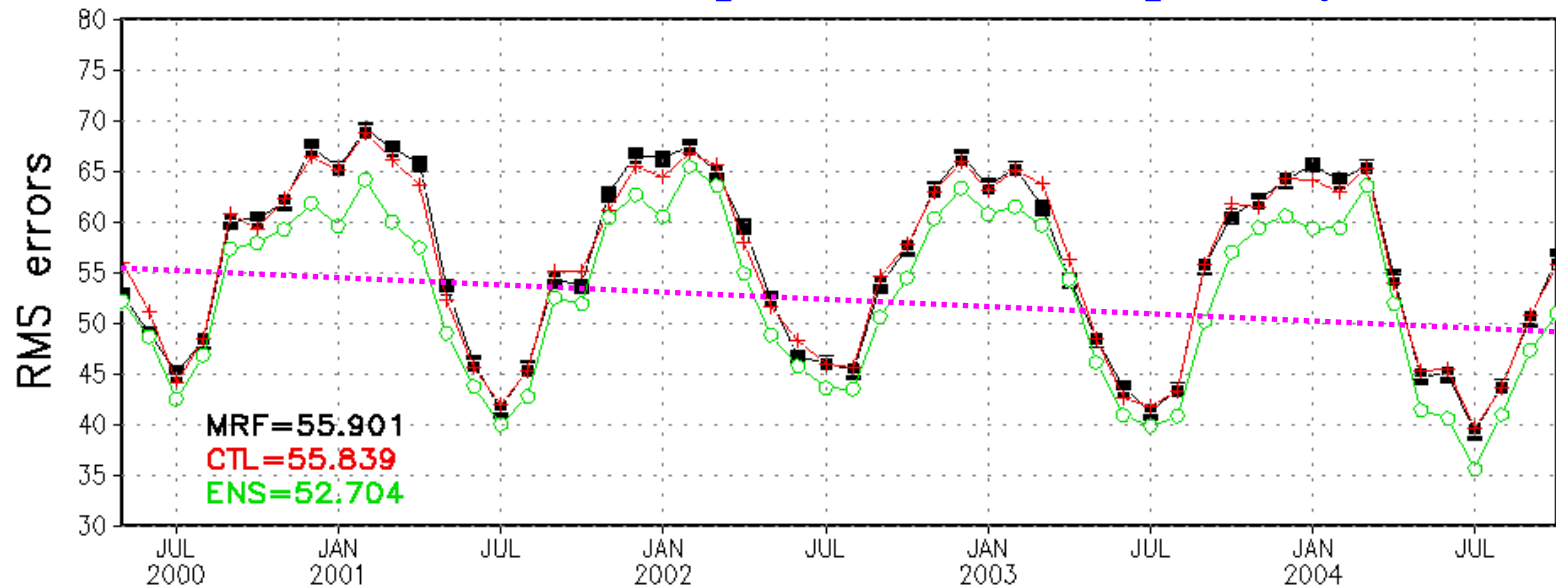
RMS Error for Northern Hemisphere (20-80N) 500hPa Height



Monthly Ave. Scores (NH 500hPa Height, 5-day forecasts)



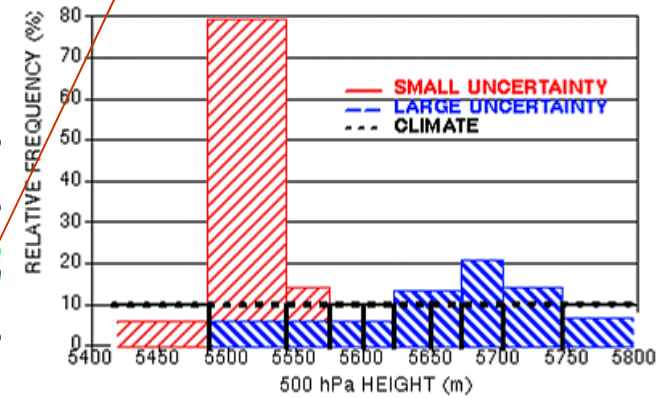
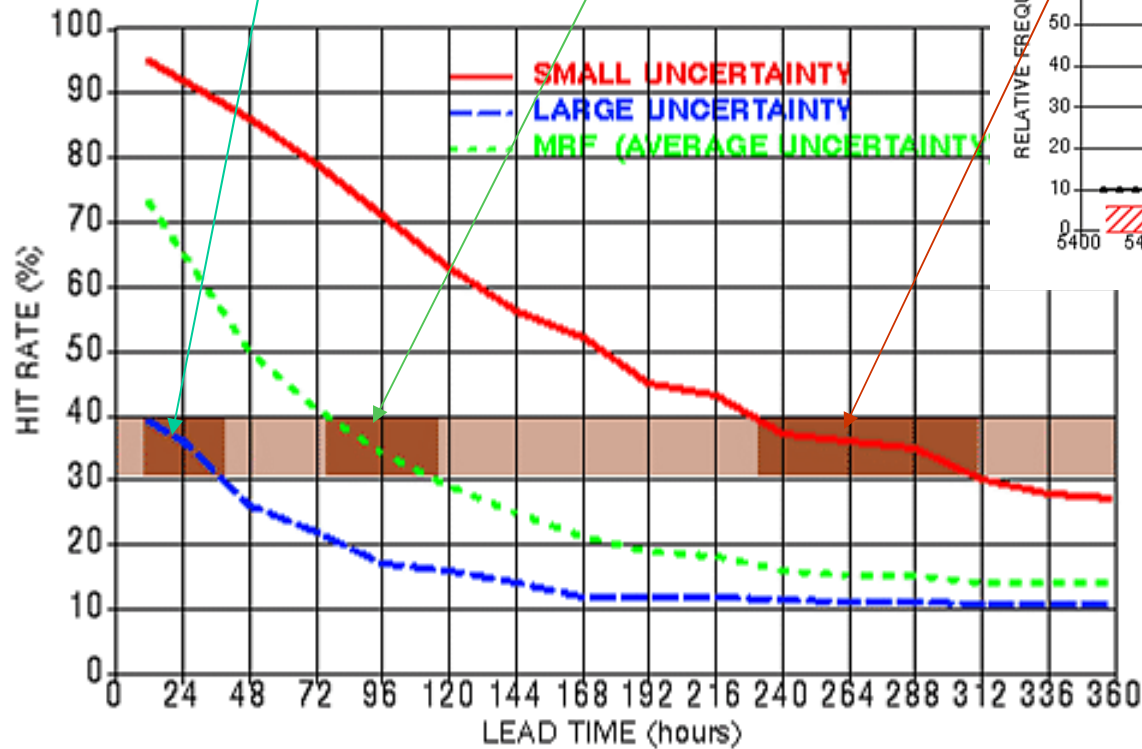
NCEP ensemble mean performance for past 5-year



Information from uncertainty forecast

... Small and large uncertainty.

1 day (large uncertainty) = 4 days (control) = 10-13 days (small uncertainty)



2. What do we like?

- Forecast accuracy?
 - The distance to observation/analysis
- Forecast consistency?
 - Forecast confidence
- Forecast skill?
 - Short/long term plan

2.1. Forecast Accuracy

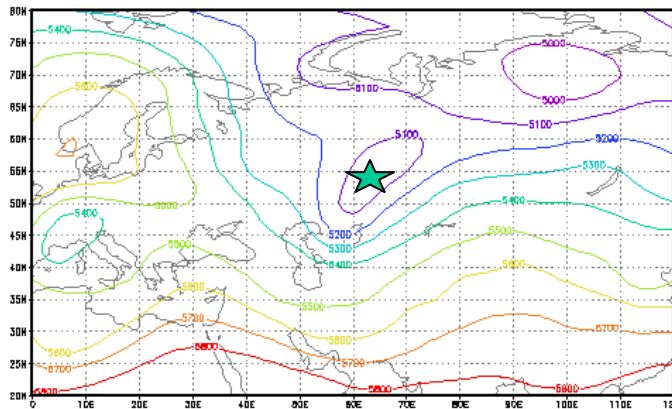
- Every one likes accurate forecast:
- Accurate forecasts should have:
 - Small difference between forecast to observation/analysis
 - Higher correlation
- Case study
 - Case dependent
 - Extreme event
- Common methods
 - For single forecast
 - The difference of forecast and observation/analysis (map)
 - For a case (a set of forecasts)
 - RMS
 - Mean errors or absolute error
 - Anomaly correlation

Subjective evaluation

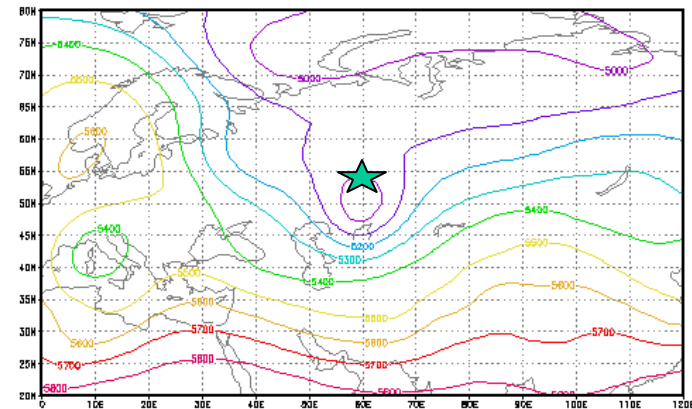
Synoptic example of 500hPa height forecast

Initial: 2003021200 Valid: 2003021700

Analysis (truth)



Forecast #1

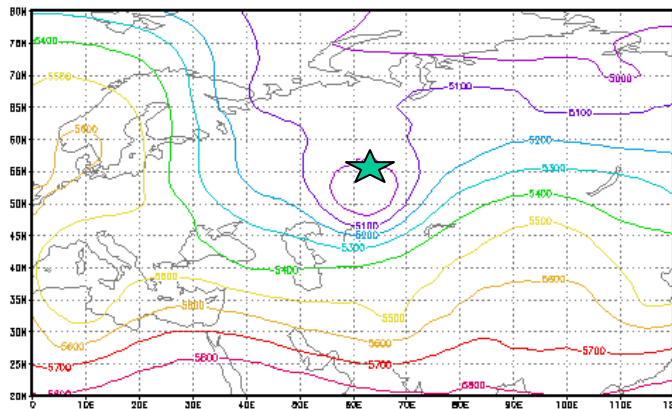


F#1=.7978

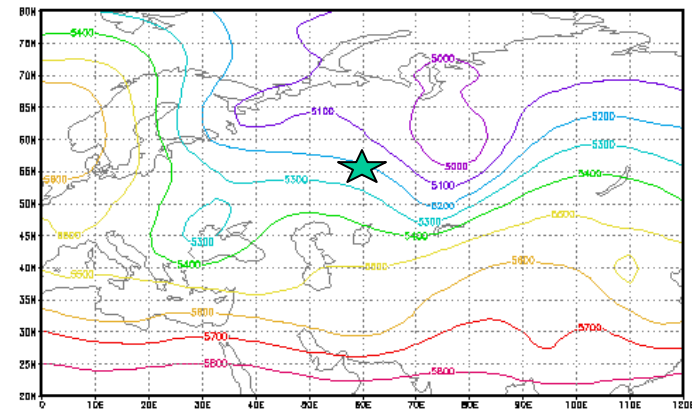
F#2=.8257

F#3=.6068

Forecast #2

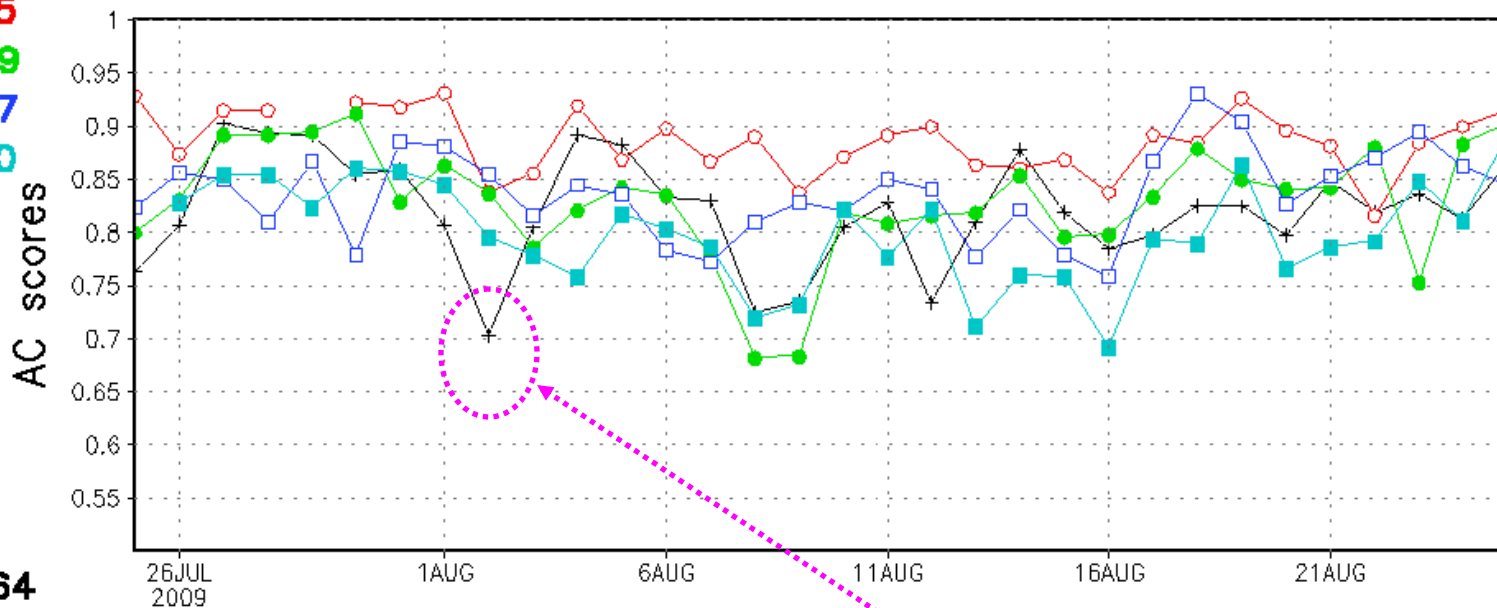


Forecast #3

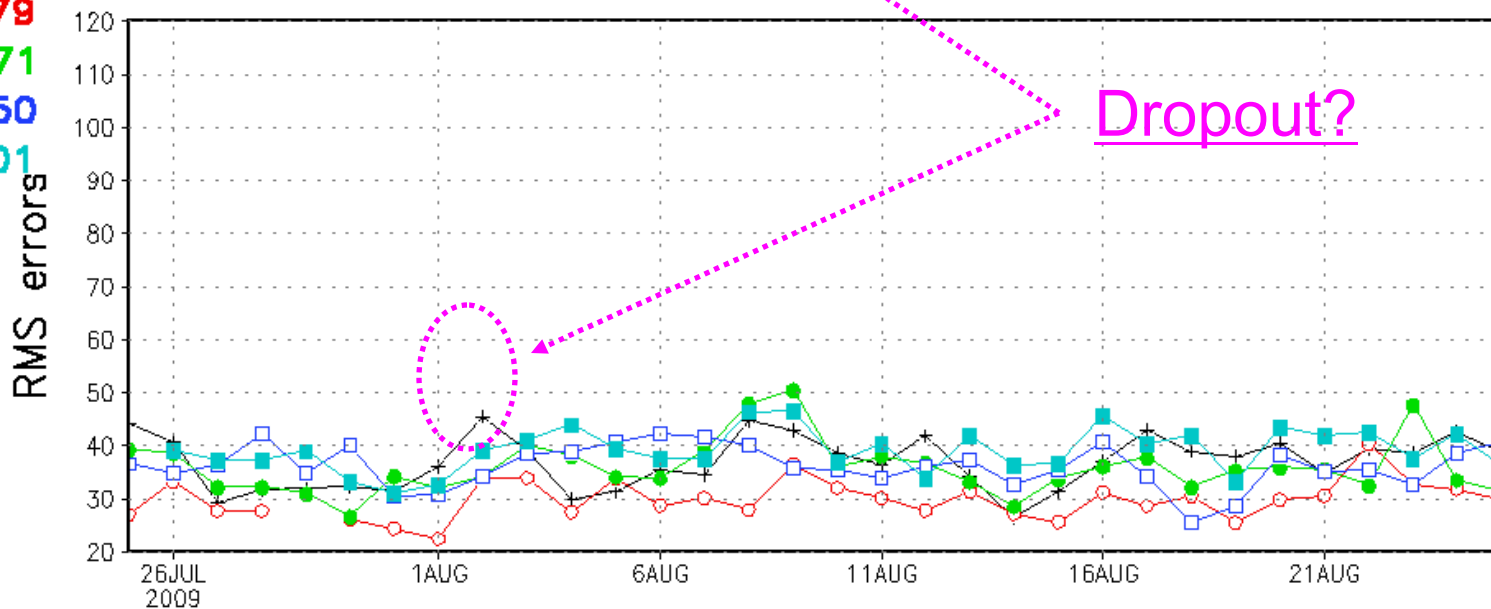


NH 500 hPa Geopotential Height at day 5 for 00Z25JUL2009 – 00Z25AUG2009

- =0.820
- =0.885
- =0.829
- =0.837
- =0.800



- =36.964
- =29.879
- =35.971
- =36.250
- =39.201

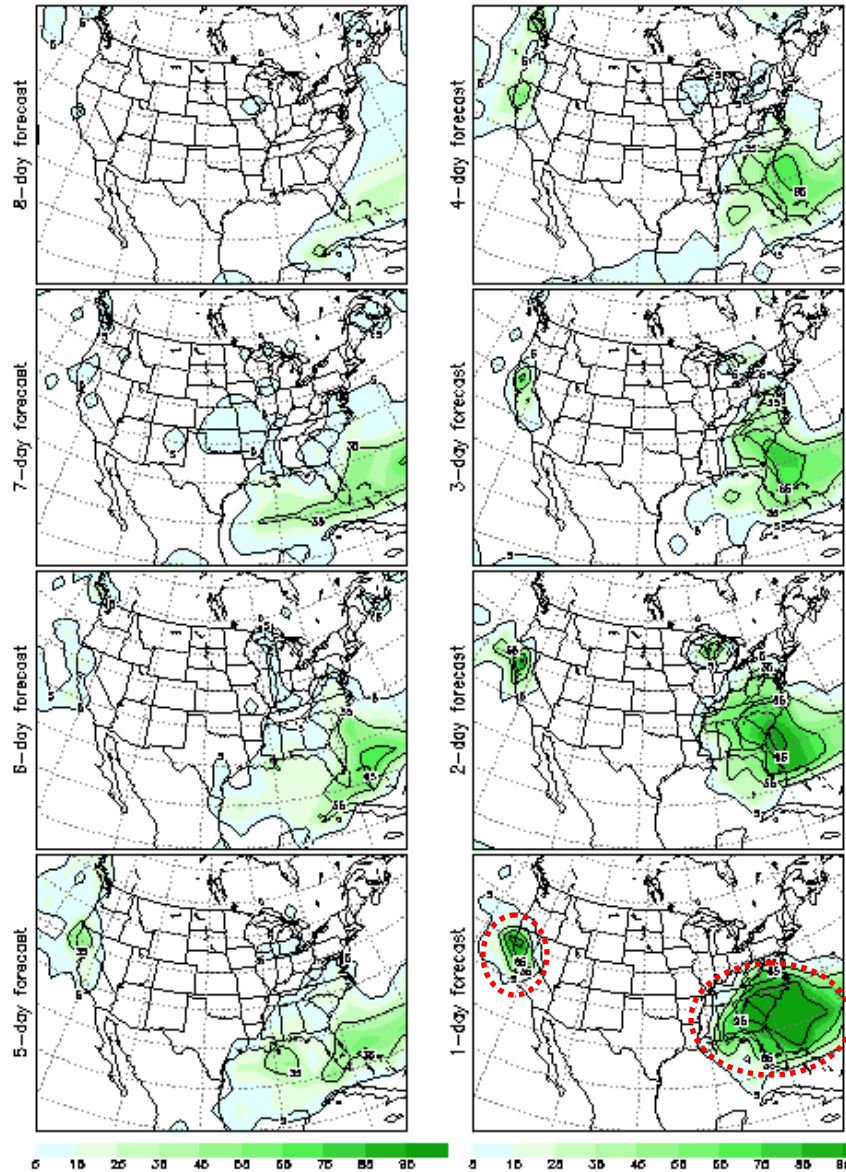


2.2 Forecast Consistency

- Expect to have the same/similar solution (valid forecast) from different initial conditions – “reliability”.
- Consistency is different from accuracy.
- User consideration (requirement), user likes consistency forecasts.
- The similar properties: deterministic (quantitative) or ensemble (probabilistic) forecasts.
- Consistency forecasts should increasing/decreasing predictability gradually from different lead times or cycles (6-hour or 24-hour)
- High consistency forecast is corresponding to high forecast accuracy and reliability (?)

Ens Prob of Precip Amount Exceeding 0.50 Inch (12.7 mm)
 Inl: 2005022700 Valld for 2005022712-2005022812

Ens Prob of Precip Amount Exceeding 0.50 Inch (12.7 mm/day)
 Inl: 2005030300 Valld for 2005030312-2005030412



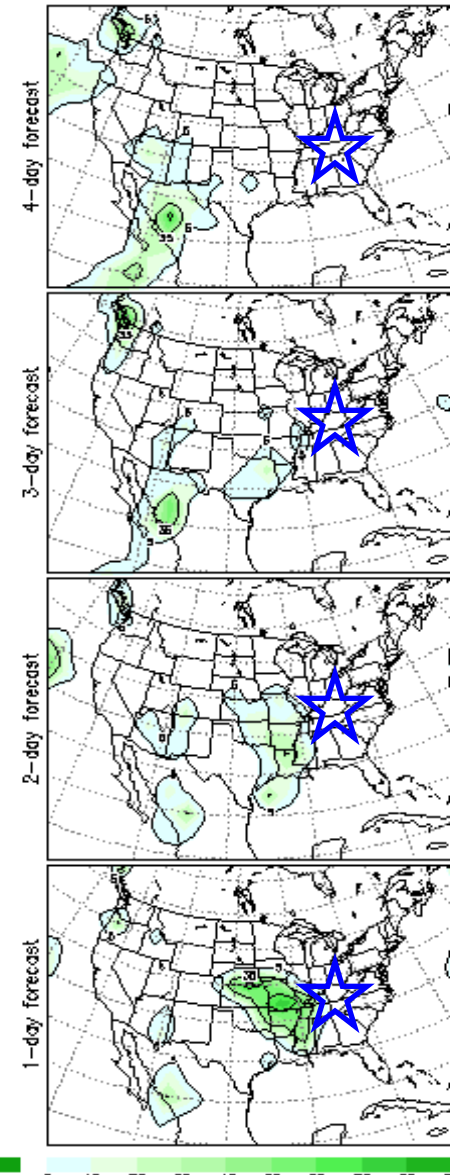
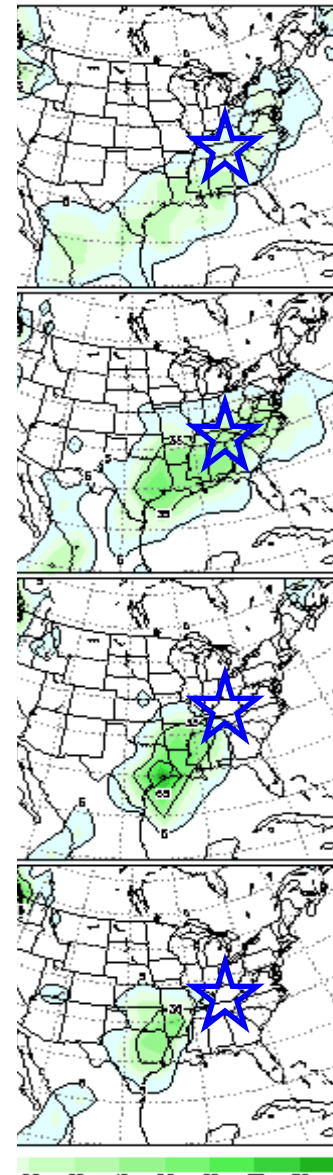
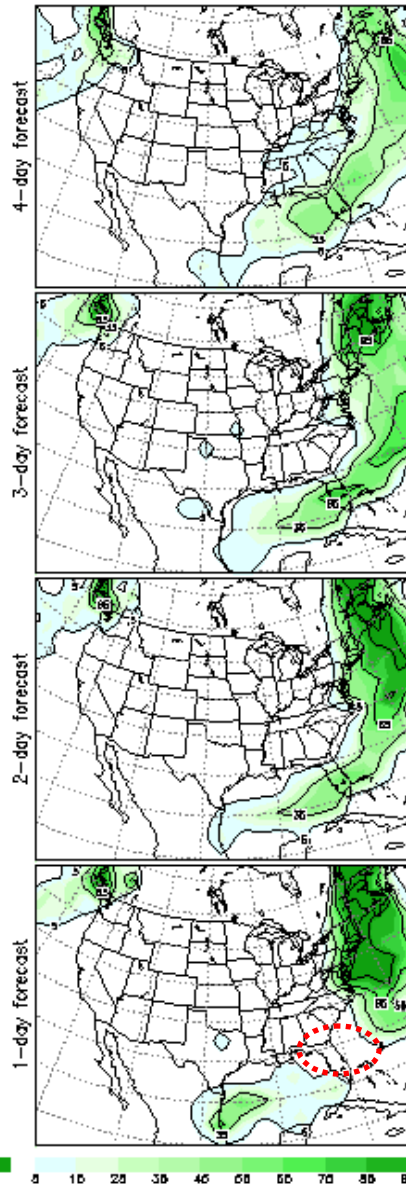
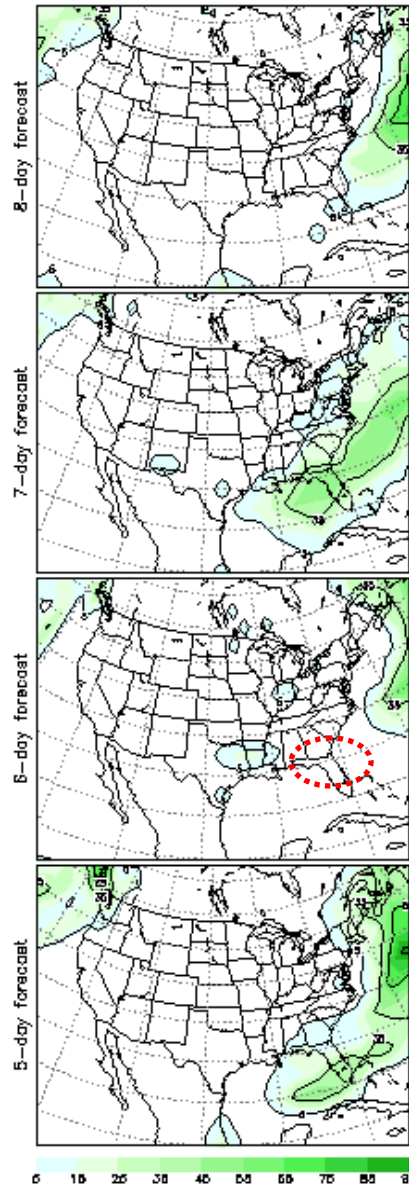
YUEJIAN ZHU, GISS/CIAC/NCEP/NOAA

YUEJIAN ZHU, GISS/CIAC/NCEP/NOAA

Good Forecasts

Ens Prob of Precip Amount Exceeding 0.50 Inch (12.7 mm/day),
 Inl: 2005030800 Valld for 2005030812-2005030912

Ens Prob of Precip Amount Exceeding 0.50 Inch (12.7 mm/day),
 Inl: 0005021200 Valld for 2005021212-2005021312

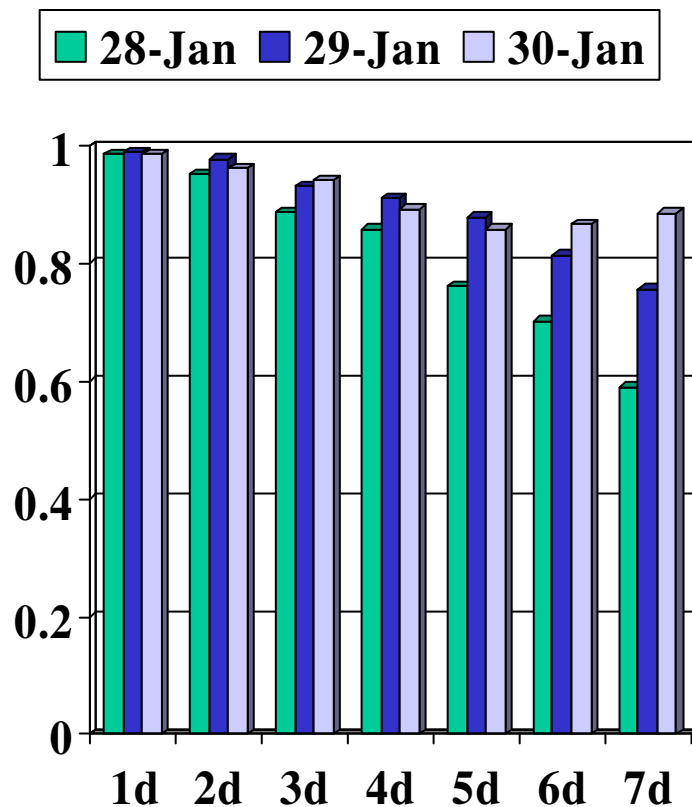


YUELIAN ZHU, GIB/CIK/NCEP/NOAA

YUELIAN ZHU, GIB/CIK/NCEP/NOAA

Poor Forecasts

Example for deterministic forecast

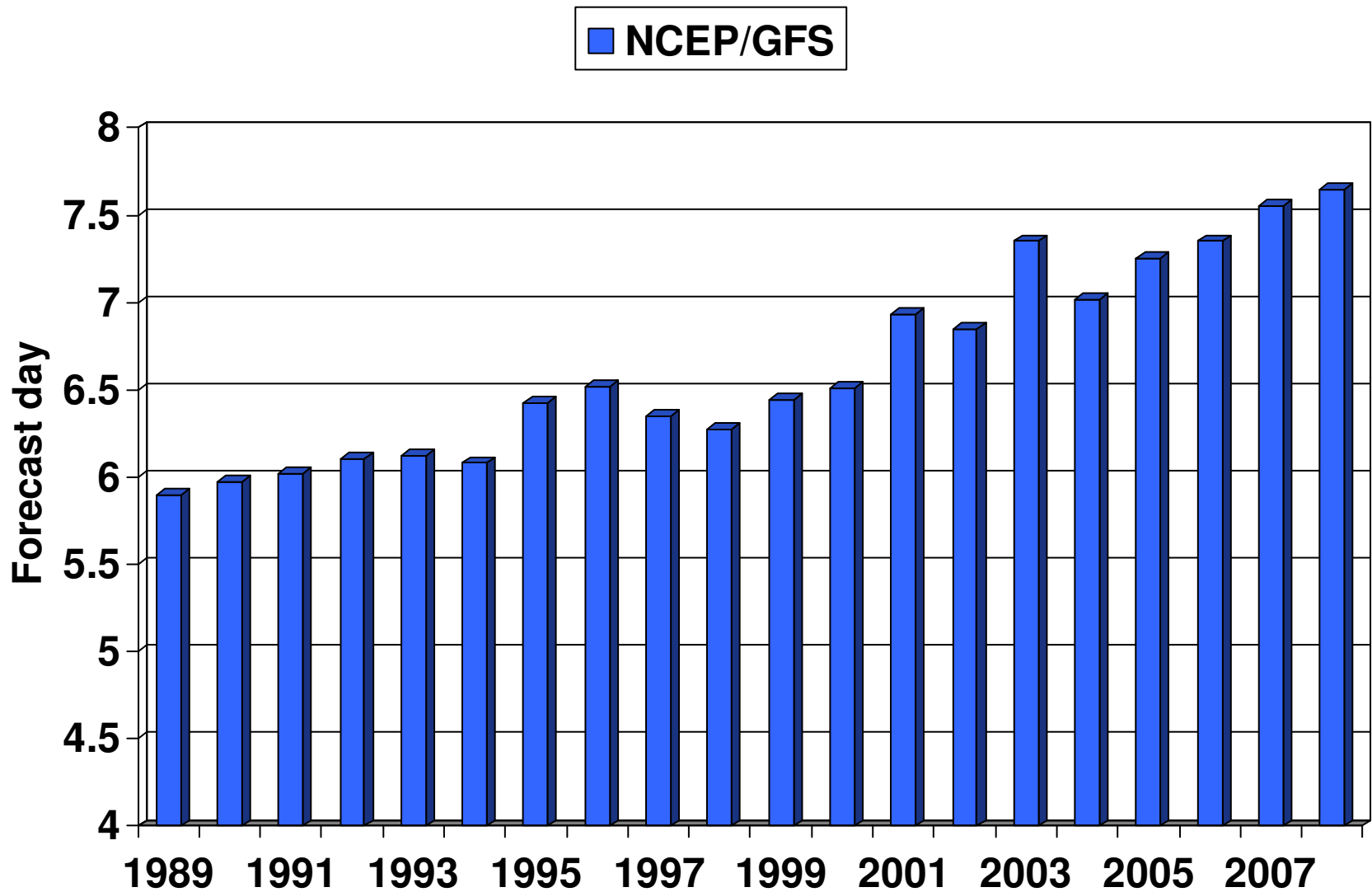


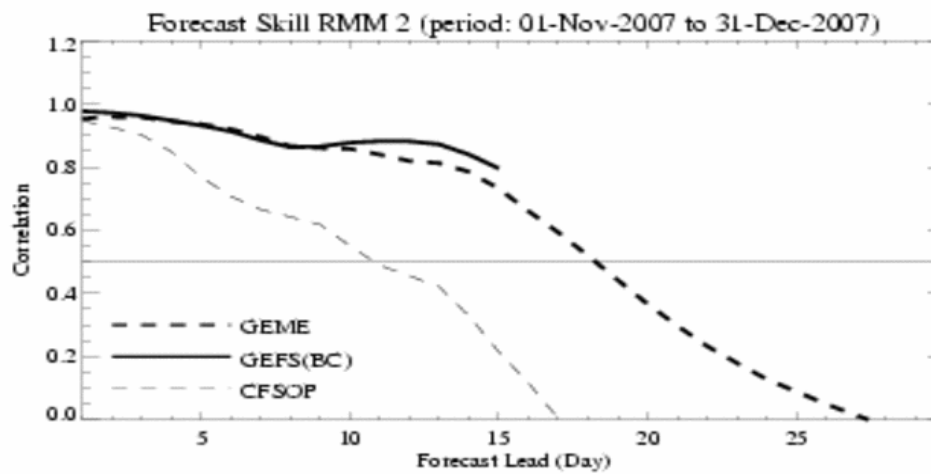
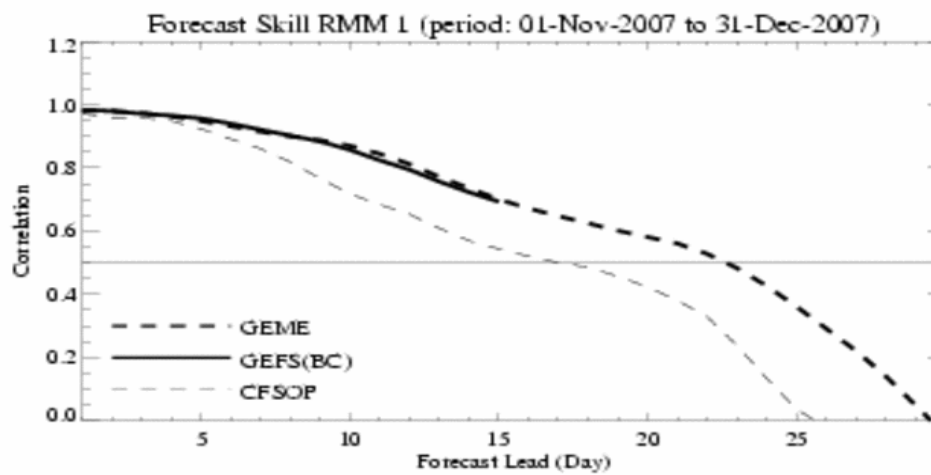
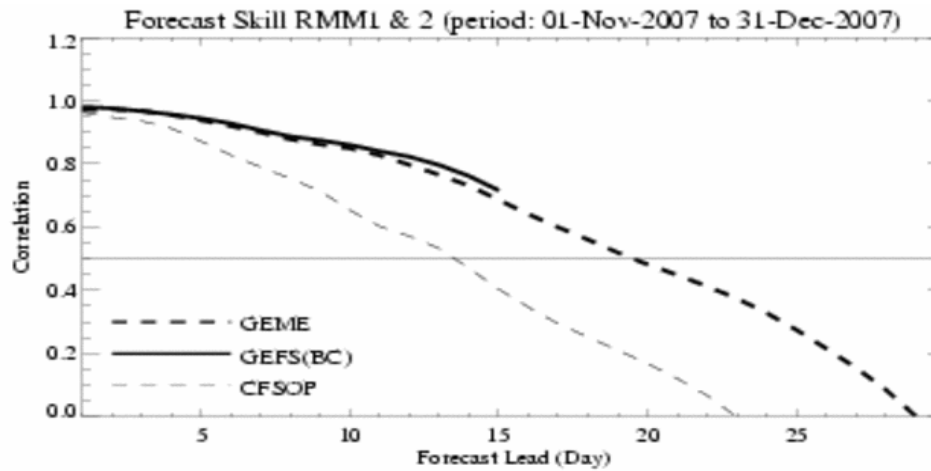
- NH 500hPa geopotential height
- PAC scores
- 1d means 24 hrs forecast .vs 48 hrs
- 2d means 48 hrs forecast against 72 hrs forecast valid at the same time, and so on

2.3 Forecast Skill

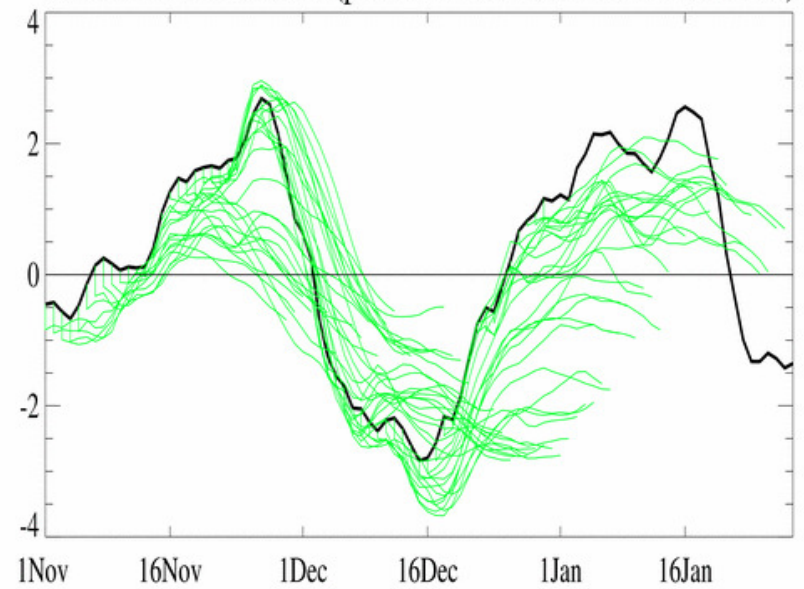
- Relative to the references (different standards)
- For med- to long-rang forecast: climatological information is a good reference
- For short-rang forecast: other high resolution forecast may be a good reference.
- Someone likes to use past forecast as a reference (past season, or same season of last year)
- Skillful forecast
 - For large scale synoptic system (60% AC)
 - For MJO prediction (50% AC or even lower)
- Other standards?

Day at which forecast loses useful skill (AC=0.6) N. Hemisphere calendar year means

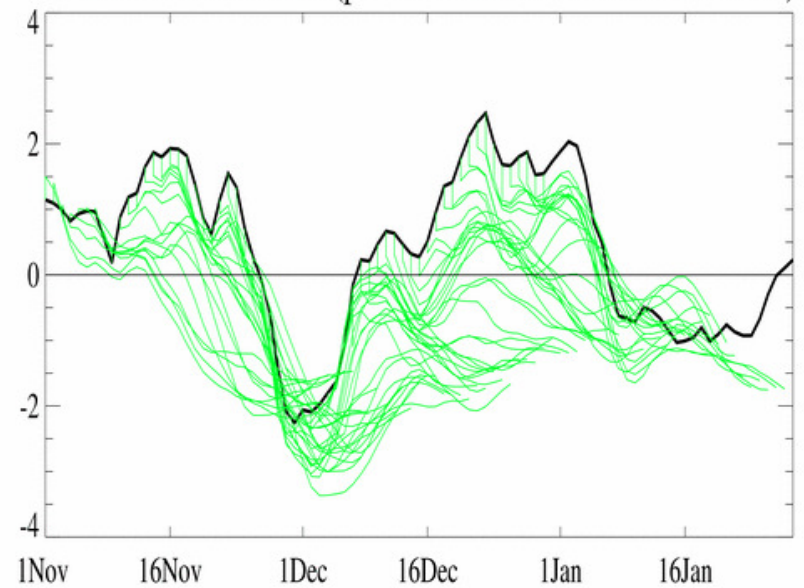




GEME Forecast RMM1 (period: 01-Nov-2007 to 31-Dec-2007)



GEME Forecast RMM2 (period: 01-Nov-2007 to 31-Dec-2007)



Courtesy of Qin Zhang

2. What do we like?

- Forecast accuracy?
 - The distance to observation/analysis
- Forecast consistency?
 - Forecast confidence
- Forecast skill?
 - Short/long term plan

3. Measurement of Deterministic Forecast

- We need:
 - Model forecast
 - Observation or analysis
 - Reference (climatology or others)
- Measurements:
 - Consistence
 - $F(t)$ against $F(t-1)$ (for same validation time)
 - Example: $t=t00z$, $t-1=t18z$ (yesterday) or $t00z$ (yesterday)
 - Accuracy and skill
 - Root Mean Square (RMS) Error
 - Anomaly Correlation (AC)
- Comparison
 - Could be a map (for case study)
 - Could be time series
 - Could be period average

Examples of measurements

1. RMS error (root mean square error)

$$RMS = \sqrt{\frac{1}{m} \sum_{i=1}^m (f(i) - a(i))^2}$$

2. ME (mean error, bias)

$$ME = \frac{1}{m} \sum_{i=1}^m (f(i) - a(i))$$

3. AE (mean absolute error)

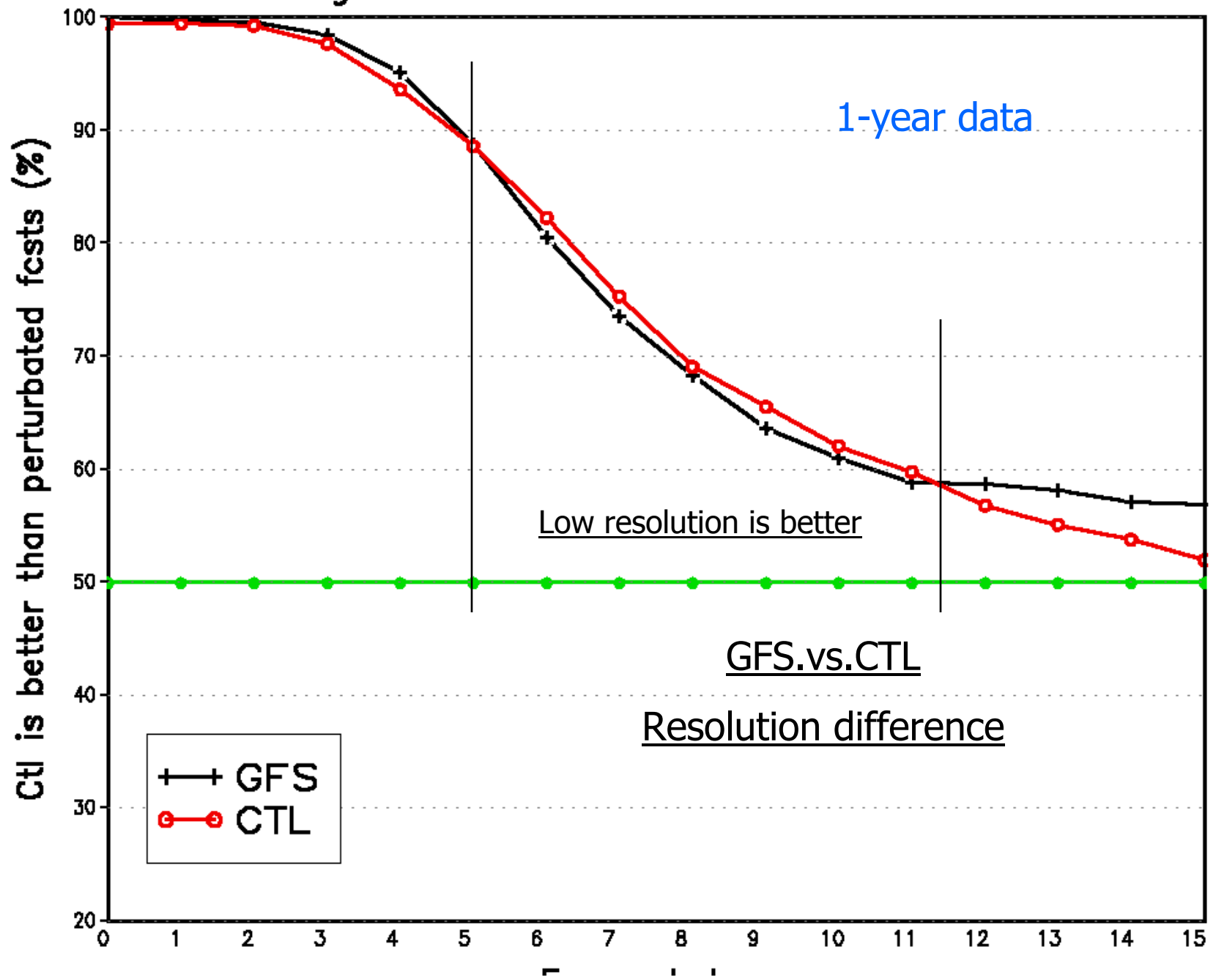
$$AE = \frac{1}{m} \sum_{i=1}^m |f(i) - a(i)|$$

4. AC (anomaly correlation)

$$AC = \frac{\frac{1}{m} \sum_{i=1}^m S_{xy}}{\sqrt{\frac{1}{m} \sum_{i=1}^m S_{xx} \cdot \frac{1}{m} \sum_{i=1}^m S_{yy}}}$$

Where f is the forecast, a is the truth (observation/analysis)

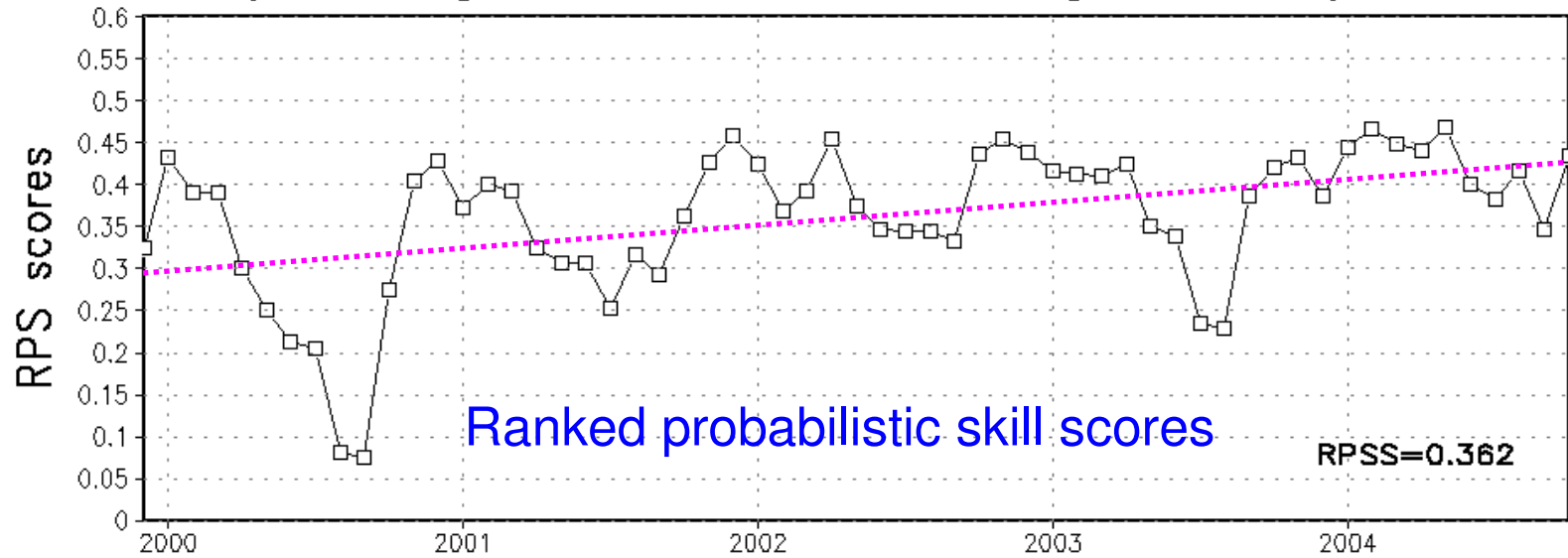
Northern Hemisphere 500 hPa Height
Average For 00Z29MAY2003 – 00Z28MAY2004



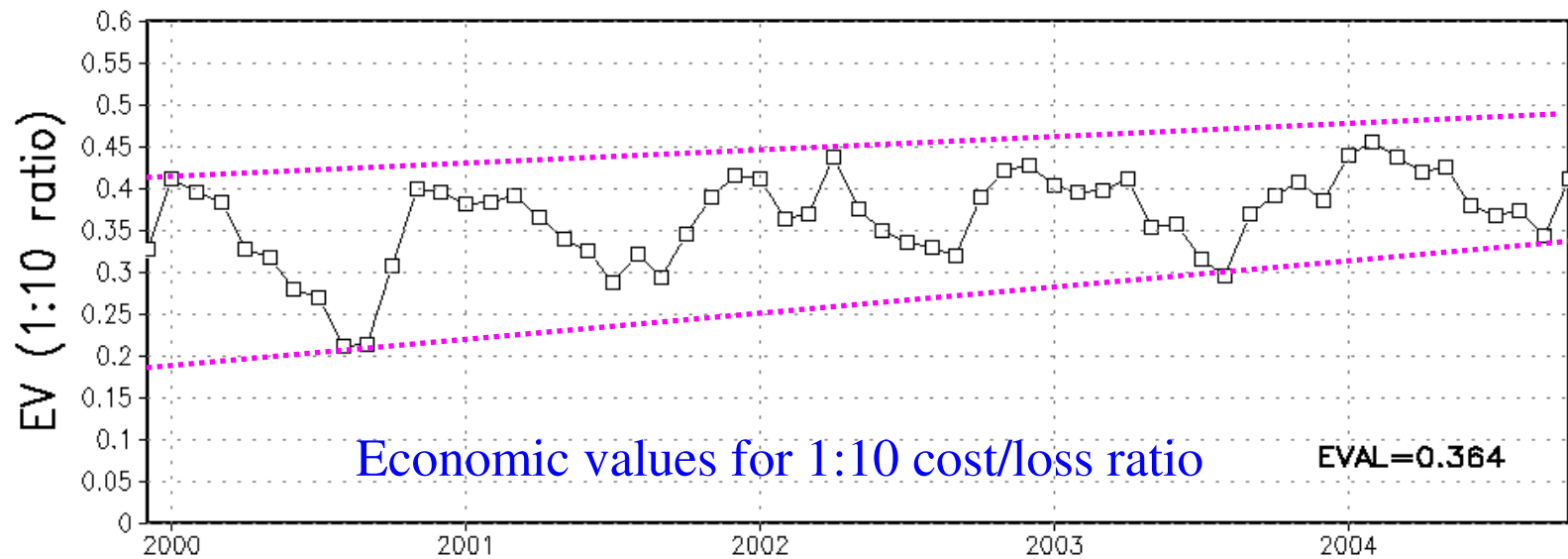
4. Measurement of Probabilistic Forecast

- We need:
 - A set of model forecast (ensemble forecast)
 - Observation or analysis
 - Reference (climatology or others)
- Measurements for ensemble mean:
 - Consistence
 - $F(t)$ against $F(t-1)$ (for same validation time)
 - Example: $t=t00z$, $t-1=t18z$ (yesterday) or $t00z$ (yesterday)
 - Performance
 - Root Mean Square (RMS) Error
 - Ensemble spread
 - Anomaly Correlation (AC)
- Measurements for ensemble distributions
 - Histogram, RPSS, CRPS, ROC, BS and etc..
- Comparison
 - Could be a map (for case study)
 - Could be time series
 - Could be period average

Monthly Average for NH 500hPa Height, 5-day forecasts



NCEP ensemble probabilistic performance for past 5-year



Ensemble Forecast

1. SPRD (ensemble spread)

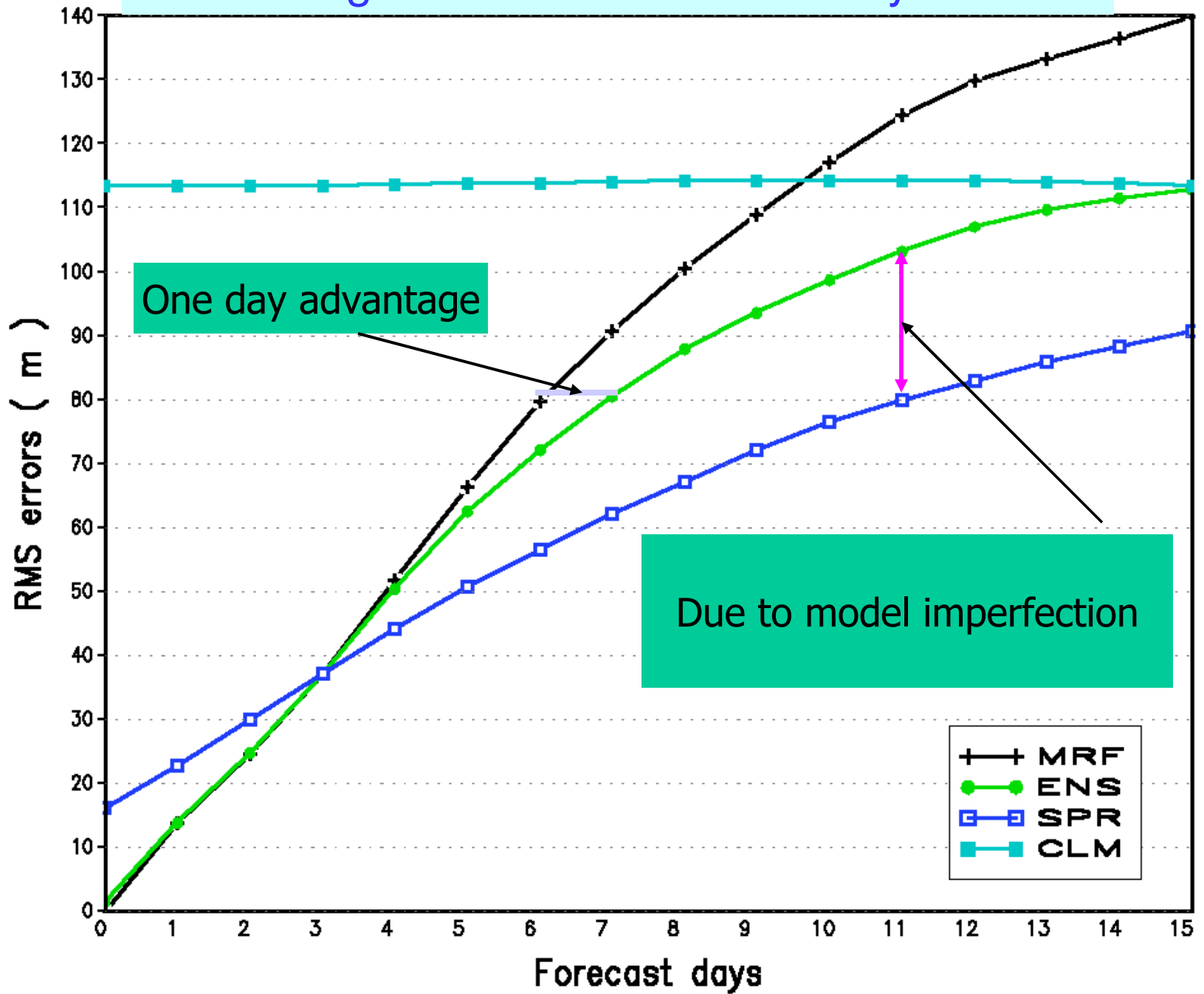
$$SPRD = \sqrt{\frac{1}{m} \sum_{i=1}^m \frac{1}{n-1} \sum_{j=1}^n (\bar{f} - f(j))^2}$$

where $\bar{f} = \frac{1}{n} \sum_{j=1}^n f(j)$ is an ensemble mean

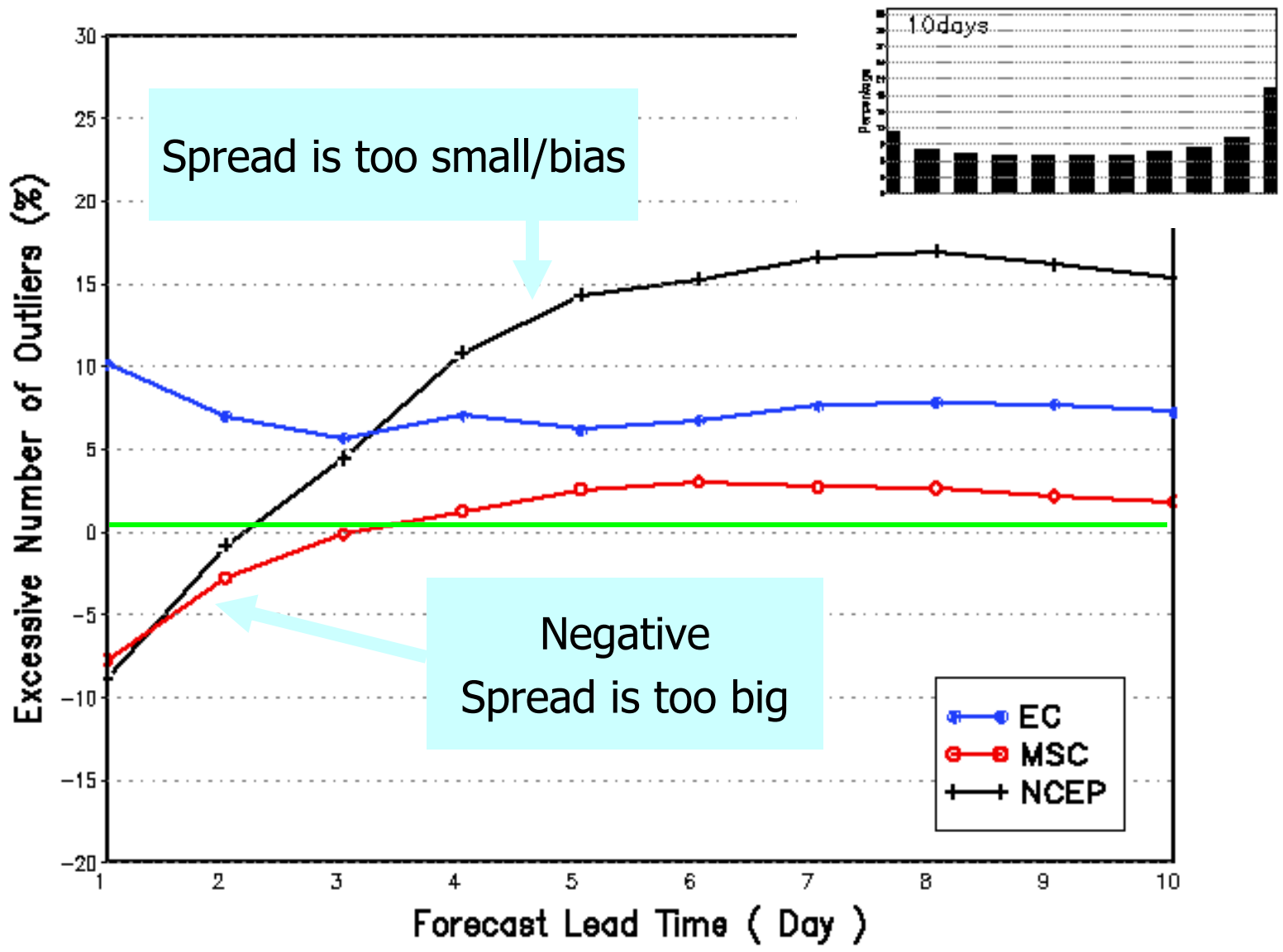
***n** is the ensemble size and **m** the sample size*

***SPRD** has the same unit as **RMSE**, it measures ensemble's uncertainty.*

What is a good ensemble forecast system?



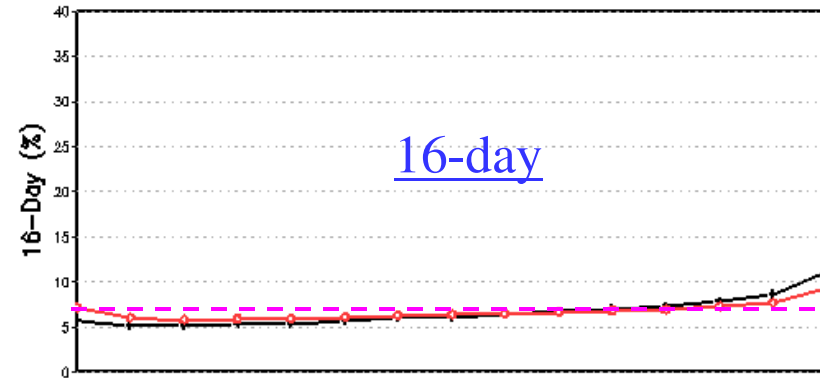
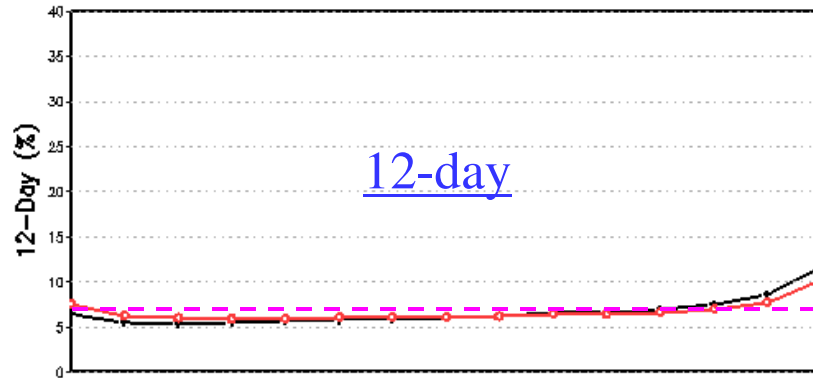
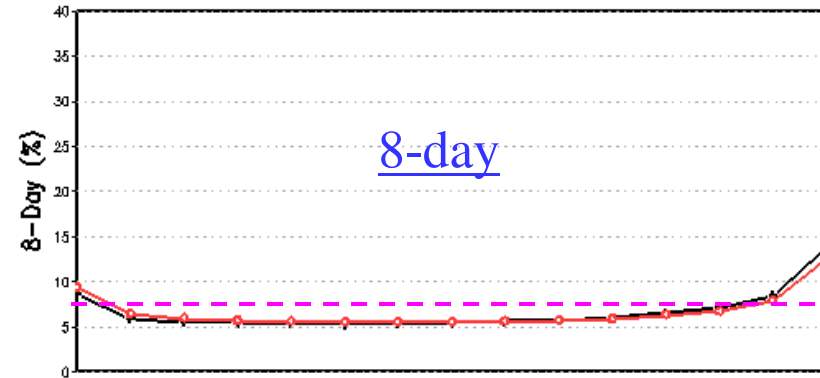
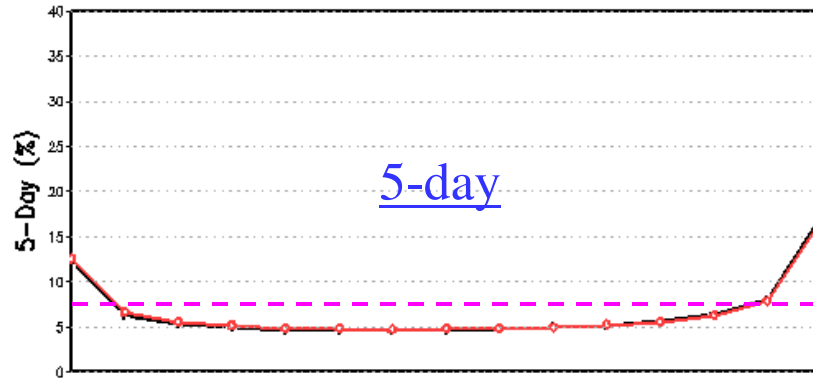
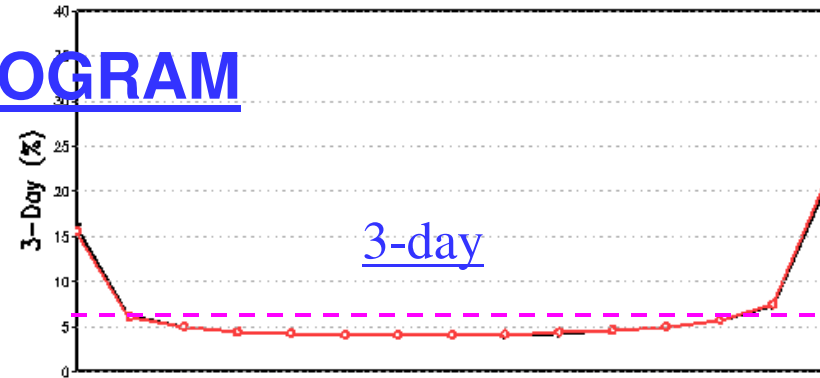
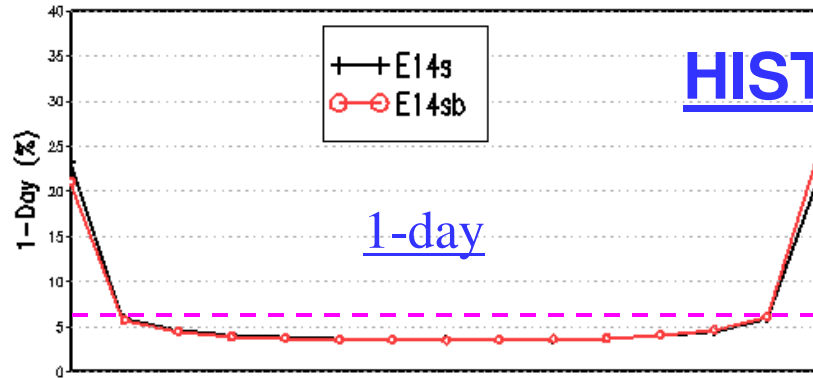
Outlier – from histogram distribution



Northern Hemisphere 2 Meter Temp. Histogram Distribution

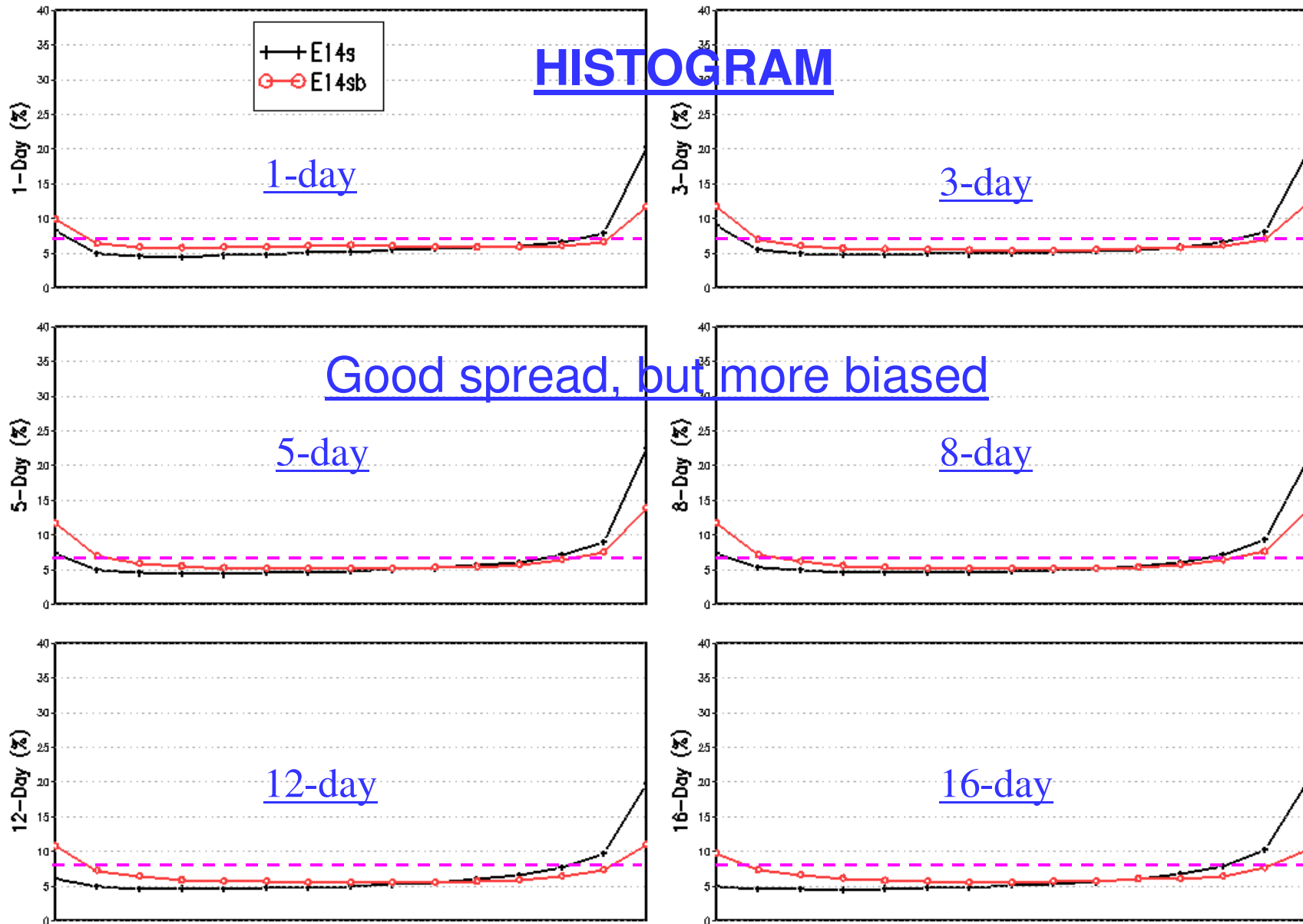
Average For 20061201 – 20070228

HISTOGRAM

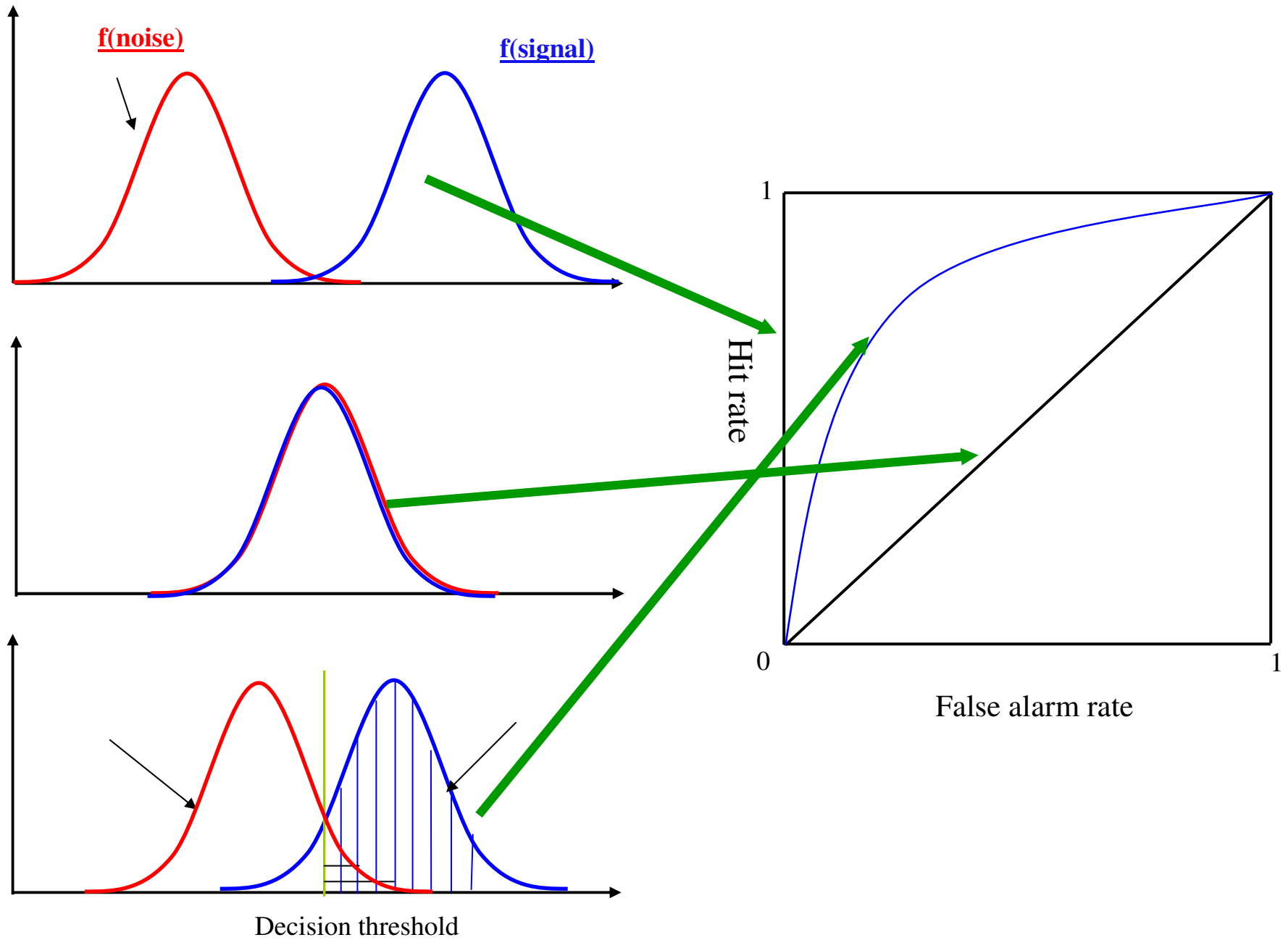


Northern Hemisphere 500hPa Height Histogram Distribution

Average For 20060601 – 20060831



Relative Operating Characteristics area (ROC area)

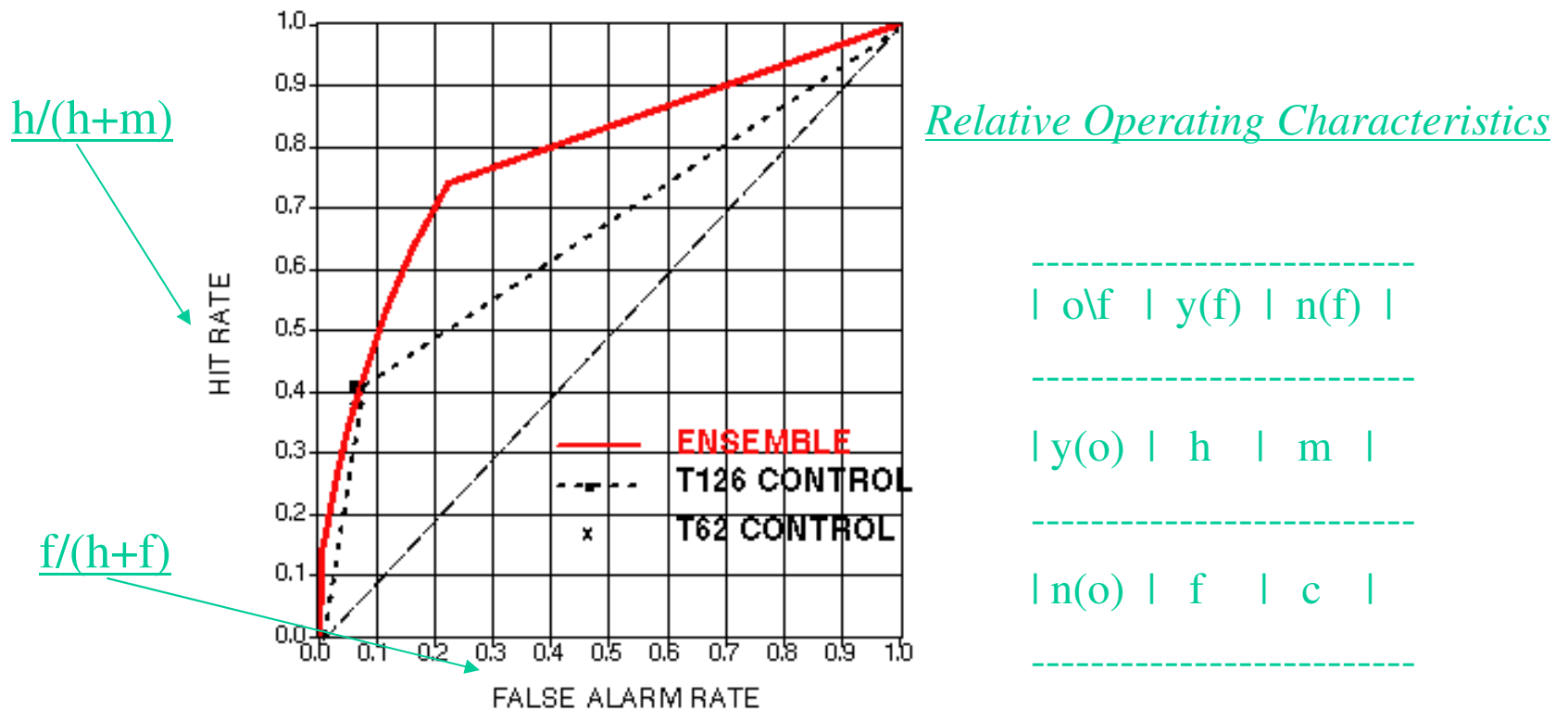


Prob. Evaluation (cost-loss analysis)

Based on hit rate (HR) and false alarm (FA) rate.

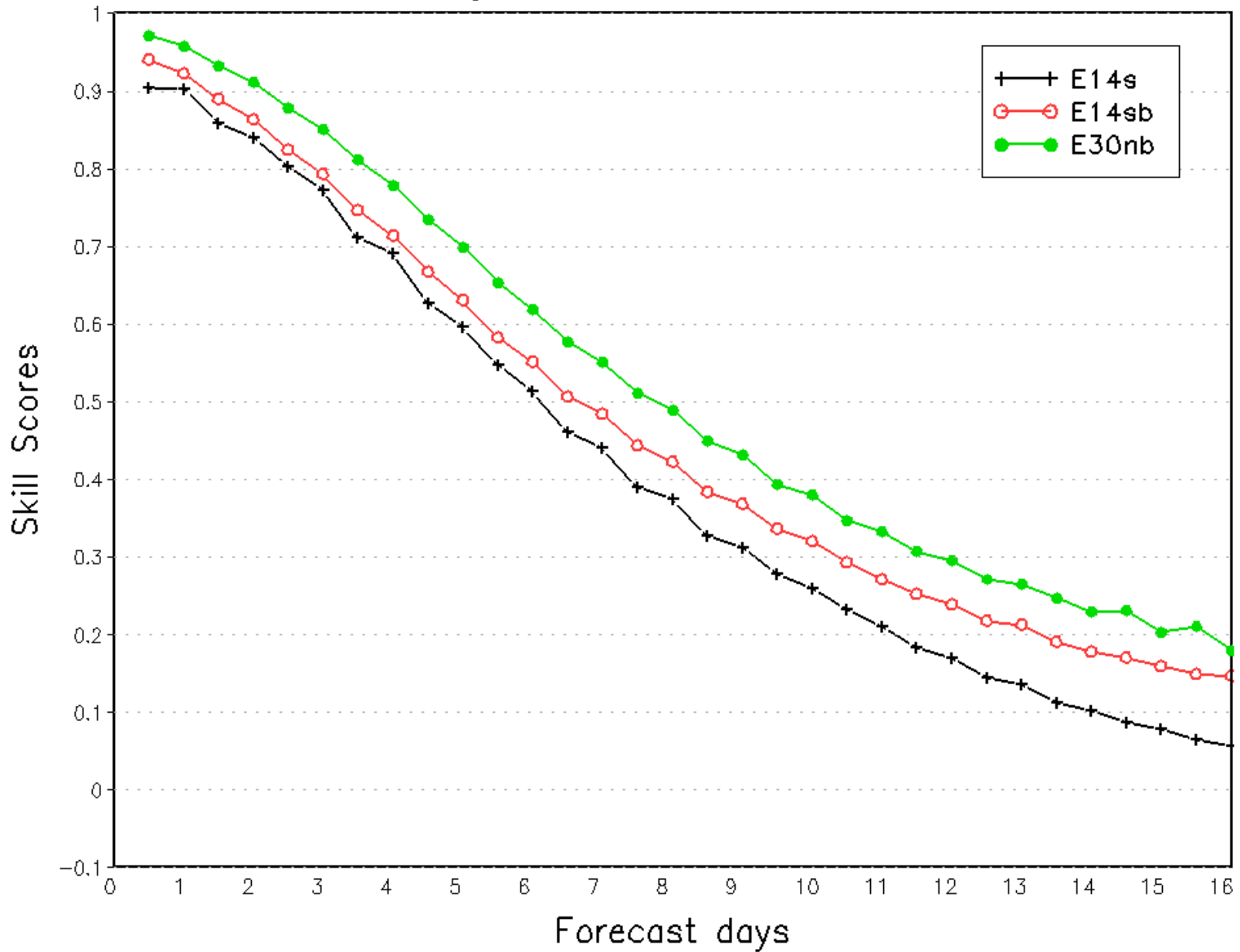
1. Relative Operating Characteristics (ROC) area - *Appl. of signal detection theory for measuring discrimination between two alternative outcome.*

$$ROC_{area} = \text{Intergrated area} * 2 \quad (0-1 \text{ normality})$$



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 occurrence of one of 10 climatologically equally likely bins) did or did not occur.

Northern Hemisphere 500hPa Height
ROC area (0-1)
Average For 20061201 - 20070228

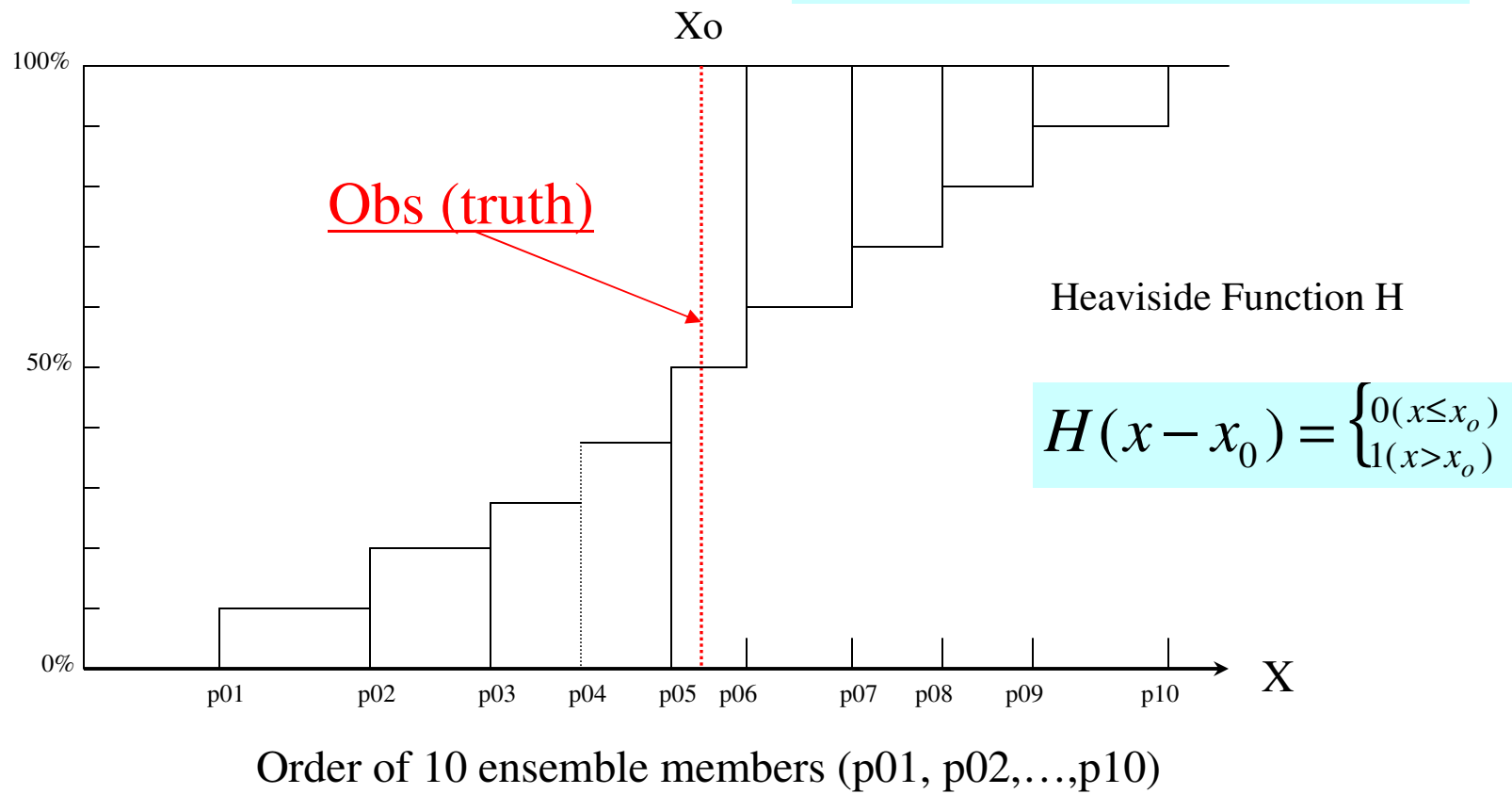


Continuous Rank Probability Score

$$CRPS = \int_{-\infty}^{+\infty} [F(x) - H(x - x_0)]^2 dx$$

CRP Skill Score is

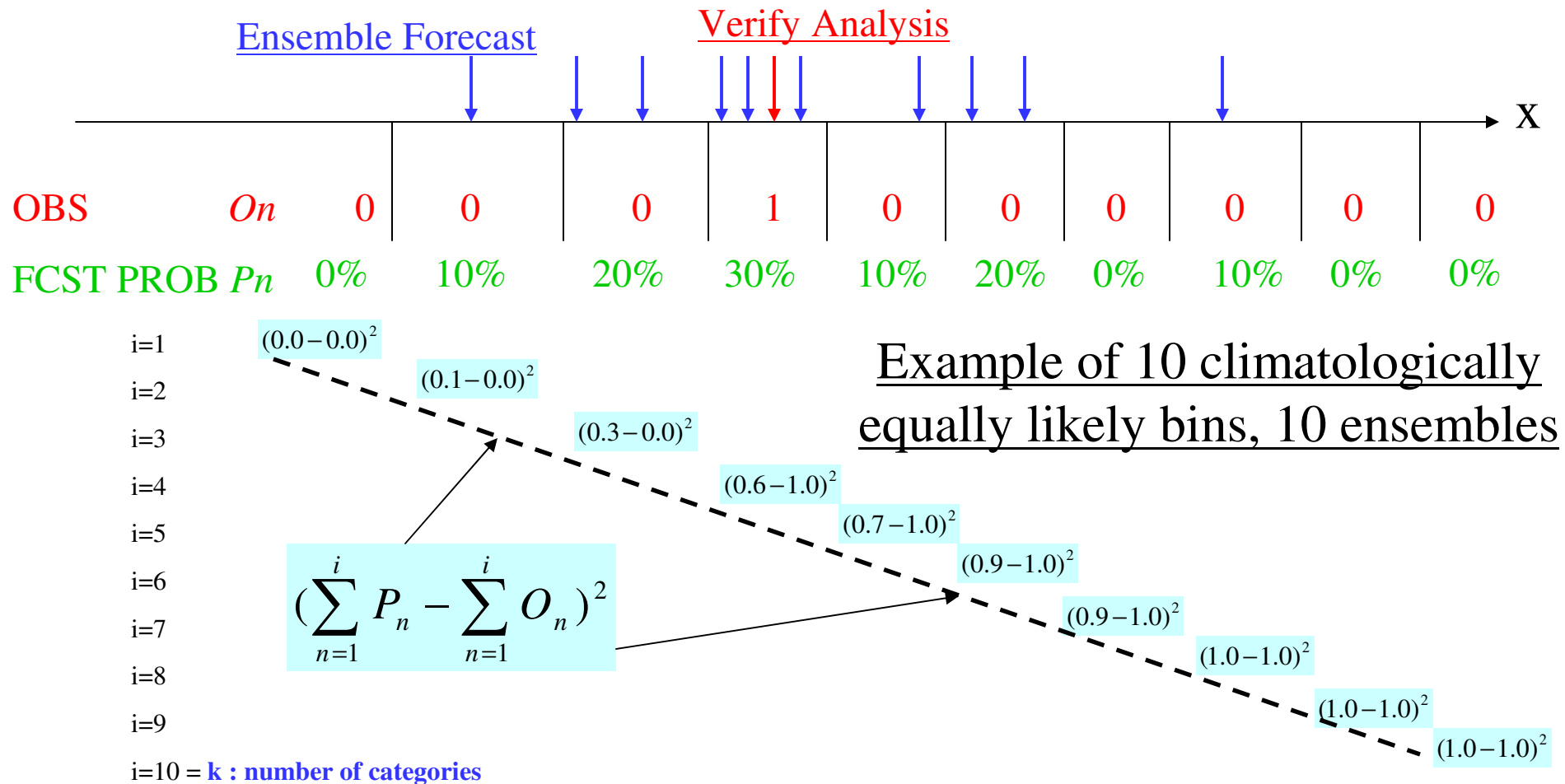
$$CRPSS = \frac{CRPS_c - CRPS_f}{CRPS_c}$$



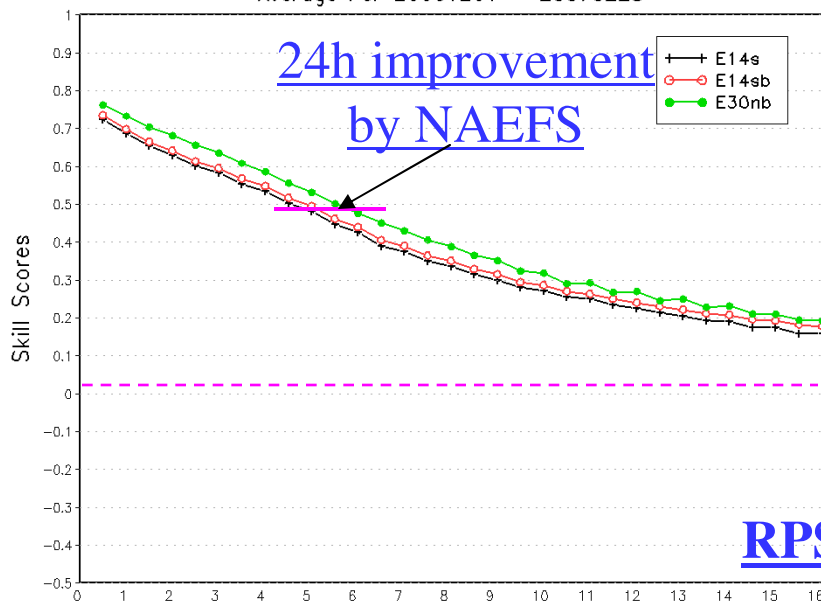
Ranked Probabilistic Score

Ranked (ordered) Probability Score (RPS) is to verify multi-category probability forecasts, to measure both reliability and resolution which based on climatologically equally likely bins

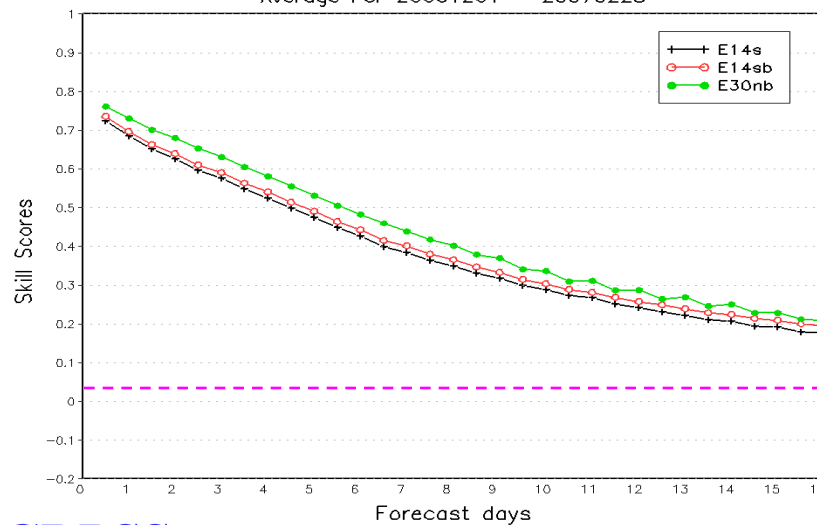
$$RPS = 1 - \frac{1}{k-1} \left[\sum_{i=1}^k \left(\sum_{n=1}^i P_n - \sum_{n=1}^i O_n \right)^2 \right] \text{ and } RPSS = \frac{RPS_f - RPS_c}{1 - RPS_c}$$



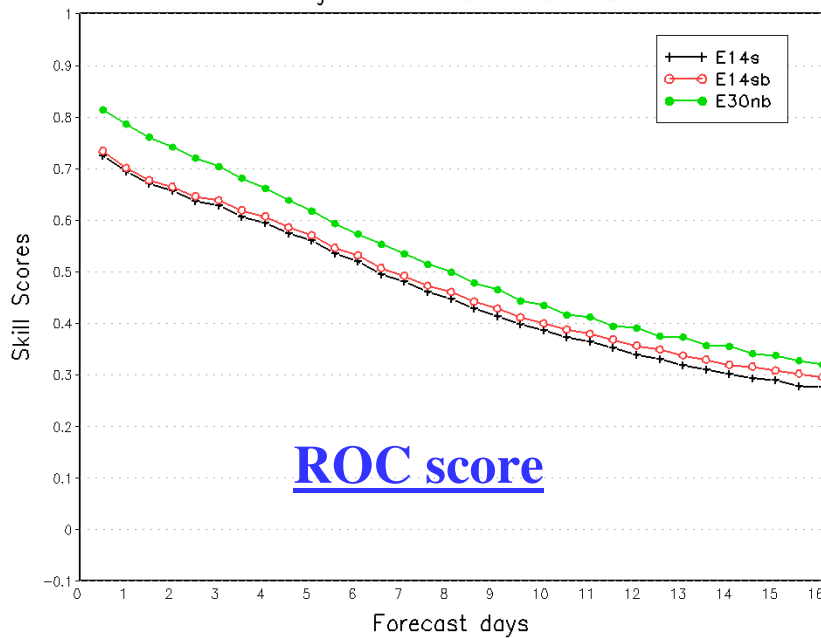
Northern Hemisphere 2 Meter Temp.
Ranked Probability Skill Scores (RPSS)
Average For 20061201 - 20070228



Northern Hemisphere 2 Meter Temp.
Continuous Ranked Probability Skill Scores
Average For 20061201 - 20070228



Northern Hemisphere 2 Meter Temp.
ROC area (0-1)
Average For 20061201 - 20070228



Winter 2006-2007

NH 2m temperature

For

NCEP raw forecast (black)

NCEP bias corrected forecast (red)

NAEFS forecast (pink)

5. Reliability and Resolution

*See <<Statistical Methods in the Atmospheric Science>> by D. S. Wilks,
Chapter 7: Forecast Verification*

1. BS (Brier Score)

$$BS = \frac{1}{n} \sum_{k=1}^n (y_k - o_k)^2$$

Where y is a forecast probability and o is an observation (probability), index k denotes a number of the n forecast event/pairs. y and o are limited from 0 to 1 in the probability sense. $BS=0$ is a perfect forecast, and $BS=1$ is missing everything

2. BSS (Brier Skill Score)

$$BSS = \frac{BS_f - BS_{ref}}{BS_{perf} - BS_{ref}} = 1 - \frac{BS_f}{BS_{ref}}$$

*ref is the reference which is mostly climatology,
 $BS_{perf}=0$ for perfect forecast, BSS is ranged from 0-1.*

Brier Score (and decomposition)

3. Algebraic Decomposition of the Brier Score

After some algebra, the Brier Score can be expressed as three separated terms

$$BS = \frac{1}{n} \sum_{i=1}^I N_i (y_k - \bar{o}_i)^2 - \frac{1}{n} \sum_{i=1}^I N_i (\bar{o}_i - \bar{o})^2 + \bar{o}(1 - \bar{o})$$

Reliability *Resolution* *Uncertainty*

↓ ↓ ↓

where $n = \sum_{i=1}^I N_i$

Conditional probability of observed and sample climatology

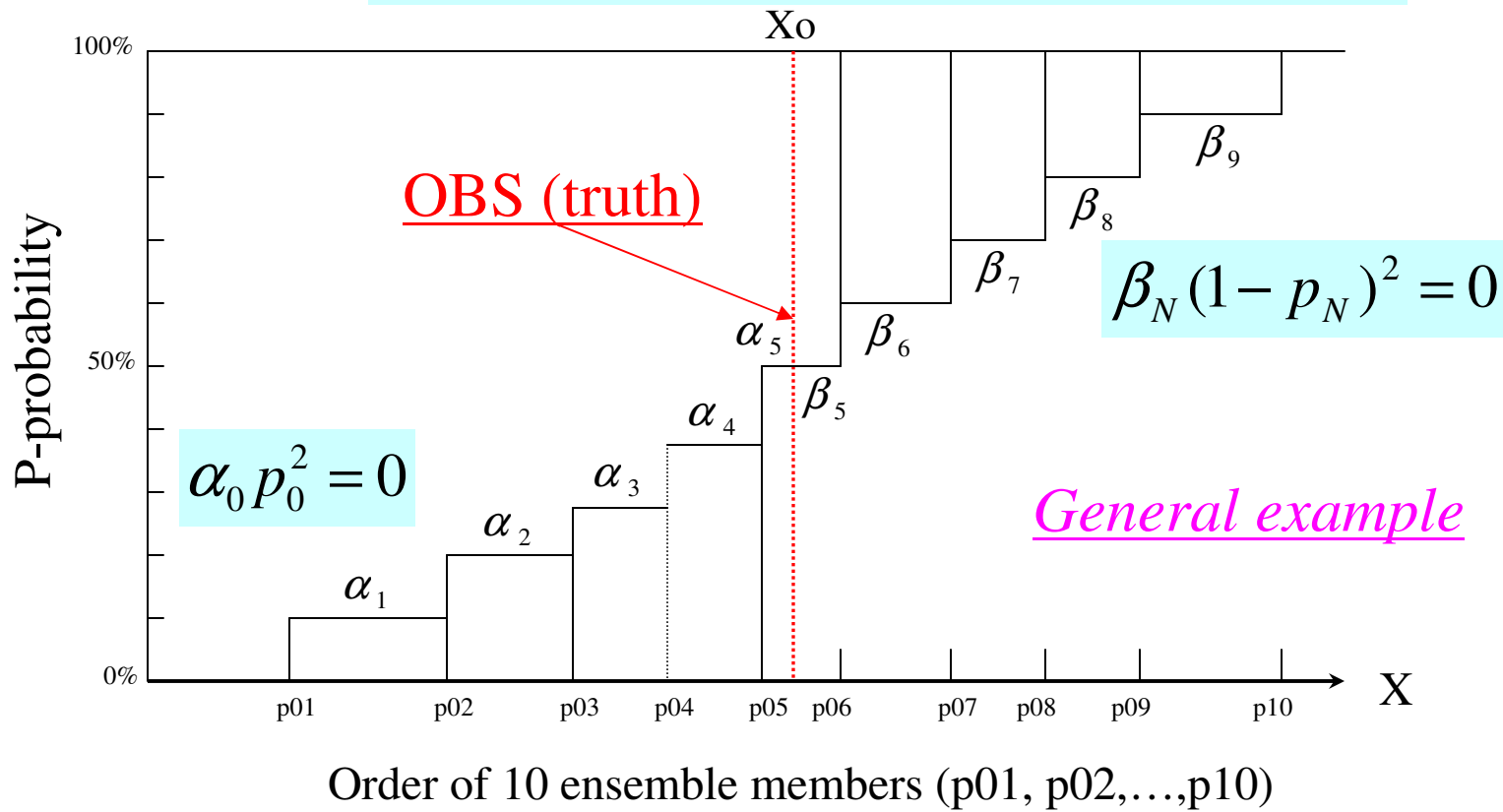
$$\bar{o}_i = p(o_1 | y_i) = \frac{1}{N_i} \sum_{k \in N_i} o_k$$

and $\bar{o} = \frac{1}{n} \sum_{k=1}^n o_k$

CRPS Decomposition

$$CRPS = \int_{-\infty}^{+\infty} [F(x) - H(x - x_0)]^2 dx$$

$$\overline{CRPS} = \sum_{i=0}^N [\overline{\alpha}_i p_i^2 + \overline{\beta}_i (1 - p_i)^2]$$



CRPS Decomposition

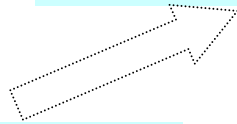
$$\overline{CRPS} = \sum_{i=0}^N [\overline{\alpha}_i p_i^2 + \overline{\beta}_i (1 - p_i)^2]$$



$$\overline{CRPS} = \sum_{i=0}^N \overline{g}_i [(1 - \overline{o}_i) p_i^2 + \overline{o}_i (1 - p_i)^2]$$



$$\overline{CRPS} = \overline{RELI} + CRPS_{pot}$$



$$\overline{RELI} = \sum_{i=0}^N \overline{g}_i (\overline{o}_i - p_i)^2$$

$$CRPS_{pot} = \overline{U} - \overline{RESO} = \sum_{i=1}^N \overline{g}_i \overline{o}_i (1 - \overline{o}_i)$$

Where:

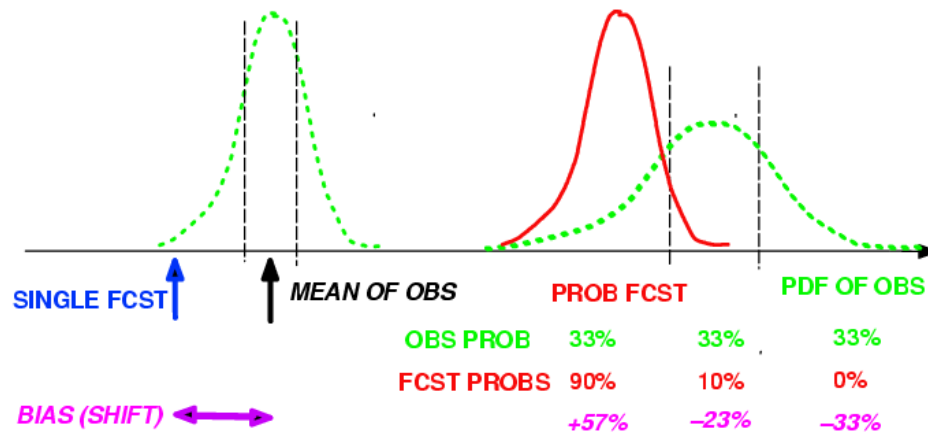
$$\overline{g}_i = \overline{\alpha}_i + \overline{\beta}_i$$

$$\overline{o}_i = \frac{\overline{\beta}_i}{\overline{\alpha}_i + \overline{\beta}_i}$$

TWO MAIN ATTRIBUTES OF FORECASTS

RELIABILITY – Lack of systematic error

(No conditional bias)



Consider cases with same forecast
 Construct pdf of corresponding obs
 If fcst identical to pdf of observations =>

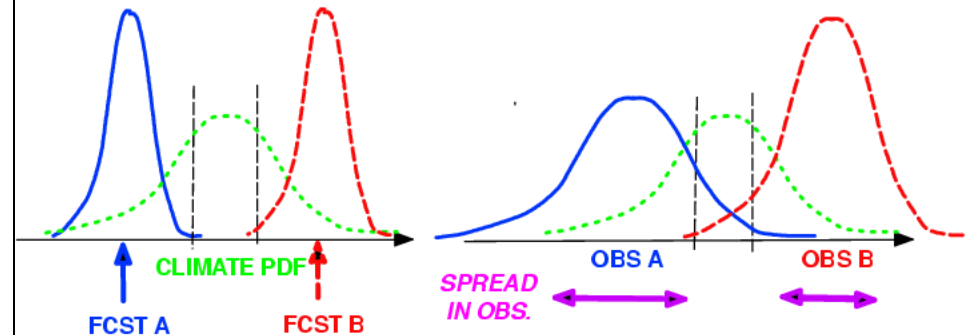
PERFECT RELIABILITY

Reliability **CAN BE** statistically corrected
 (assuming stationary processes)

Climate forecasts are perfectly reliable –

RELIABILITY IN ITSELF HAS NO FCST VALUE

RESOLUTION – Different forecasts precede different observed events



Consider different classes of fcst events
 If all observed classes are preceded by distinctly different forecasts =>

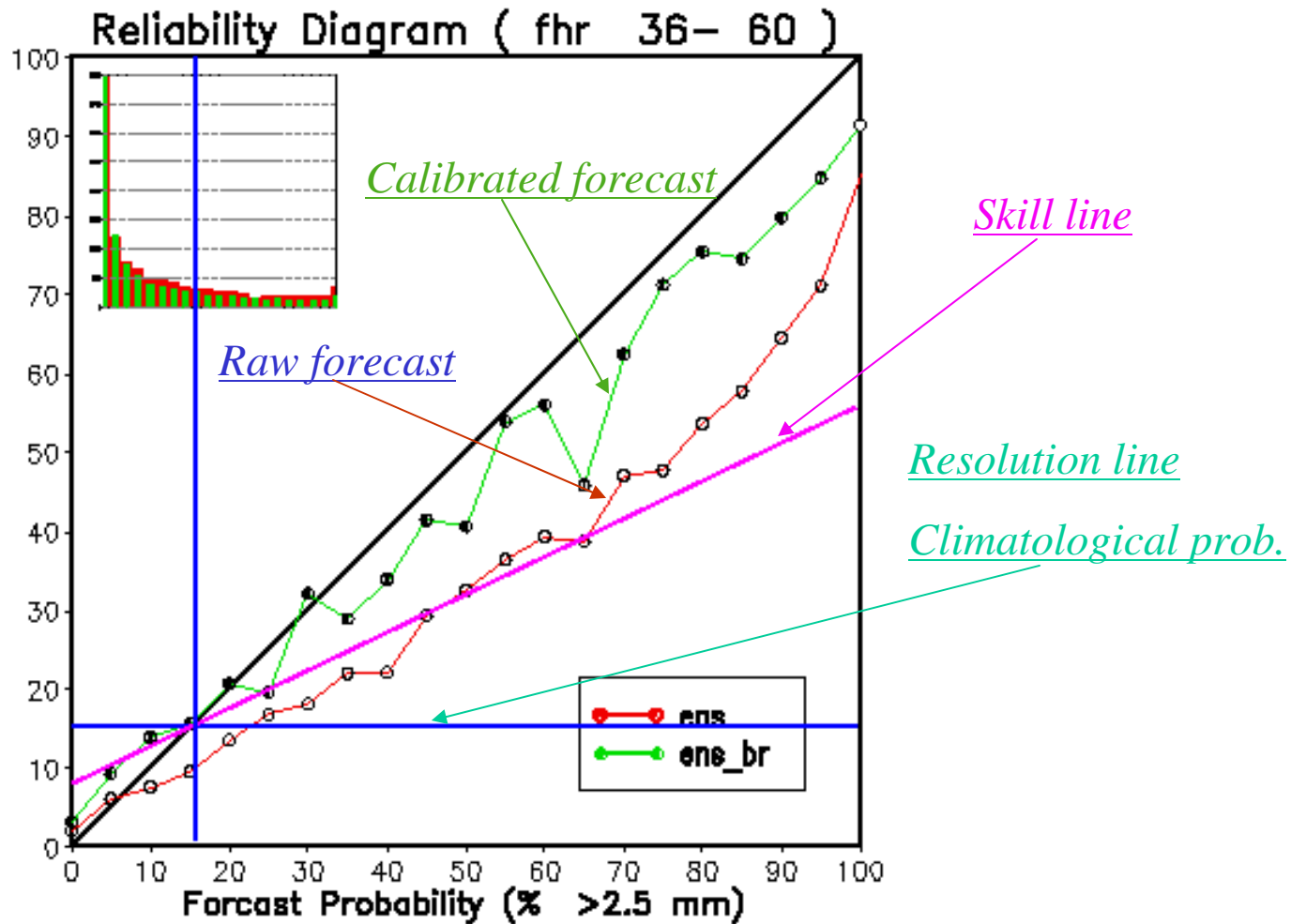
PERFECT RESOLUTION

Resolution **CANNOT BE** statistically corrected

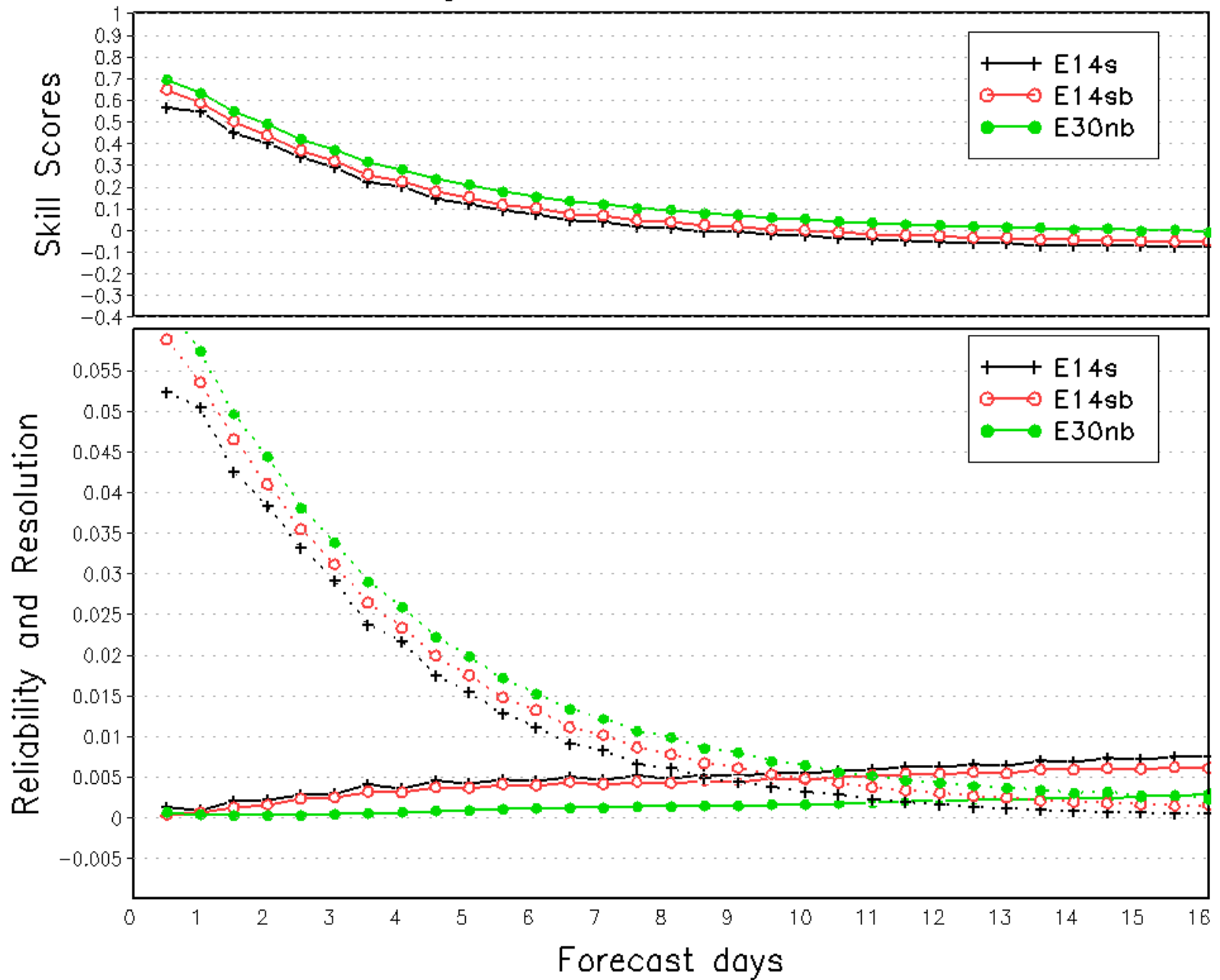
INTRINSIC VALUE OF FCST SYSTEM

Prob. Evaluation (multi-categories)

4. Reliability and possible calibration (remove bias) : For period precipitation evaluation



Northern Hemisphere 500hPa Height Brier Skill Scores (BSS) Average For 20061201 – 20070228



Category Forecast: Precipitation Evaluation

1. Frequency Bias (FBI)

$$FBI = \frac{(h + f)}{(h + m)}$$

2. ETS (equitable threat score)

$$ETS = \frac{h - R}{h + f + m - R}$$

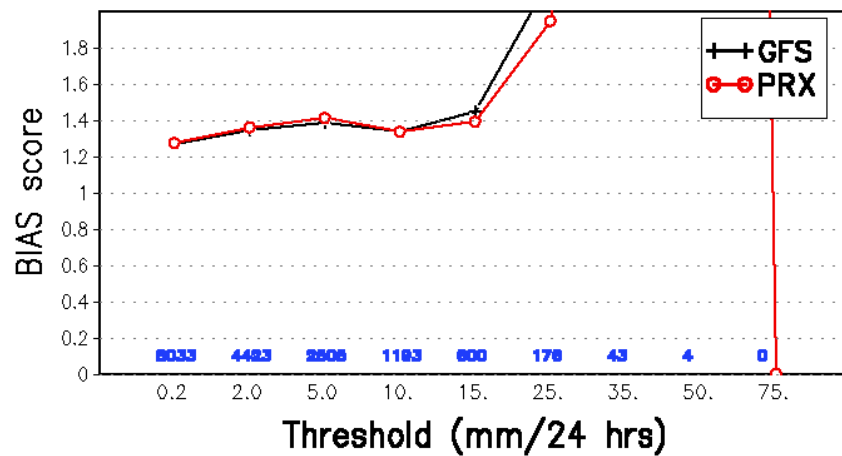
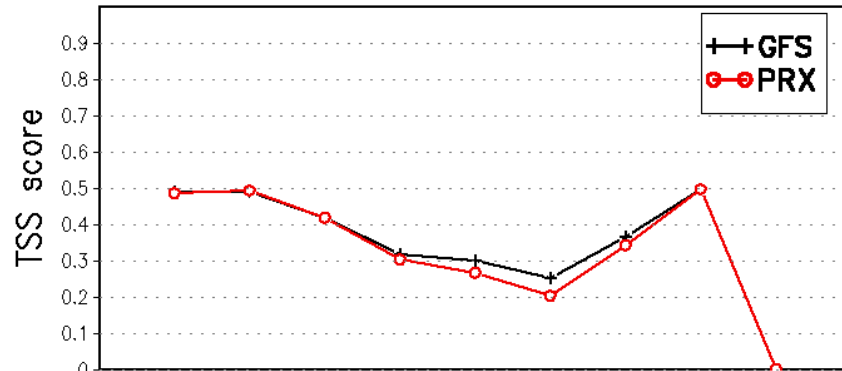
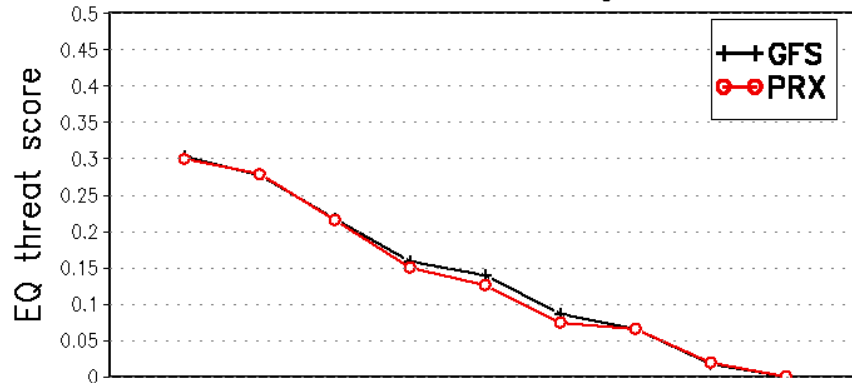
where $R = \frac{(h+f) \cdot (f+m)}{(h+f+m+c)}$ is randomly forecasting rate

	<i>OBS</i> (Yes)	<i>OBS</i> (No)
<i>FCST</i> (Yes)	<i>Hit</i> (<i>h</i>)	<i>False alarm</i> (<i>f</i>)
<i>FCST</i> (No)	<i>Miss</i> (<i>m</i>)	<i>Correct Reject</i> (<i>c</i>)

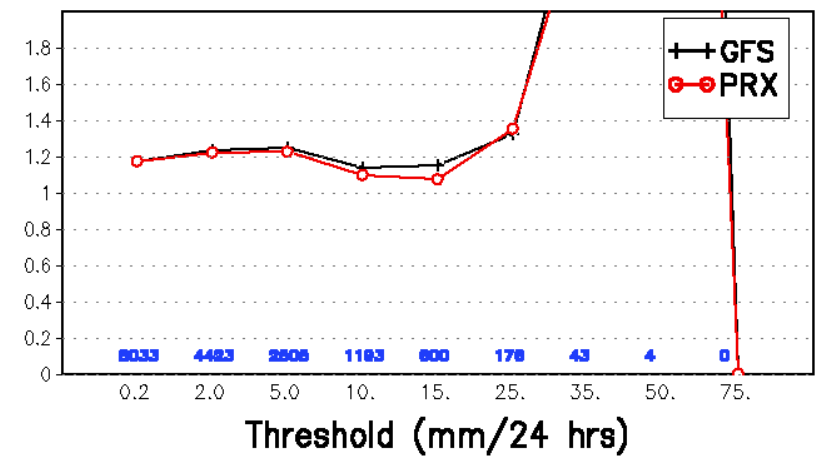
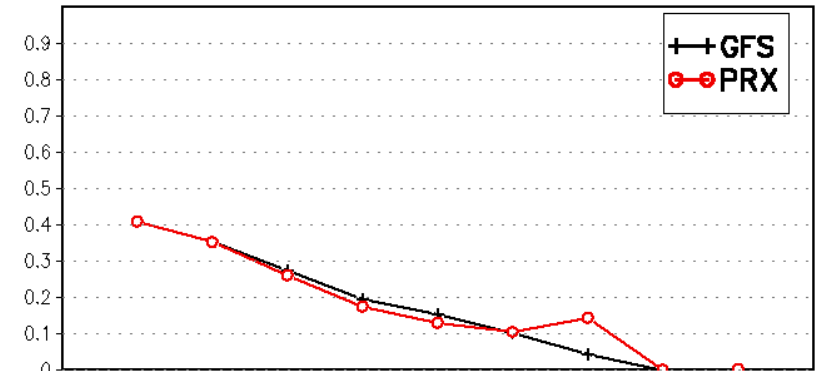
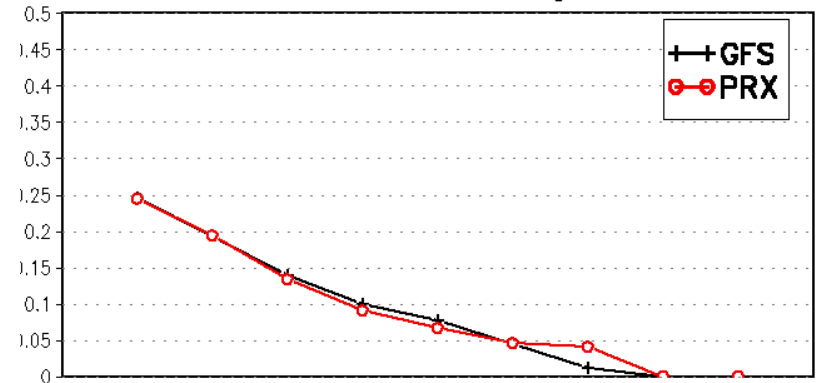
3. TSS (true skill statistic or Hanssen-Kuipers discriminant)

$$TSS = \frac{h \cdot c - f \cdot m}{(h+m) \cdot (f+c)} = \frac{h}{h+m} - \frac{f}{f+c}$$

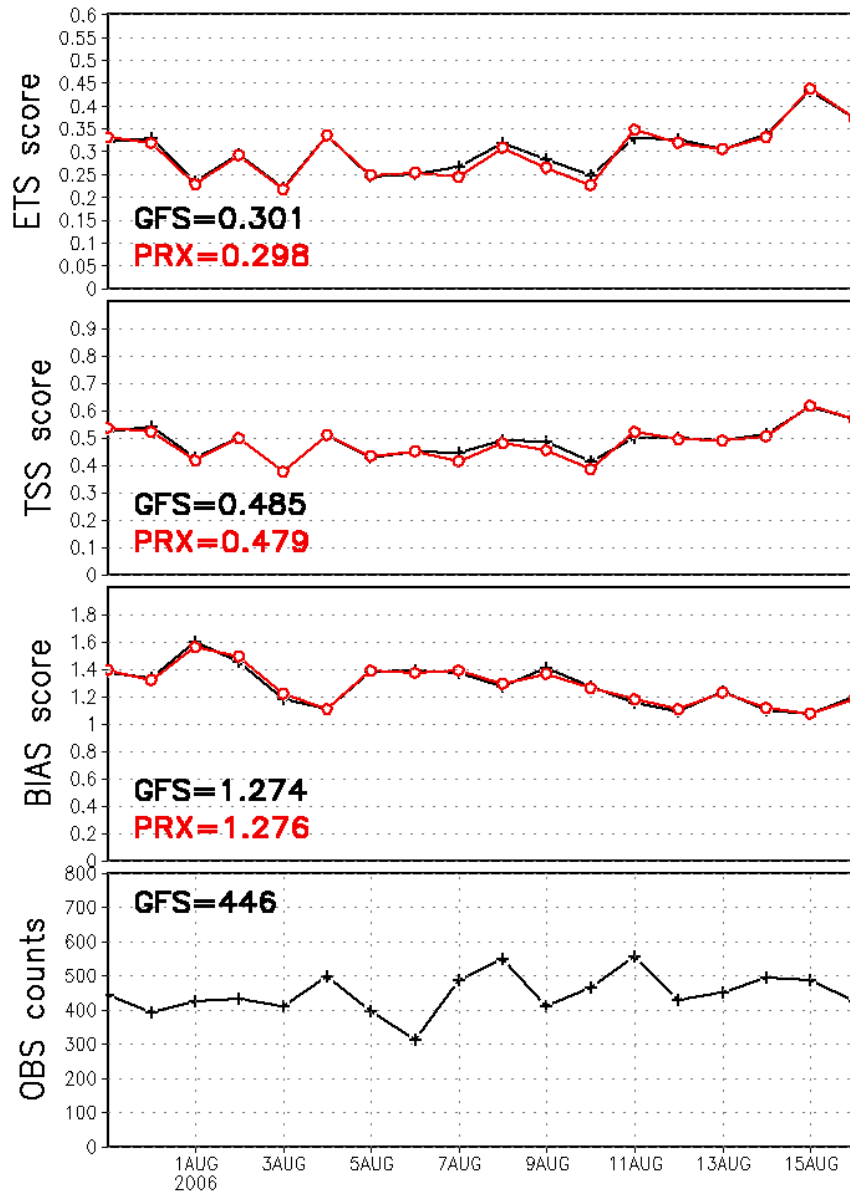
North America
00Z30JUL2006 – 00Z16AUG2006
12–36 hrs average



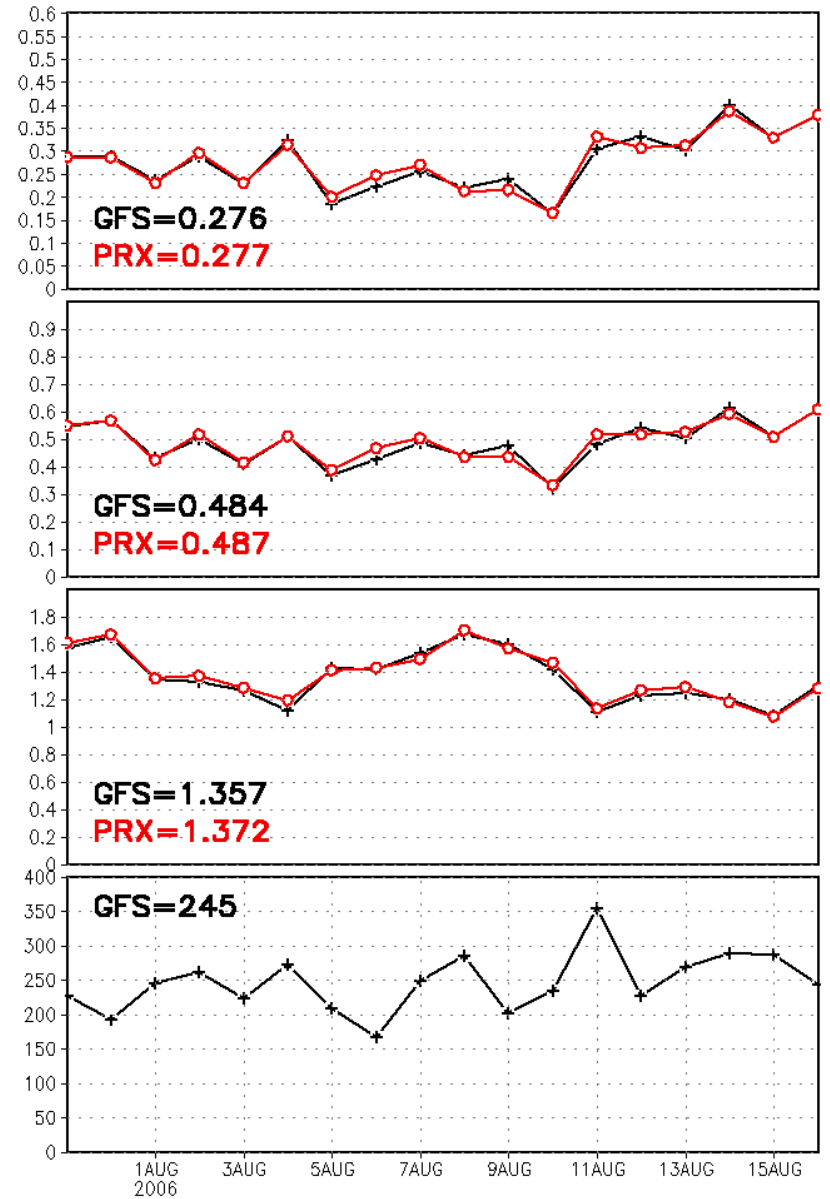
North America
00Z30JUL2006 – 00Z16AUG2006
60–84 hrs average



North America
 00Z30JUL2006 - 00Z16AUG2006
 (12-36) hrs avg (Threshold ≥ 0.2 mm/24 hrs)



North America
 00Z30JUL2006 - 00Z16AUG2006
 (12-36) hrs avg (Threshold ≥ 2.0 mm/24 hrs)



Economic Value of Forecast

TABLE. Contingency table indicating the costs and losses accrued by the use of weather forecasts, depending on forecast and observed events.

	<i>Yes (FCST)</i>	<i>No (FCST)</i>
<i>Yes (OBS)</i>	<i>Hit (h)</i> <i>Mitigated Loss (C+Lu)</i>	<i>Miss (m)</i> <i>Loss (L=Lp+Lu)</i>
<i>No (OBS)</i>	<i>False Alarm (f)</i> <i>Cost (C)</i>	<i>Correct Reject (c)</i> <i>No Cost (N)</i>

Zhu and etc.. 2002: BAMS

1. Expected Expense:

$$E_{forecast} = h(C + L_u) + fC + m(L_p + L_u)$$

$$E_{climate} = \text{Min}[oL, o(C + L_u) + (1 - o)C]$$

$$E_{perfect} = o(C + L_u)$$

2. Economic Value:

$$V = \frac{E_{climate} - E_{forecast}}{E_{climate} - E_{perfect}}$$

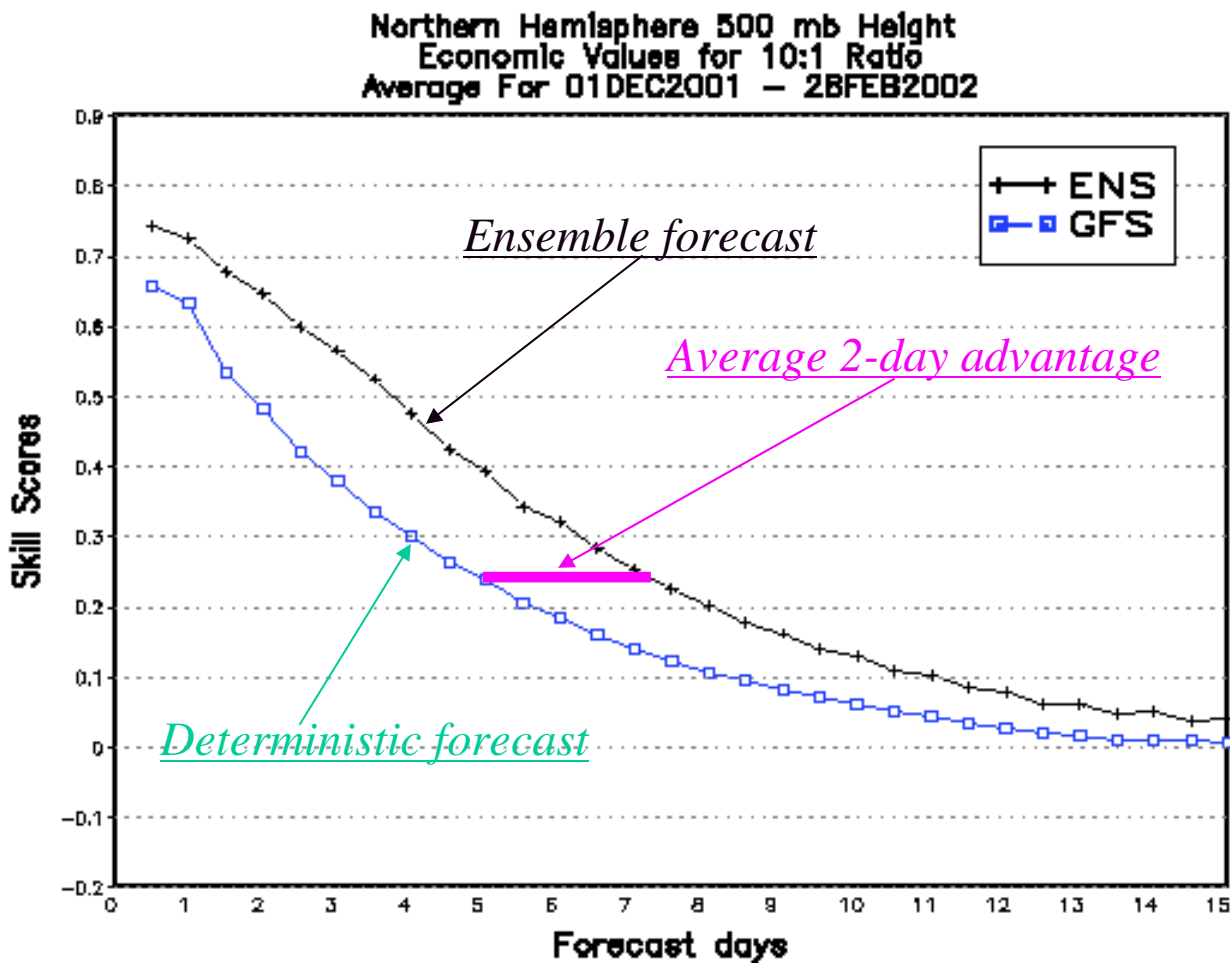
or

$$V = \frac{\text{Min}[o, r] - (h + f)r - m}{\text{Min}[o, r] - or}$$

Where o is the climatological frequency of the event (or o=h+m), r=C/Lp which is the ratio of the cost of protection to the amount of potential loss that can be protected

Prob. Evaluation (cost-loss analysis)

Based on hit rate (HR) and false alarm (FA) analysis
.. Economic Value (EV) of forecasts

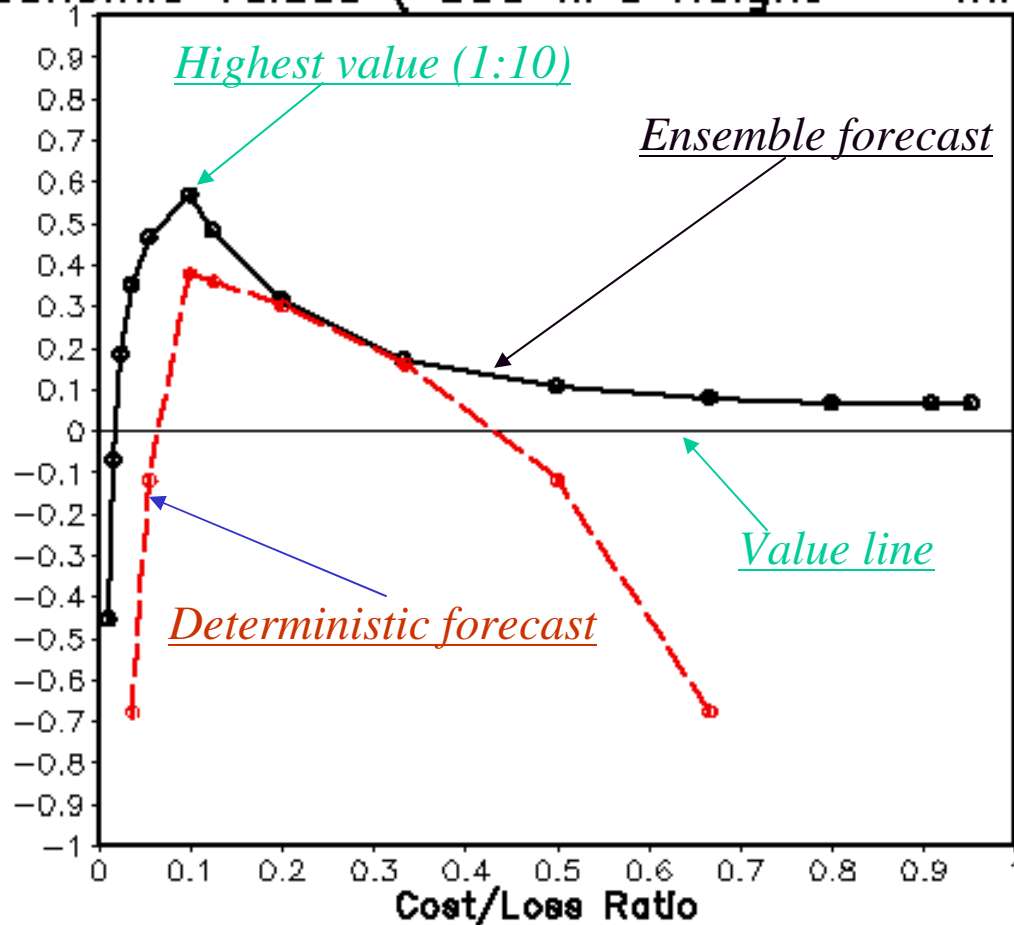


Prob. Evaluation (cost-loss analysis)

2. Economic Value (EV) of forecasts.

Given a particular forecast, a user either does or does not take action

Economic Values (500 hPa Height -- fhr 72)



Decision Theory Example

Critical Event: sfc winds > 50kt

Cost (of protecting): \$150K

Loss (if damage): \$1M

		Forecast?	
		YES	NO
Observed?	YES	<i>Hit</i> \$150K	<i>Miss</i> \$1000K
	NO	<i>False Alarm</i> \$150K	<i>Correct Rejection</i> \$0K

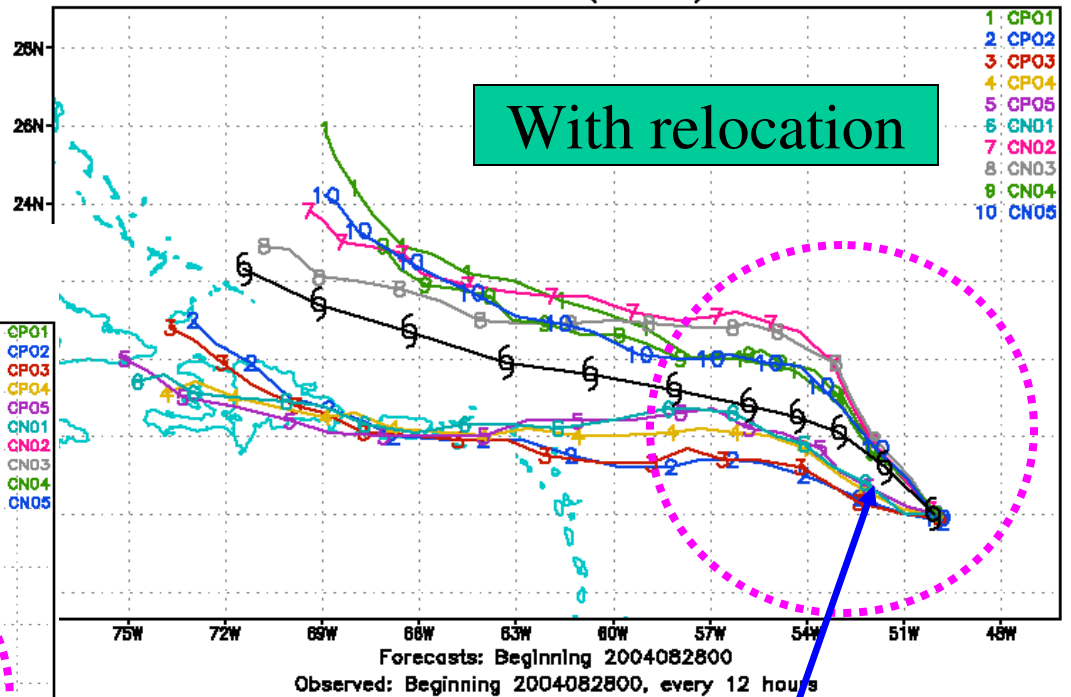
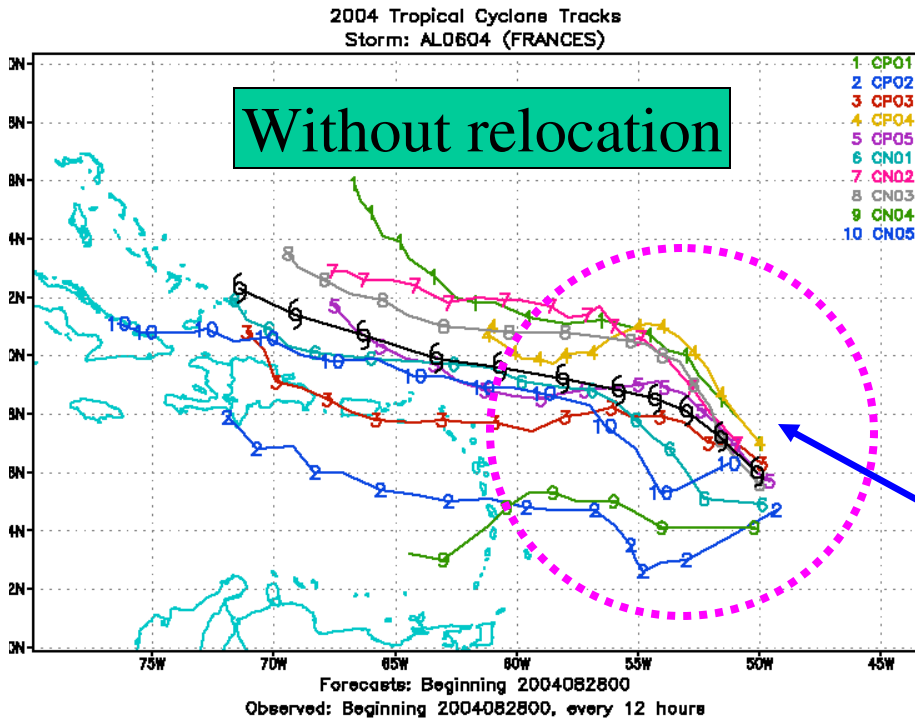
Case	Deterministic	Observation	Cost (\$K)	Probabilistic	Cost (\$K) by Threshold for Protective Action					
	Forecast (kt)	(kt)		Forecast	0%	20%	40%	60%	80%	100%
1	65	54	150	42%	150	150	150	1000	1000	1000
2	58	63	150	71%	150	150	150	150	1000	1000
3	73	57	150	95%	150	150	150	150	150	1000
4	55	37	150	13%	150	0	0	0	0	0
5	39	31	0	3%	150	0	0	0	0	0
6	31	55	1000	36%	150	150	1000	1000	1000	1000
7	62	71	150	85%	150	150	150	150	150	1000
8	53	42	150	22%	150	150	0	0	0	0
9	21	27	0	51%	150	150	150	0	0	0
10	52	39	150	77%	150	150	150	150	0	0
Total Cost:			\$ 2,050		\$1,500	\$1,200	\$1,900	\$2,600	\$3,300	\$5,000

Optimal Threshold = 15%

Hurricane Track Plots (case 1)

Frances (08/28)

2004 Tropical Cyclone Tracks
Storm: AL0604 (FRANCES)

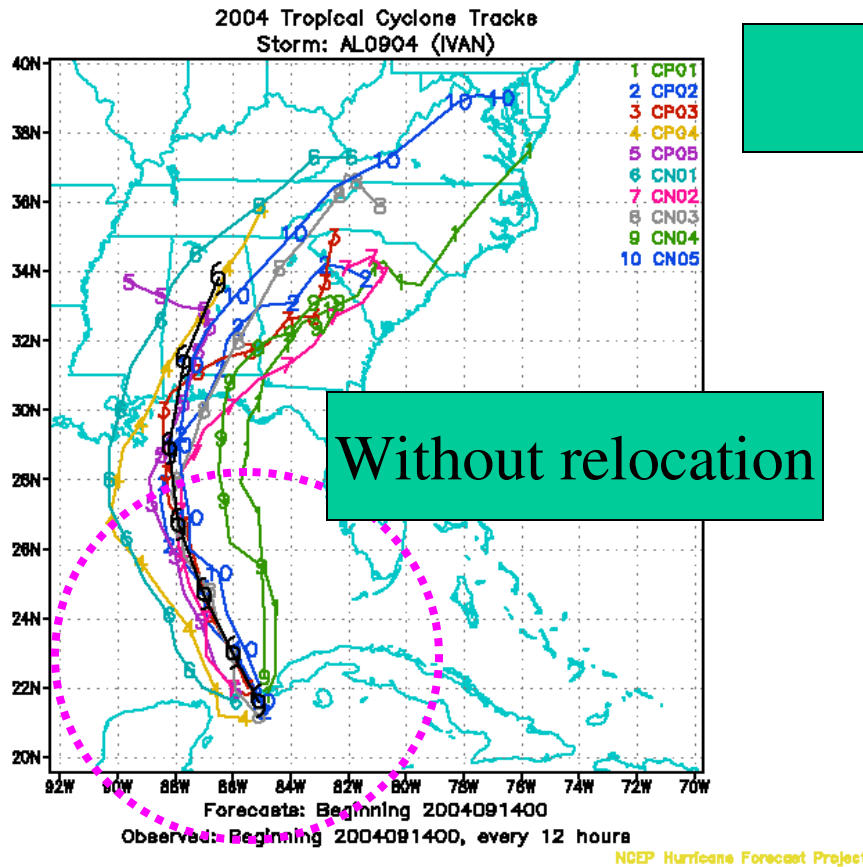


NCEP Hurricane Forecast Project

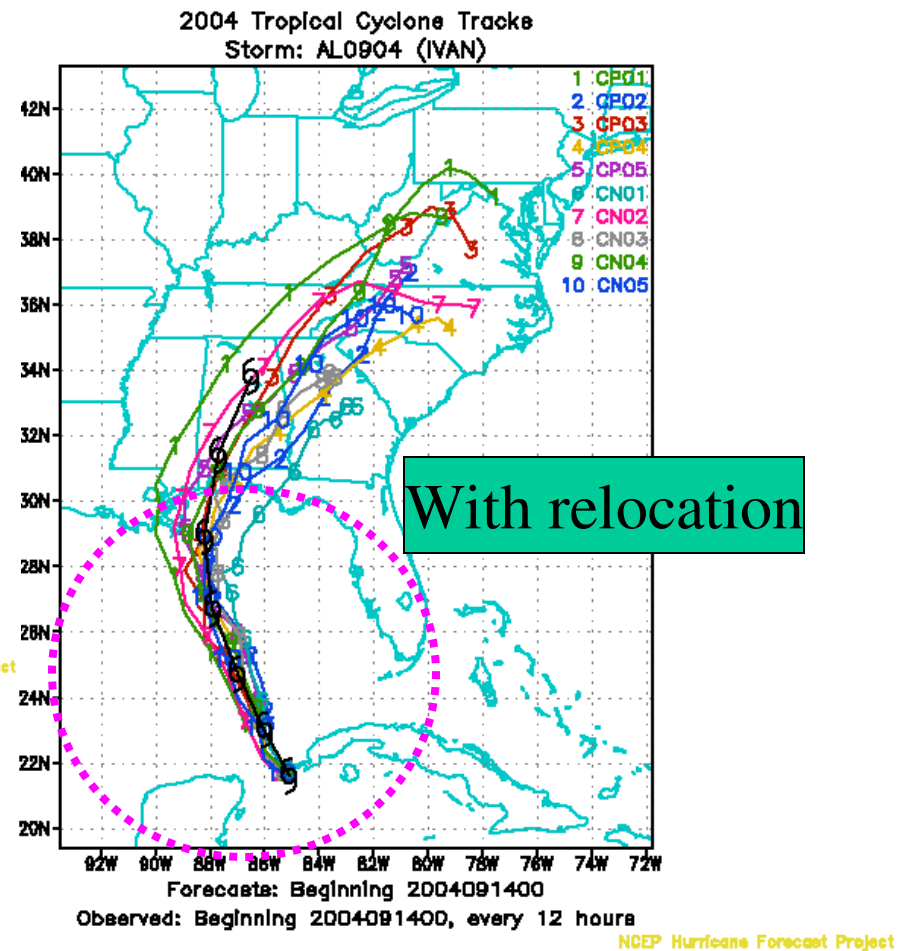
Reduced initial spread

Large initial spread

Hurricane Tracks Plots (case 2)

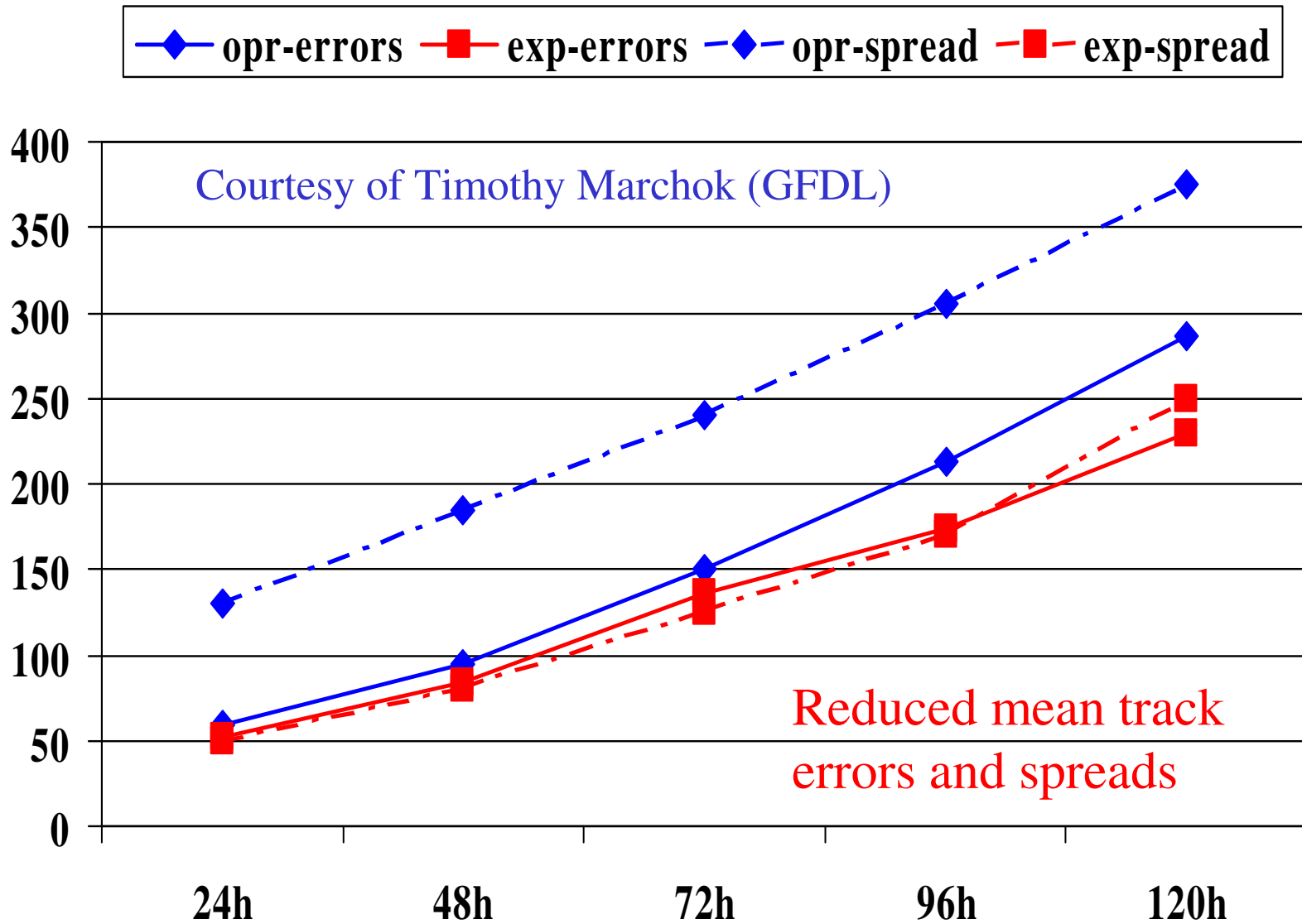


Ivan (09/14)



Track errors and spreads

2004 Atlantic Basin (8/23-10/1)



References

1. Zhu and Toth, 2008: "Ensemble based probabilistic verification" AMS, 2008
2. Zhu, 2007: "Objective evaluation of precipitation forecast" Special collection, Beijing, China
3. Zhu, 2005: "*Ensemble forecast: A new approach to uncertainty and predictability*" AAS
4. Wilks, 1995: "Statistical methods in atmospheric science" Academic Press
5. Toth, Talagrand, and Zhu, 2006: "*The attributes of forecast system*" book chapter. Cambridge University Press
6. Toth, Talagrand, Candille and Zhu, 2003: "*Probability and ensemble forecasts*" book chapter.
7. Zhu, 2004: "*Probabilistic forecasts and evaluations based on a global ensemble prediction system*" In book of Observation, theory and modeling of atmospheric variability,
8. Zhu, Iyenger, Toth, Tracton and Marchok, 1996: "*Objective evaluation of the NCEP global ensemble forecasting system*" AMS conference proceeding.
9. Hersbach, Hans, 2000: "*Decomposition of the Continuous Ranked Probability Score for ensemble prediction system*" Weather and Forecasting.
10. Toth, Zhu and Marchok, 2001: "*The use of ensembles to identify forecasts with small and large uncertainty*". Weather and Forecasting
11. Zhu, Toth, Wobus, Reardon and Mylne, 2002: "*The economic value of ensemble-based weather forecasts*" BAMS
12. Buizza, Houtekamer, Toth, Pellerin, Wei and Zhu, 2005: "*Assessment of the status of global ensemble prediction*" MWR
13. + more related articles

Strike probability!!!

Prob. Evaluation (useful tools)

2. Information content:

Use 10 climatologically equally likely bins to define events

Entropy = $P \log_2 P$:

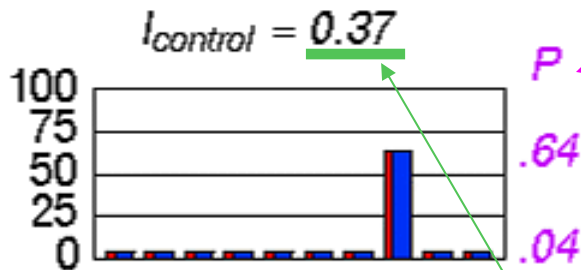
Example from large & small uncertainty

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

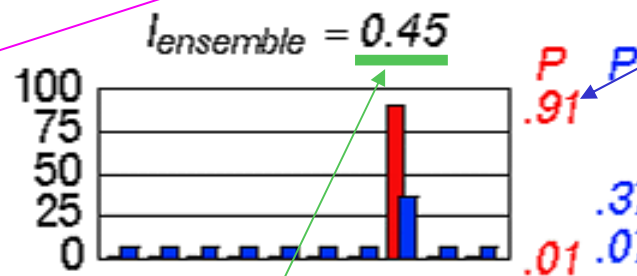
deterministic forecast

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

small uncertainty



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



Ensemble can differentiate between well and less predictable situations

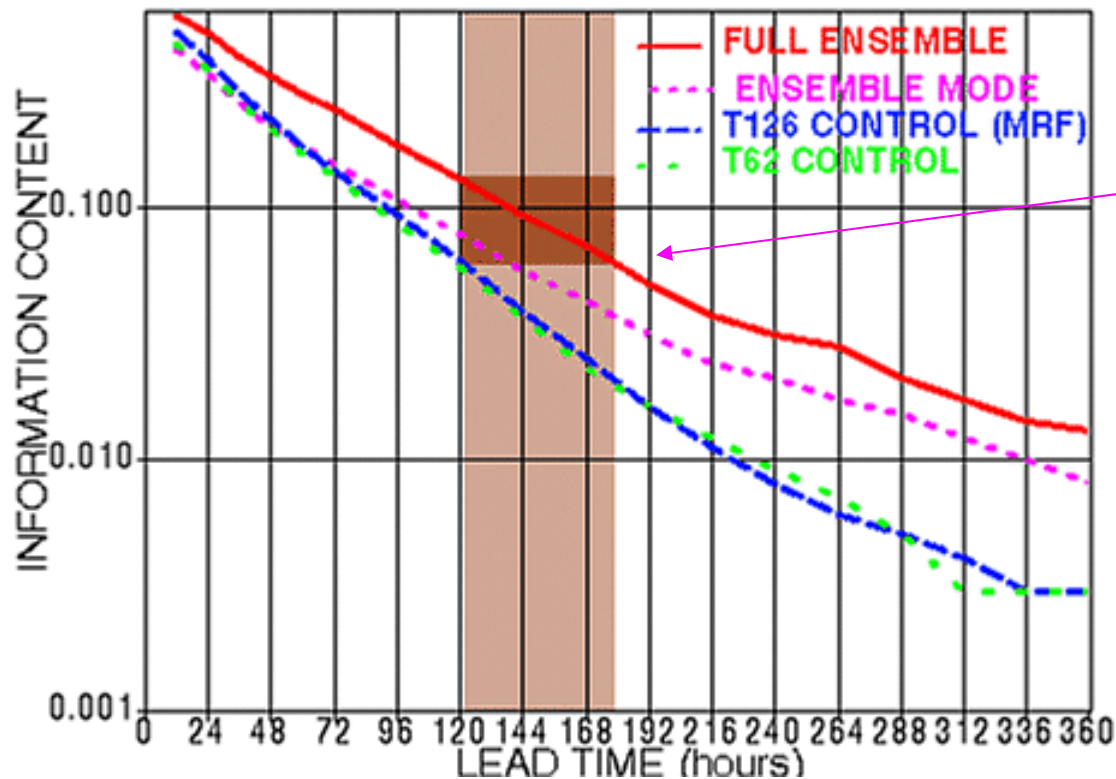
information content

large uncertainty

Prob. Evaluation (useful tools)

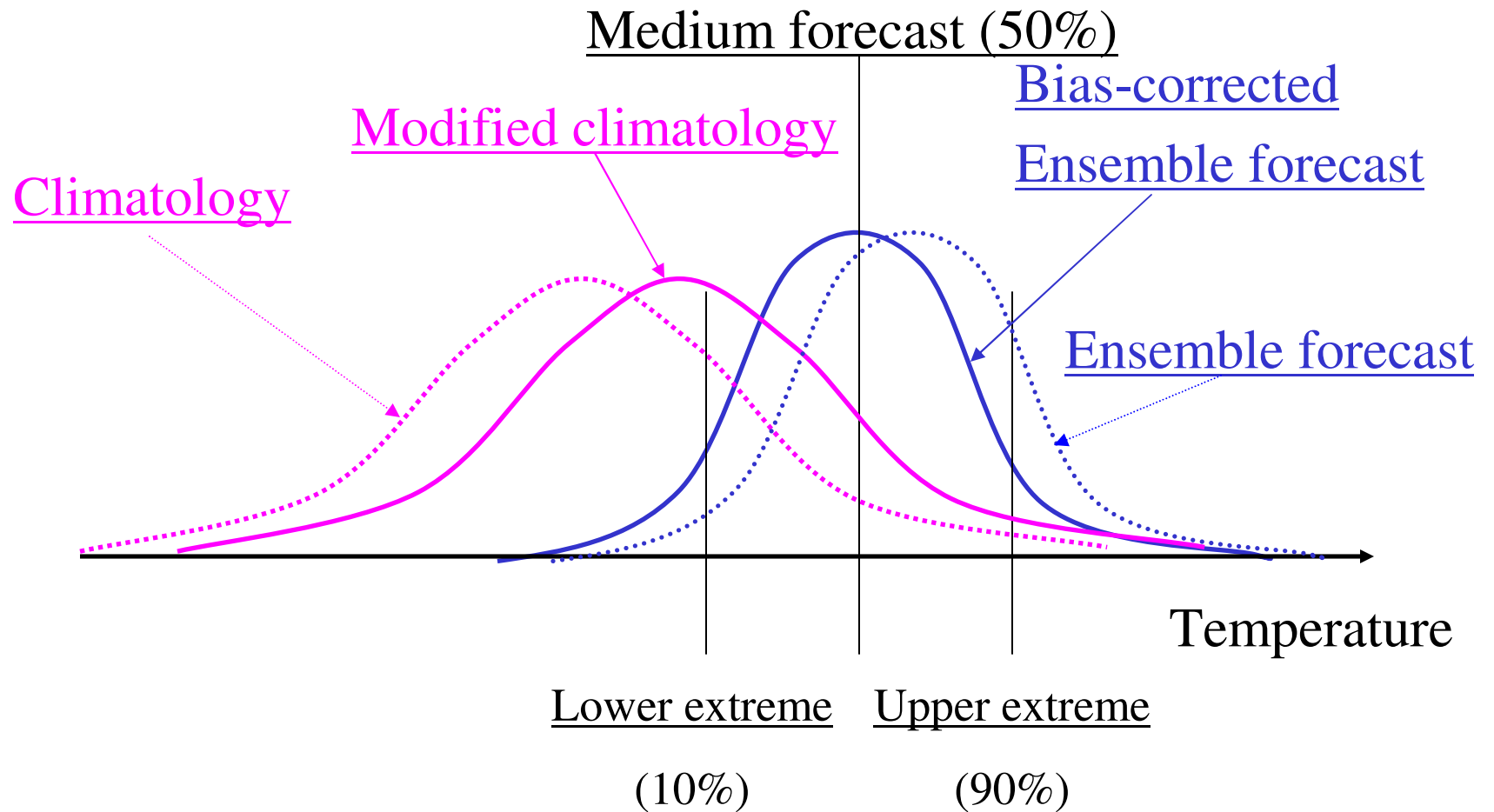
2. Information content:

Statistics show a **7.5-day** fully probabilistic forecast or **6-day** categorical forecast has as much information content as **5-day** control forecast. Or fully probabilistic forecast has more than **twice** as much information content at **day-5**.



ensemble mode
considers as
most frequent
forecasts

Schematic diagram for forecast anomalies



Climatology is generated from NCEP/NCAR reanalysis

(40 years from 1958 to 1997)