Ensemble Verification

Yuejian Zhu

Environmental Modeling Center
NOAA/NWS/NCEP

Acknowledgements:
Zoltan Toth GSD/ESRL

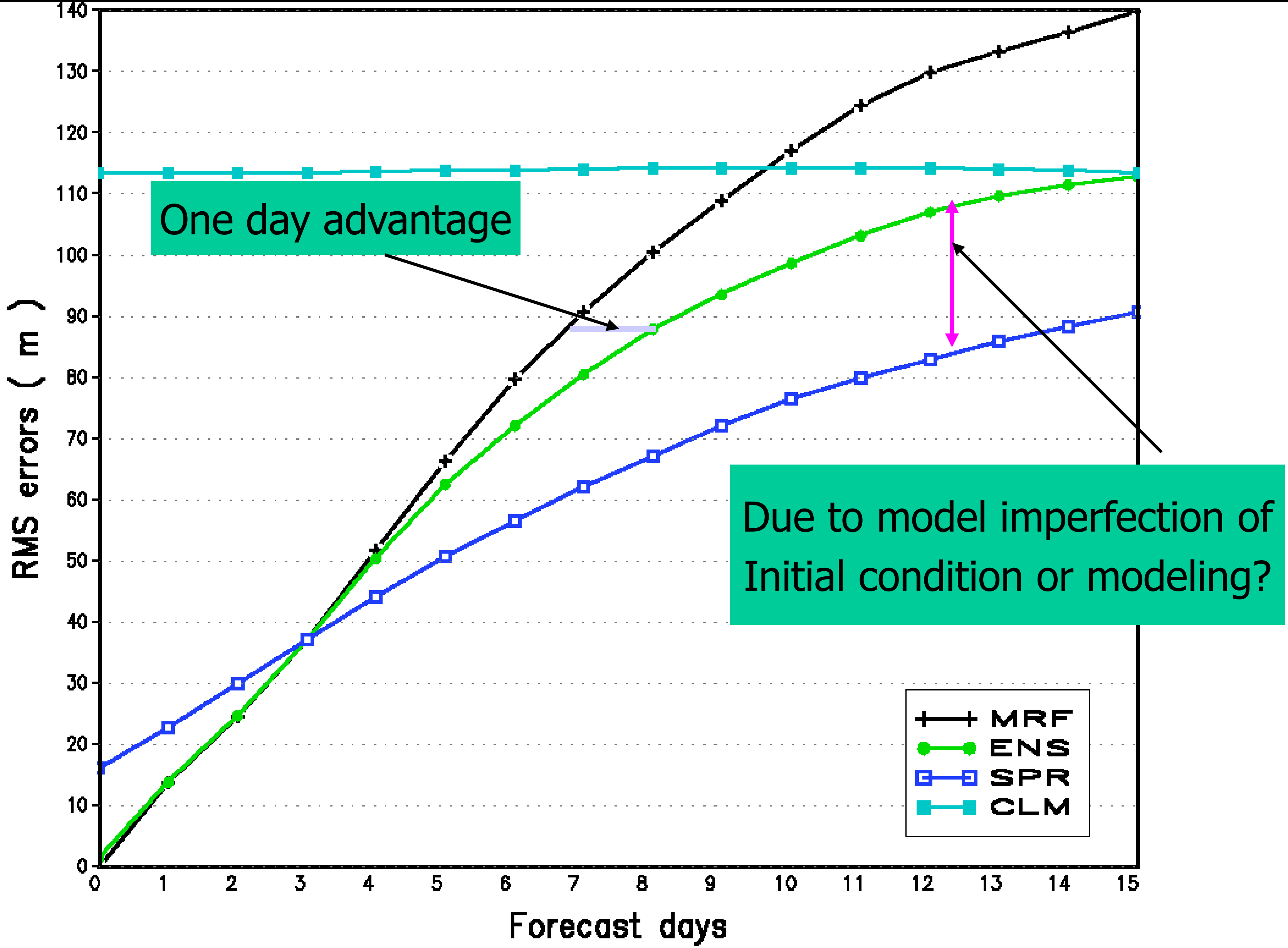
Outlines

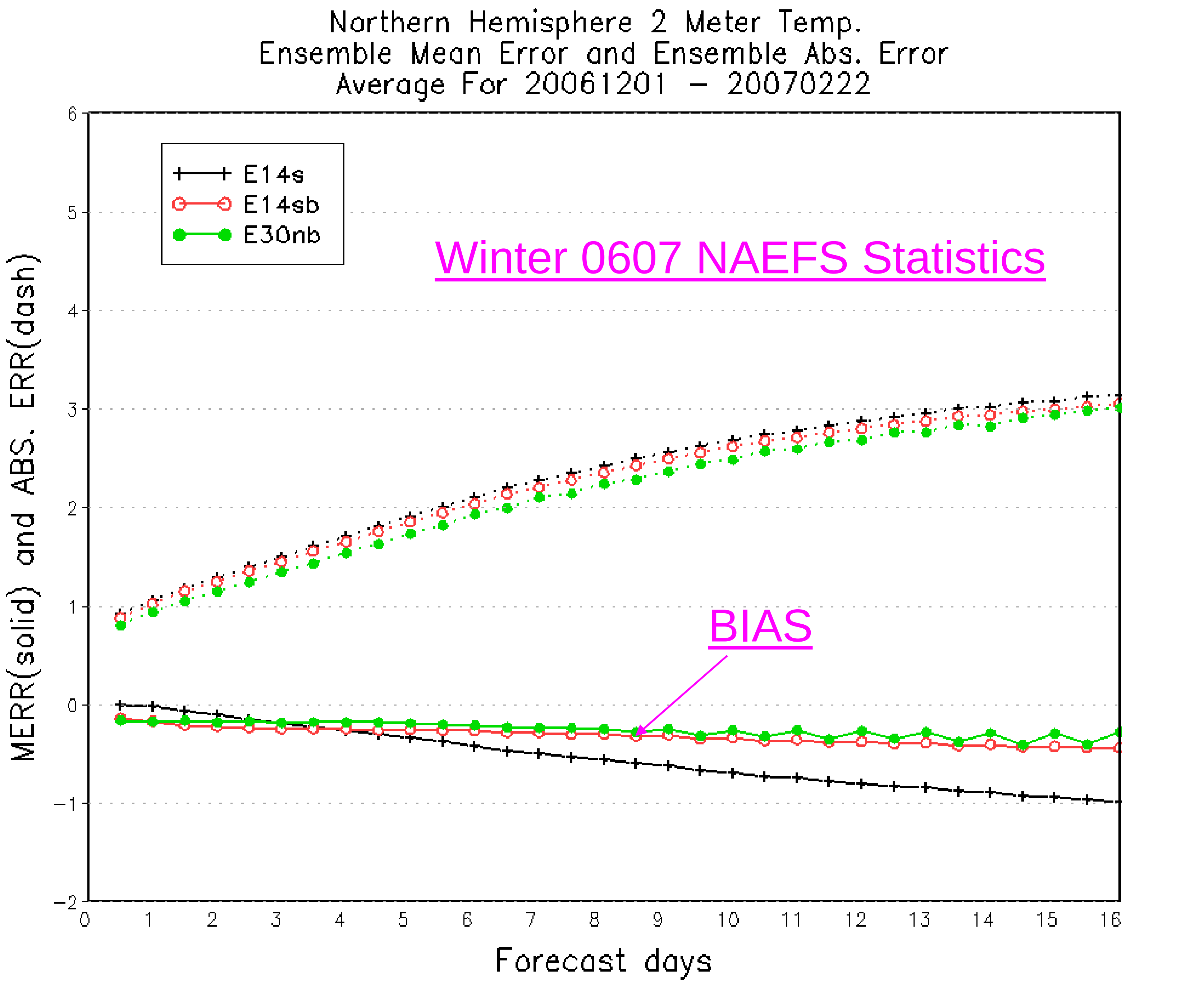
- Climatological Data
- RMS and Spread
  - Mean Error and Absolute Error
- Histogram and Outlier
- RPS and RPSS
- CRPS and CRPSS
- BSS (Resolution and Reliability)
- ROC (Hit Rate and False Alarm Rate)
- Economic Value (cost-loss analysis)

Climatological Data

- NCEP/NCAR 40 years (1958-1997) reanalysis
  - Will use CFSRR data (1979-2009) soon
- Monthly Sampling
  - For example: 40*30=1200
- 10 equally-a-likely, based on sampling
- Projected to verify date
- All forecast skills will base on 10 equally-a-likely climatological bins.
  - Except for defined threshold

1. RMS .vs Spread: To measure ensemble performance





Northern Hemisphere 2 Meter Temp.
Ensemble Mean Error and Ensemble Abs. Error
Average For 20061201 – 20070222
E14s
E14sb
E30nb
Winter 0607 NAEFS Statistics
BIAS
MERR(solid) and ABS. ERR(dash)
Forecast days

2. Talagrand Distribution - simple measurement

Talagrand Distribution (histogram distribution):

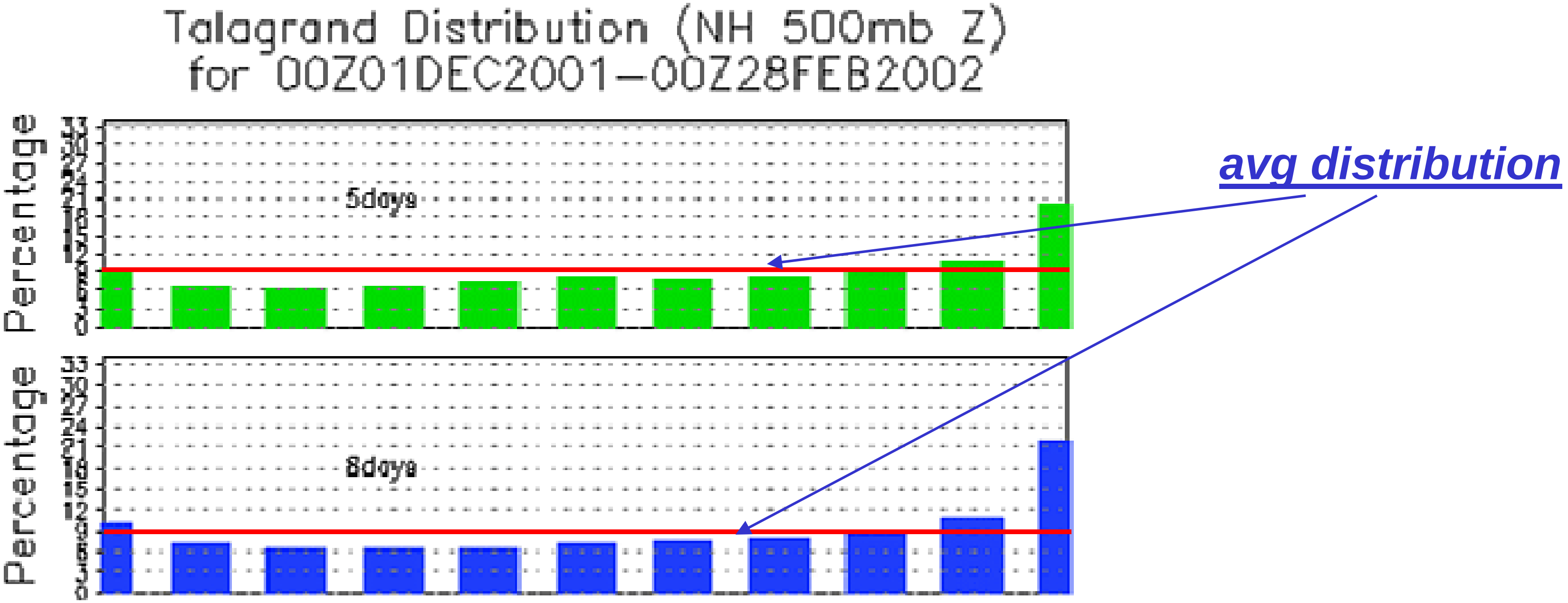
Sorting forecast in order, to check where the analysis is falling

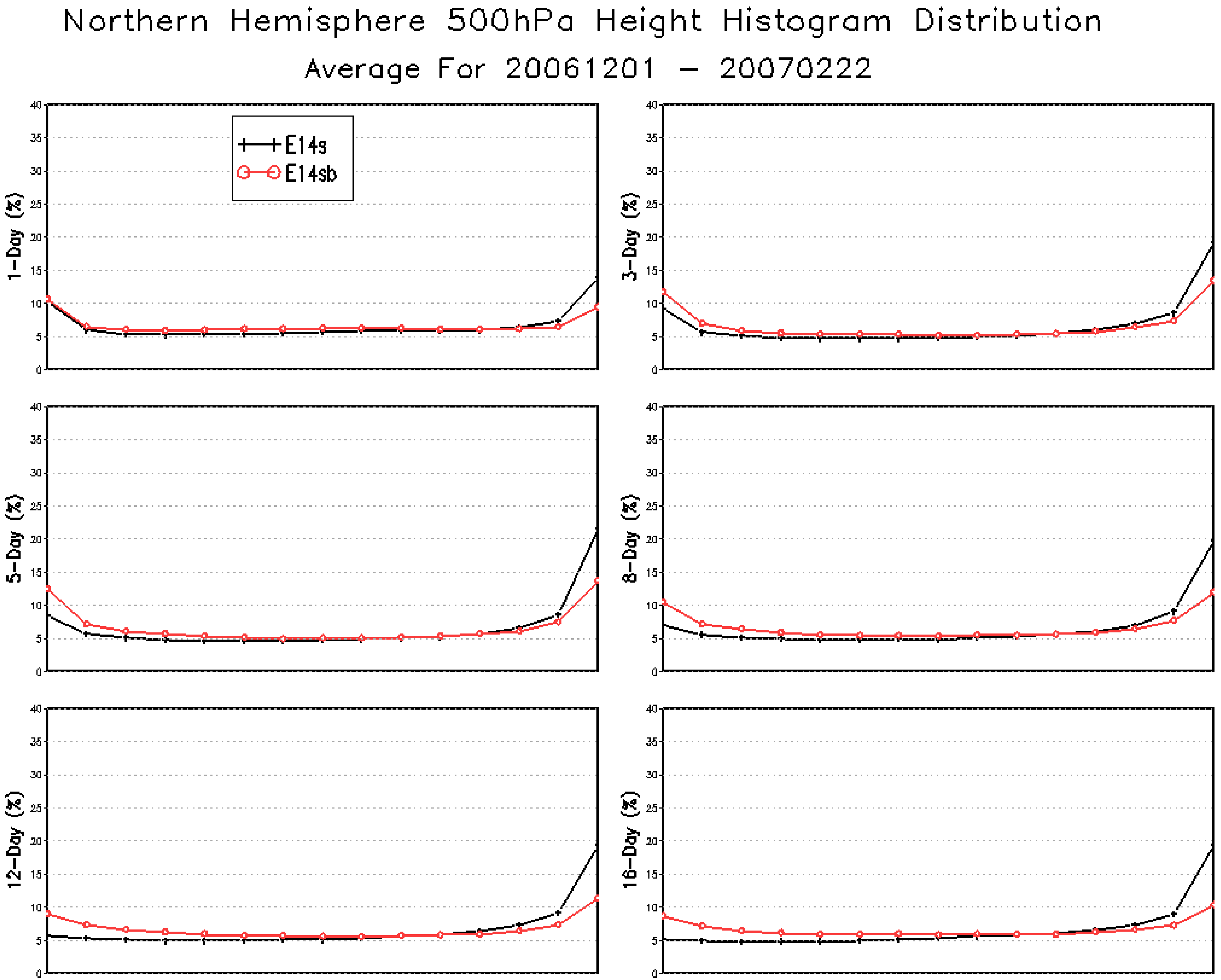
Reliability measurement, system bias detected.

positive/negative biased for forecasting model,

example of these forecasts --> cold bias,

assume analysis is bias-free (perfect). Common -"U" sharp




Northern Hemisphere 500hPa Height Histogram Distribution
Average For 20061201 – 20070222
E14s
E14sb
1-Day (%)
3-Day (%)
5-Day (%)
8-Day (%)
12-Day (%)
16-Day (%)
0
5
10
15
20
25
30
35
40

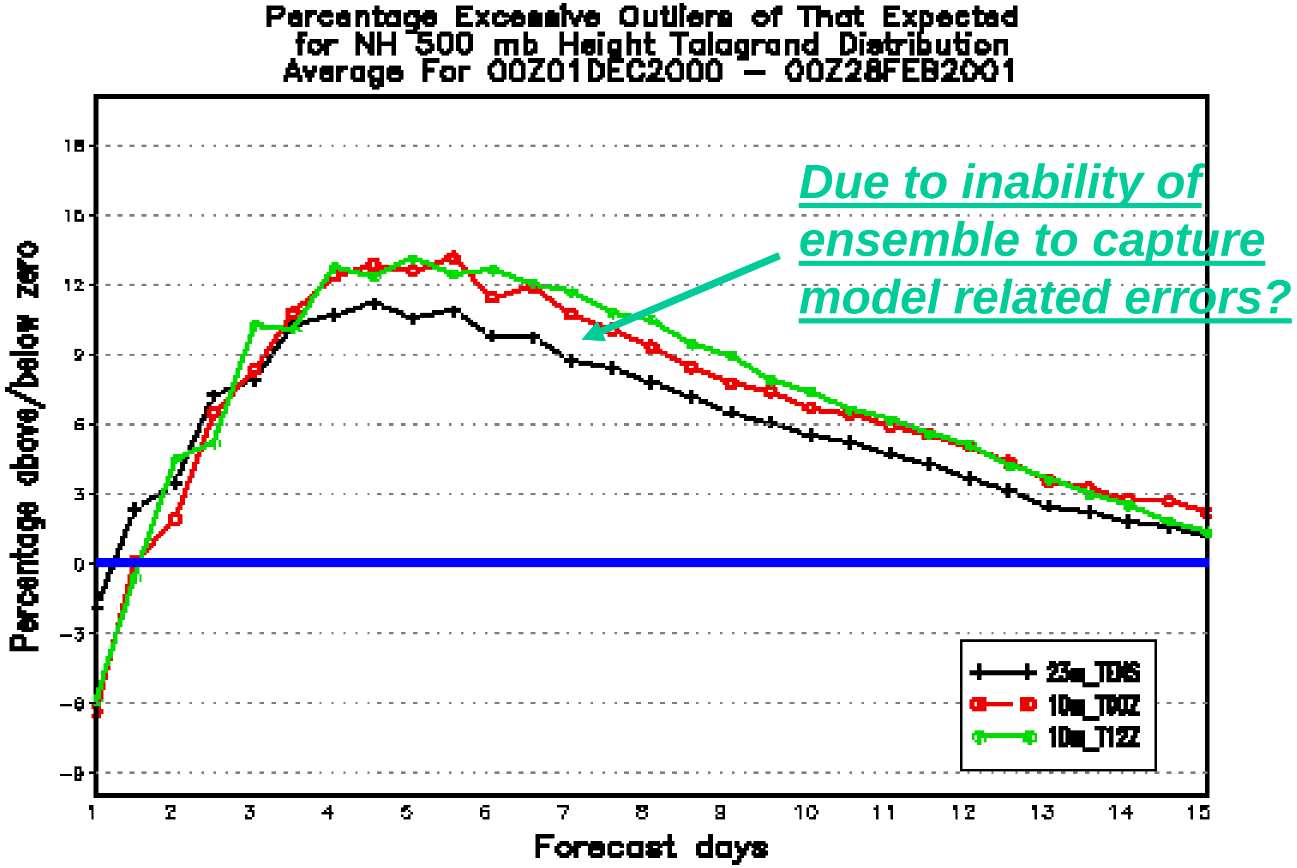
Talagrand Distribution - simple measurement

Talagrand distribution (continue).

. Outlier evolution by different leading time

.. Adding up two outliers subtract the average.

… Ideal forecasts will have zero outliers.



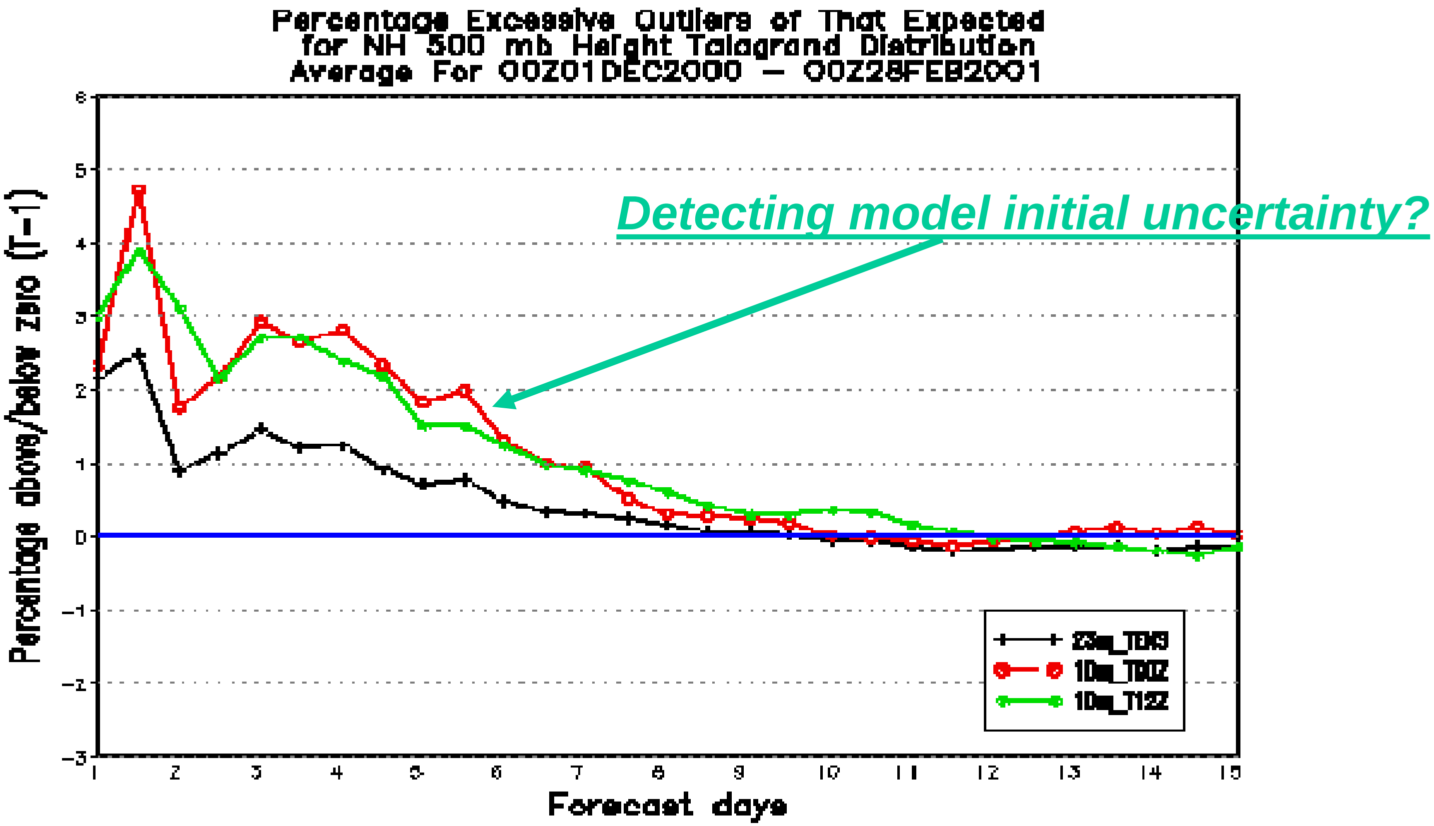
Talagrand Distribution - simple measurement

Outlier --> diagnostic

forecasts .vs. next forecasts (f+24hrs valid at same time)

assume forecasting model is perfect, f+24.

perfect forecast system will expect the outliers are zero.



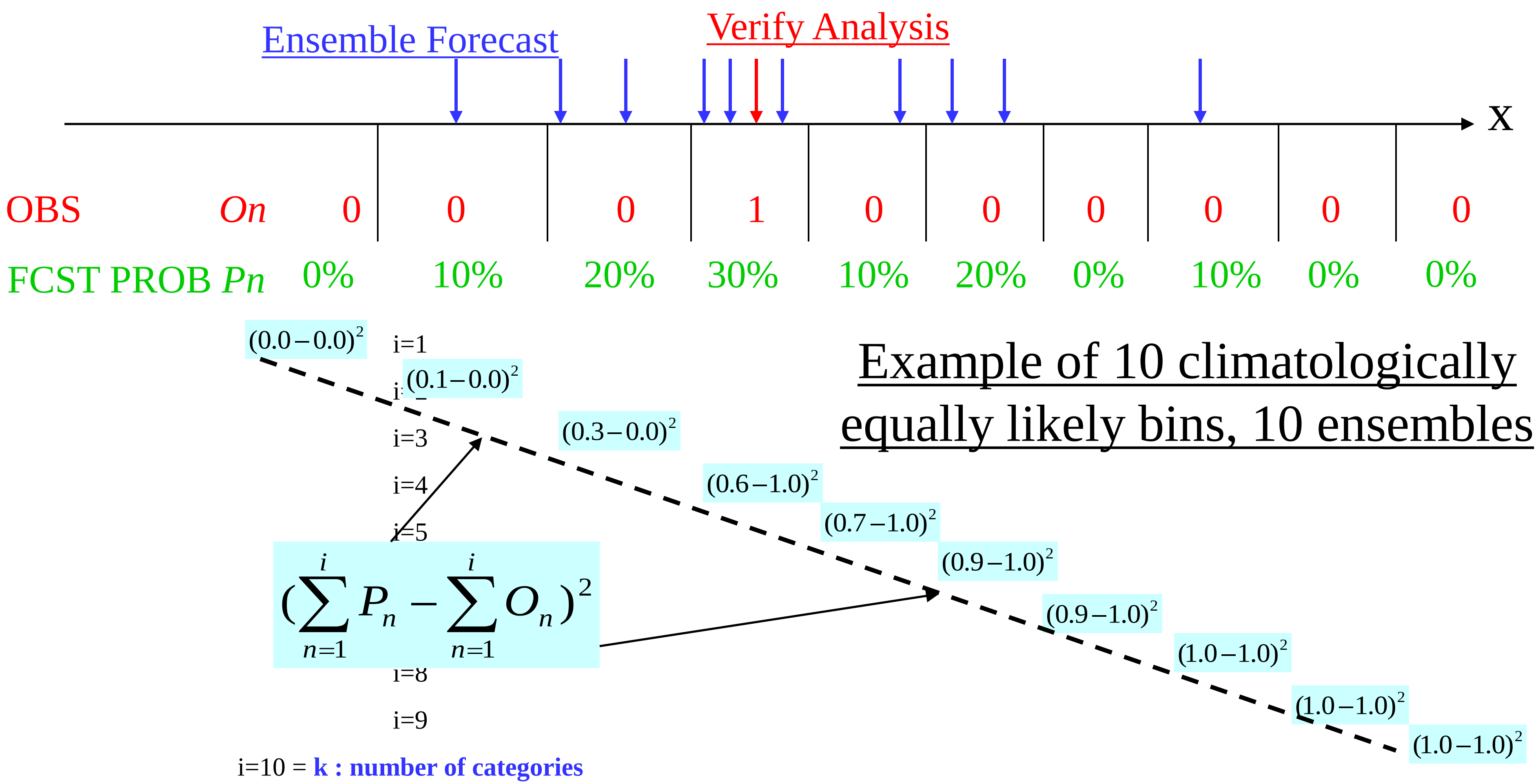
3. Ranked Probabilistic Score - multi-categories

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

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

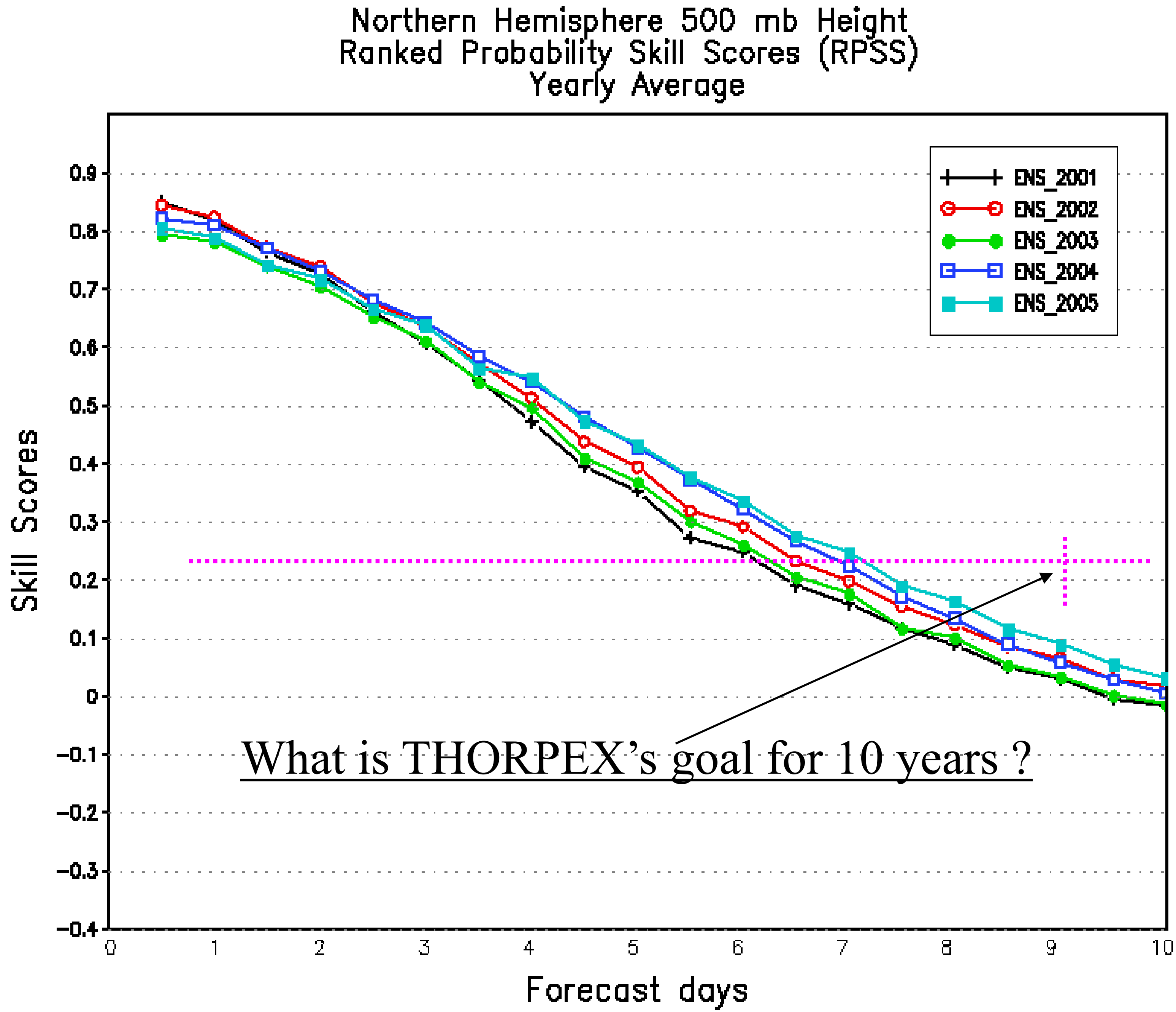
and

$$RPSS = \frac{RPS_f - RPS_c}{1 - RPS_c}$$



$$\left(\sum_{n=1}^{i} P_n - \sum_{n=1}^{i} O_n\right)^2$$

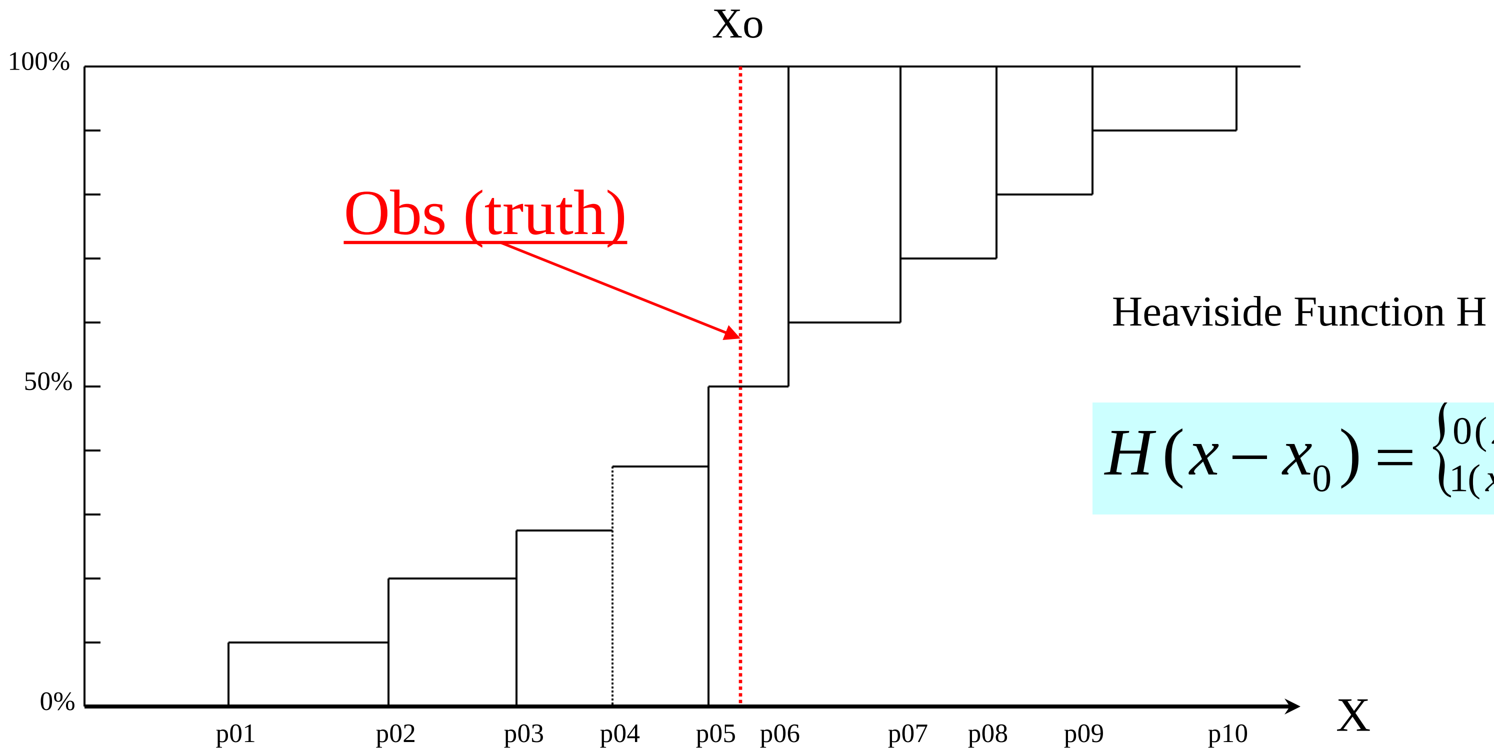
Example of 10 climatologically equally likely bins, 10 ensembles


Northern Hemisphere 500 mb Height
Ranked Probability Skill Scores (RPSS)
Yearly Average
ENS_2001
ENS_2002
ENS_2003
ENS_2004
ENS_2005
Skill Scores
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
0 1 2 3 4 5 6 7 8 9 10
Forecast days
What is THORPEX's goal for 10 years ?

4. Continuous Rank Probability Score – multi-categories

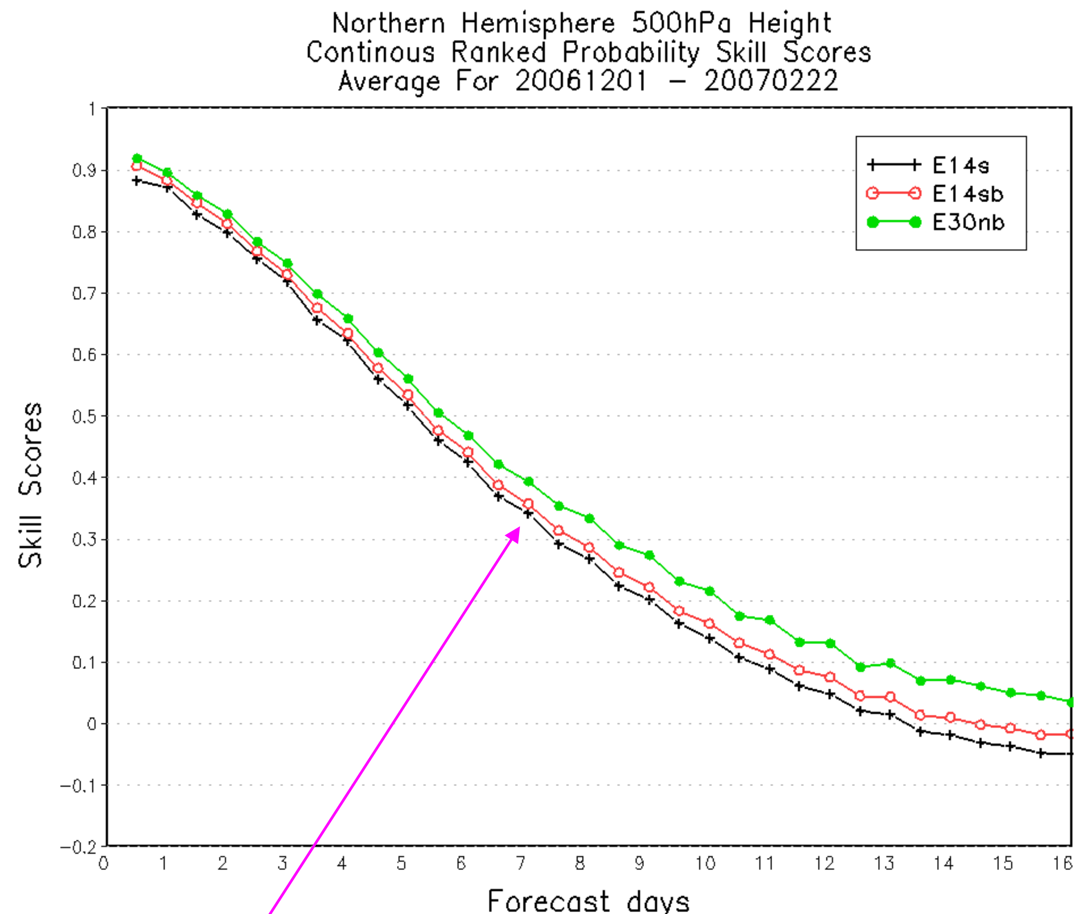
$$CRPS = \int_{-\infty}^{+\infty} [F(x) - H(x - x_0)]^2 \, dx$$

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

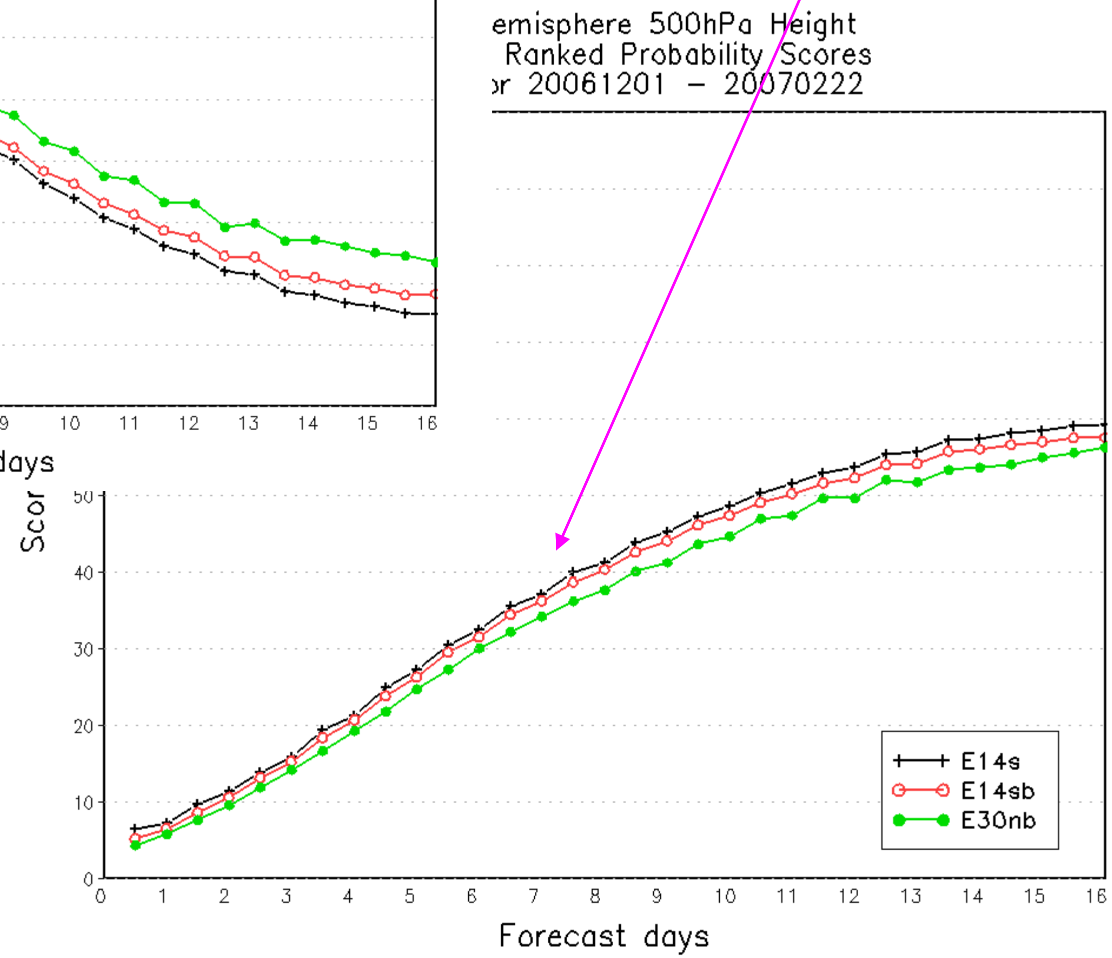


$$H(x - x_0) = \begin{cases} 0 \; (x \le x_o) \\ 1 \; (x > x_o) \end{cases}$$

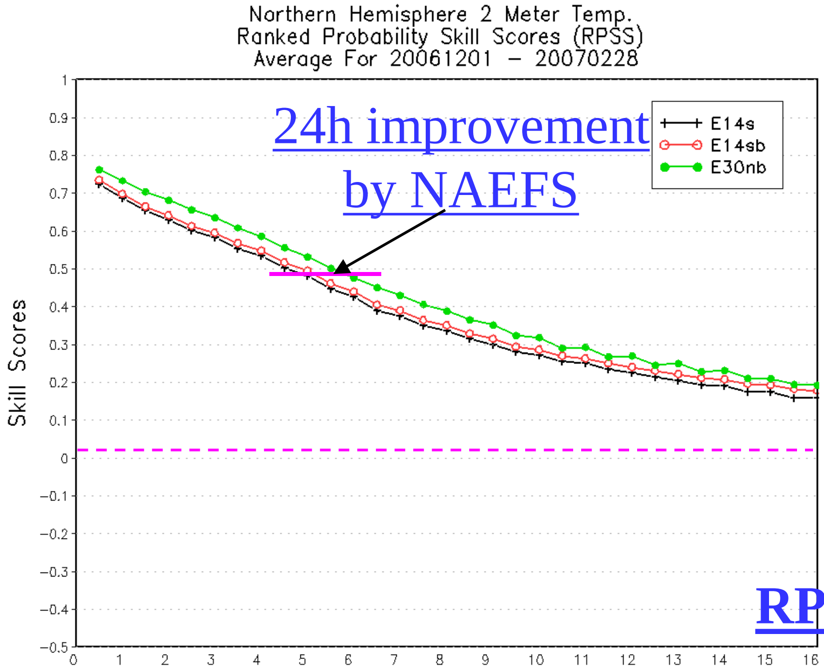
Order of 10 ensemble members (p01, p02,…,p10)

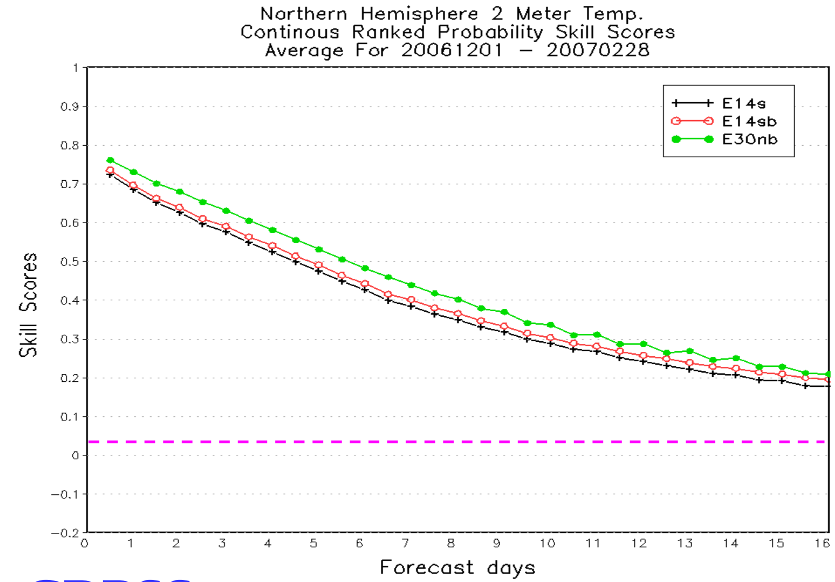


CRPS for winter 0607

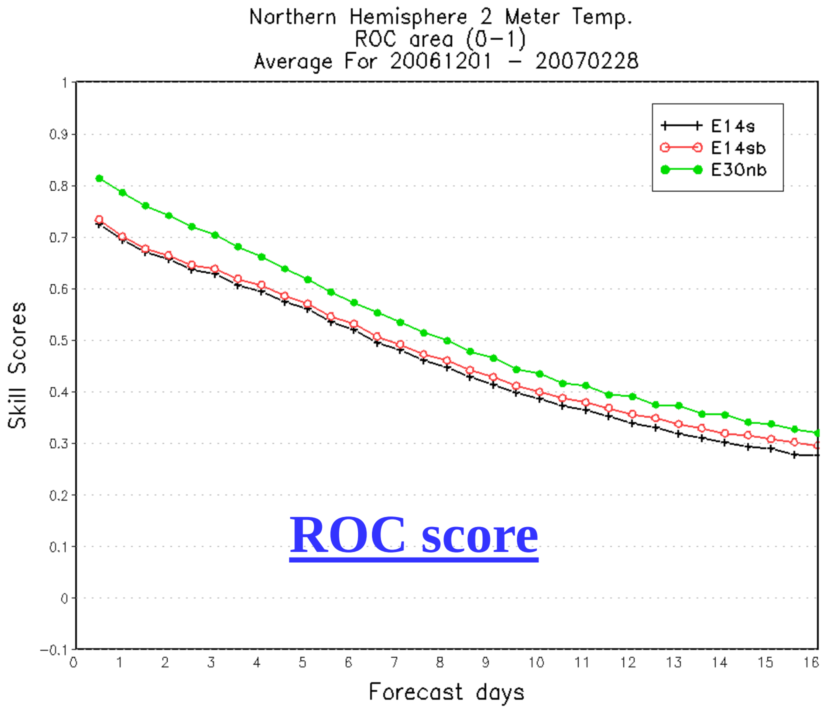


CRPSS for winter 0607





RPSS .vs CRPSS



Winter 2006-2007

NH 2m temperature

For

NCEP raw forecast (black)

NCEP bias corrected forecast (red)

NAEFS forecast (pink)

5. Brier Score: 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,*
***BSperf**=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

Reliability **Resolution** **Uncertainty**

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i (y_i - \overline{o_i})^2 - \frac{1}{n}\sum_{i=1}^{I} N_i (\overline{o_i} - \overline{o})^2 + \overline{o}(1-\overline{o})$$

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

Conditional probability of observed and sample climatology

$$\overline{o_i} = p(o_1 \mid y_i) = \frac{1}{N_i}\sum_{k \in N_i} o_k$$

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

Brier Score (and decomposition)

4. Example for BS calculation

	Ens(1)	Ens(2)	Ens(3)	Ens(4)	Ens(5)	anl
Point 1	25	23	20	24	28	23
Point 2	21	23	30	25	20	28
Point 3	27	20	28	19	19	27
Point 4	29	27	31	29	27	28
Point 5	20	26	18	20	21	19

By considering three equally likely bins: Cb<22, 22<=Cn<26 and Ca>26

Bins	C b< 22		22<= Cn <26		Ca>=26	
	F	A	F	A	F	A
Point 1	0.2	0.0	0.6	1.0	0.2	0.0
Point 2	0.4	0.0	0.4	0.0	0.2	1.0
Point 3	0.6	0.0	0.0	0.0	0.4	1.0
Point 4	0.0	0.0	0.0	0.0	1.0	1.0
Point 5	0.8	1.0	0.0	0.0	0.2	0.0
Summary	0.120		0.064		0.216	

The average Brier Score is 0.133 for this case, BS=0.133 (range from 0 to 1)

Brier Score (and decomposition)

5. Example for BS decomposition calculation

Probability	0/5	1/5	2/5	3/5	4/5	5/5
Point 1		Cb,Ca		Cn,A		
Point 2		Ca,A	Cb,Cn			
Point 3	Cn		Ca,A	Cb		
Point 4	Cb,Cn					Ca,A
Point 5	Cn	Ca			Cb,A	
Counts (A)	0	1	1	1	1	1
Counts (F)	4	4	3	2	1	1

$$BS = \frac{1}{n}\sum_{i=1}^{I} N_i (y_i - \overline{o_i})^2 - \frac{1}{n}\sum_{i=1}^{I} N_i (\overline{o_i} - \overline{o})^2 + \overline{o}(1-\overline{o})$$

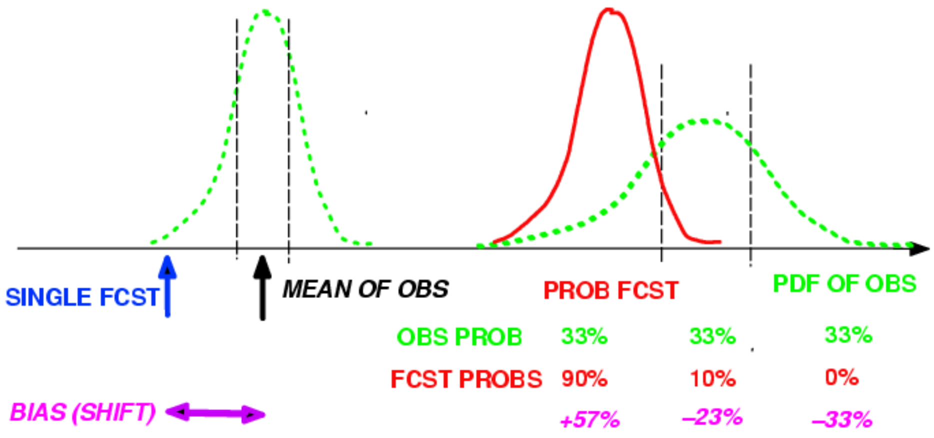
OBS PROB	0.0000	0.2500	0.3333	0.5000	1.0000	1.0000
FCST PROB	0.0000	0.2000	0.4000	0.6000	0.8000	1.0000
Climatology	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
Weights	4/15	4/15	3/15	2/15	1/15	1/15
RELI distance	0.0000	0.0500	-0.0667	-0.1000	0.2000	0.0000
RESO distance	-0.3333	-0.0833	0.0000	0.1667	0.6667	0.6667
Reliability	0.0000	0.0007	0.0009	0.0013	0.0027	0.0000
Resolution	0.0296	0.0019	0.0000	0.0037	0.0296	0.0297

Rel=0.0056, Res=0.0889, Unc=0.2222, BS=0.1333 (same as previous one)

TWO MAIN ATTRIBUTES OF FORECASTS

RELIABILITY – Lack of systematic error

(No conditional bias)



Consider cases with same forecast

Construct pdf of corresponding observtns

If fcst identical to pdf of observations =>

PERFECT RELIABILITY

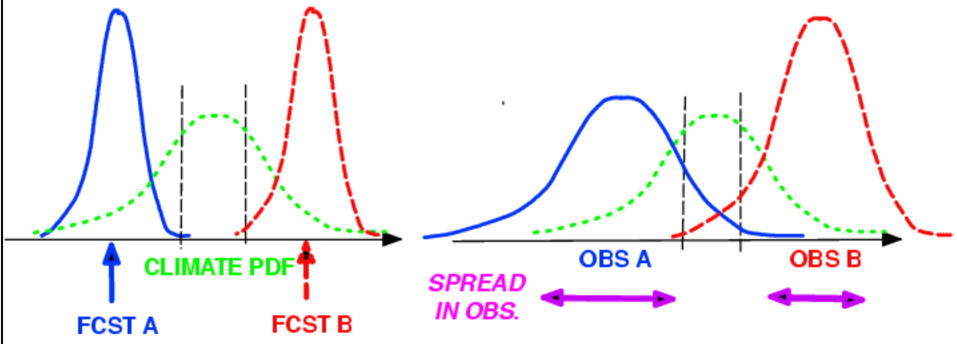
Reliability ***CAN BE*** *statistically corrected*

(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

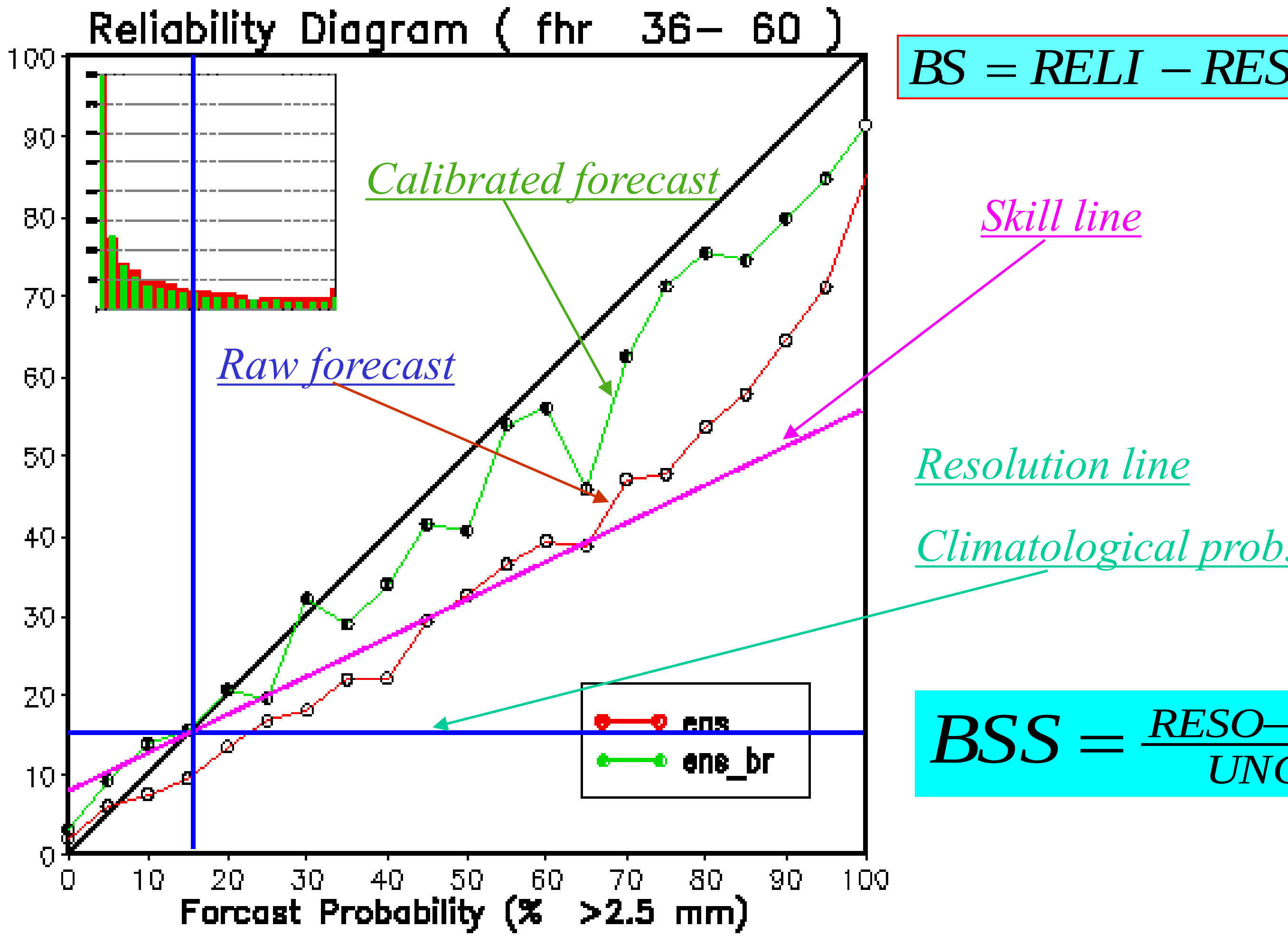
Resolution ***CANNOT BE*** *statistically corrected*

INTRINSIC VALUE OF FCST SYSTEM

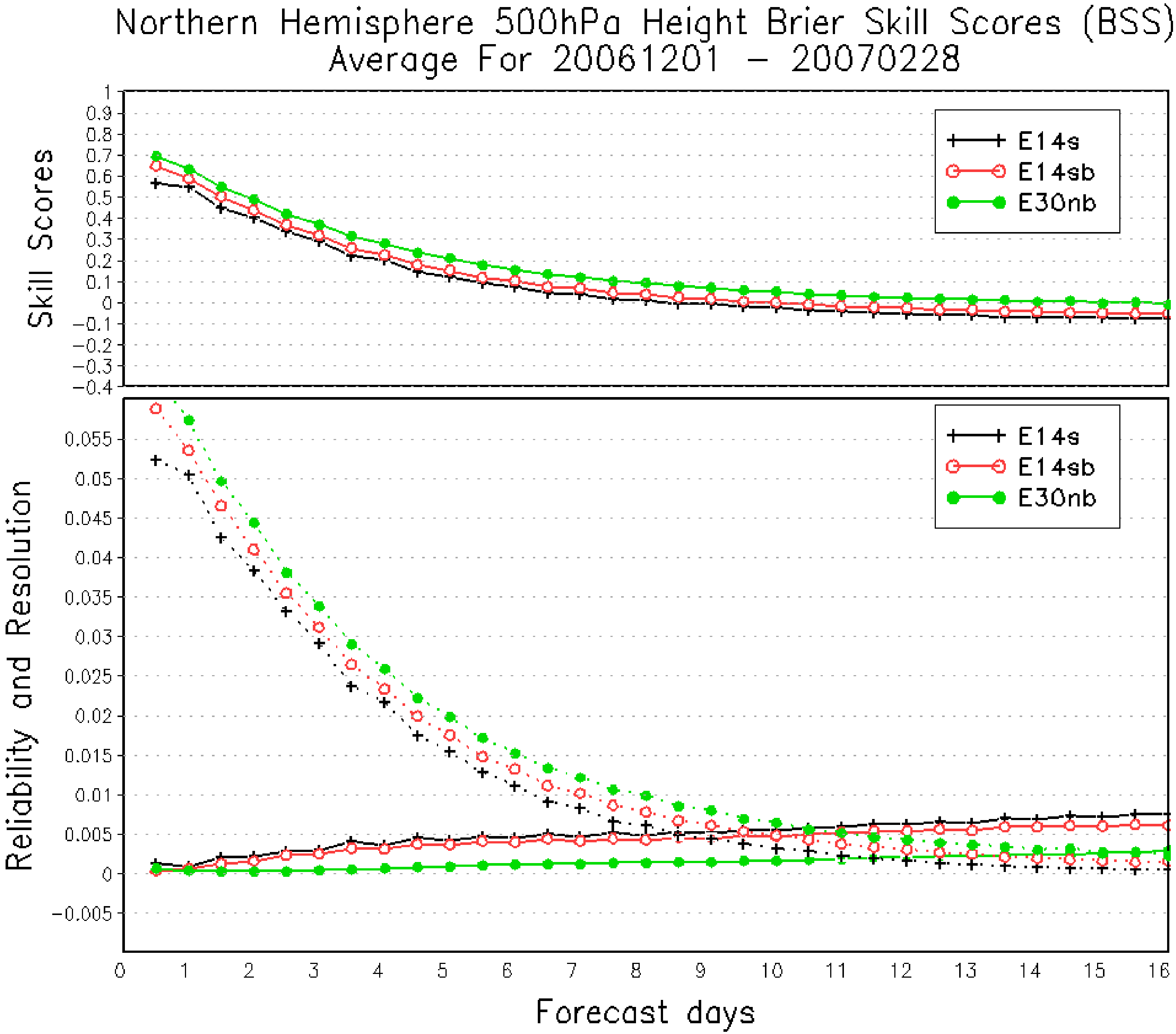
Brier Score (and decomposition)

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

$$BS = RELI - RESO + UNCE$$



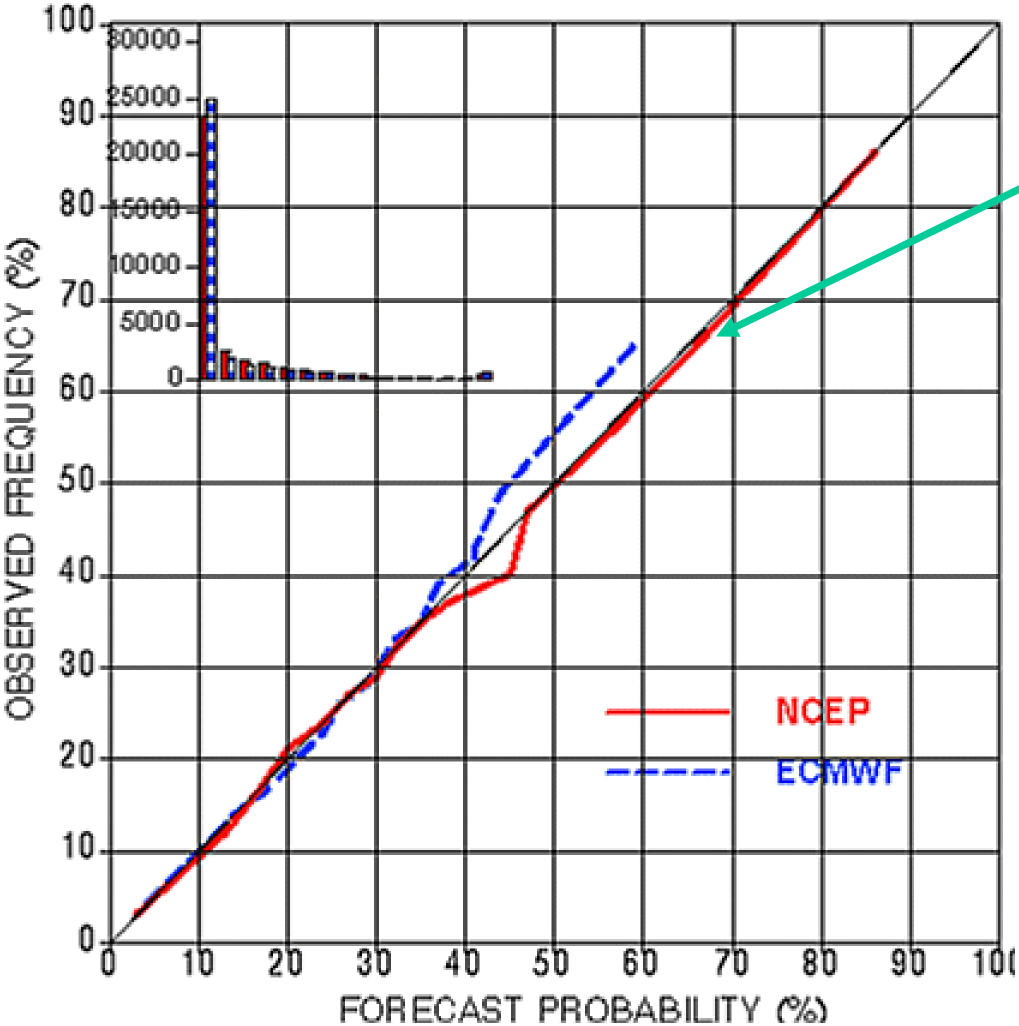
$$BSS = \frac{RESO - RELI}{UNCE}$$



Northern Hemisphere 500hPa Height Brier Skill Scores (BSS)
Average For 20061201 – 20070228
Skill Scores
E14s
E14sb
E30nb
Reliability and Resolution
Forecast days

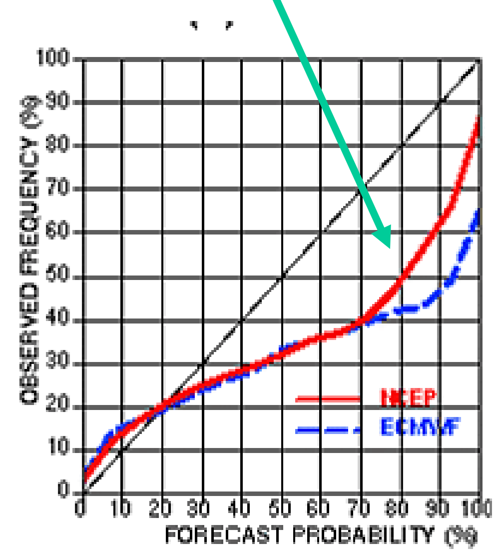
Brier Score (and decomposition)

4. Reliability and possible probabilistic calibration:
re-label fcst prob by obs frequency associated with fcst

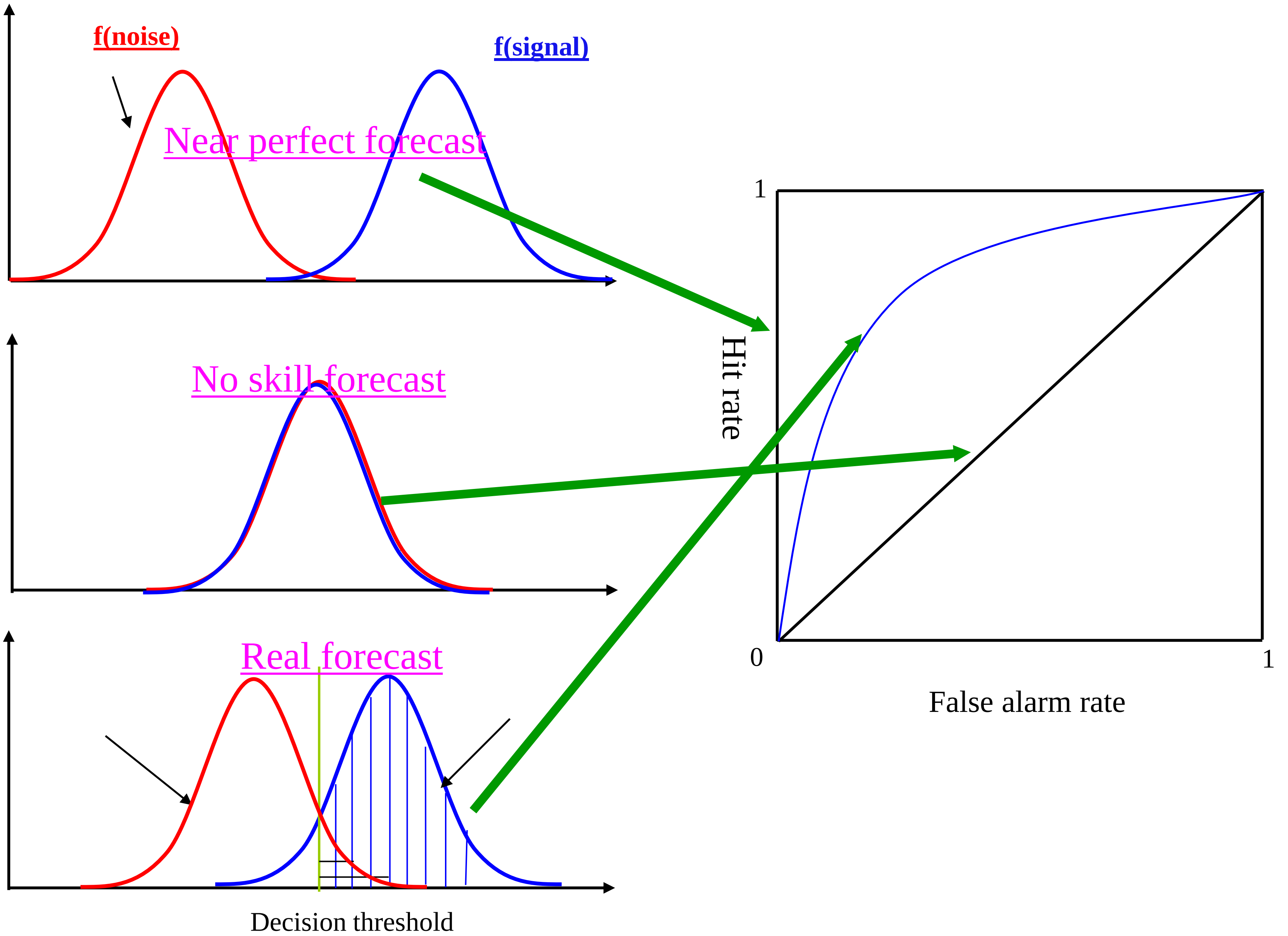


calibrated

Un-calibrated



6. Relative Operating Characteristics area (ROC area)



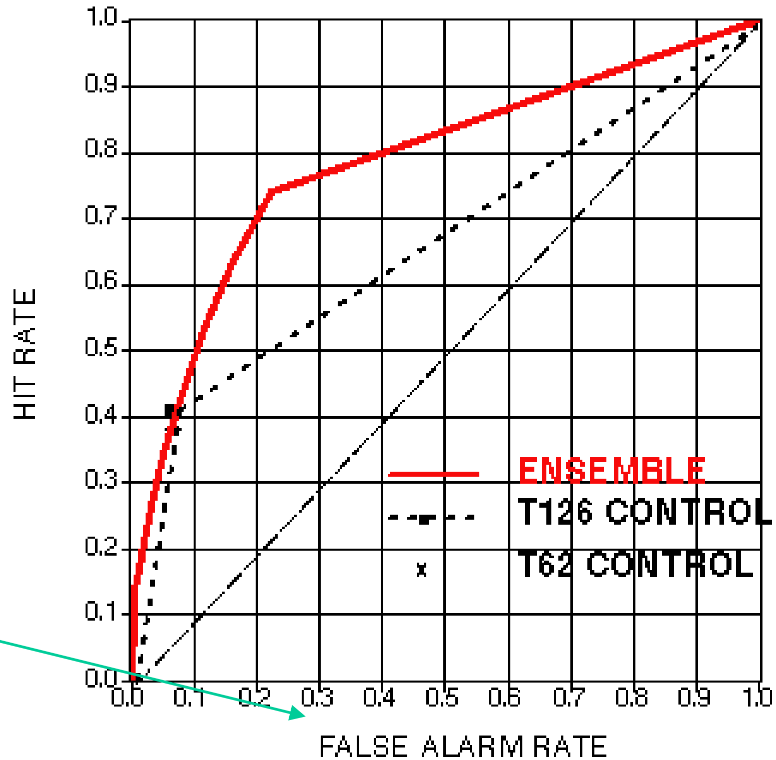
ROC area (cost-loss analysis)

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.***

ROCarea = Intergrated area * 2 (0-1 normality)

h/(h+m)



f/(h+f)

Relative Operating Characteristics

o\f	y(f)	n(f)
y(o)	h	m
n(o)	f	c

*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.*

7. 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.

Zhu and etc.. 2002: BAMS

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

1. Expected Expense:

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

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

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

2. Economic Value:

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

or

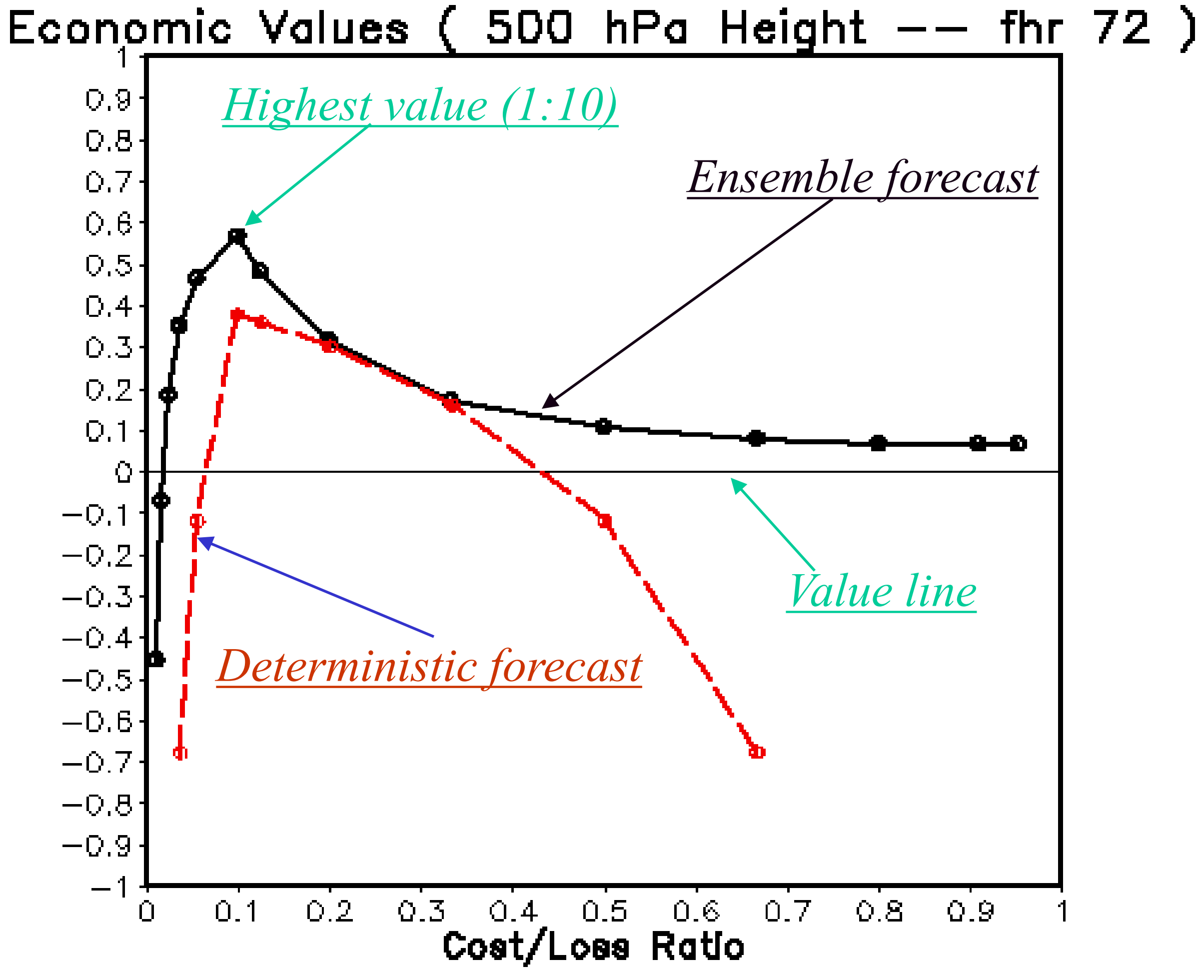
$$V = \frac{Min[o, r] - (h + f)r - m}{Min[o, r] - or}$$

*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*

Economic Value (cost-loss analysis)

Economic Value (EV) of forecasts.

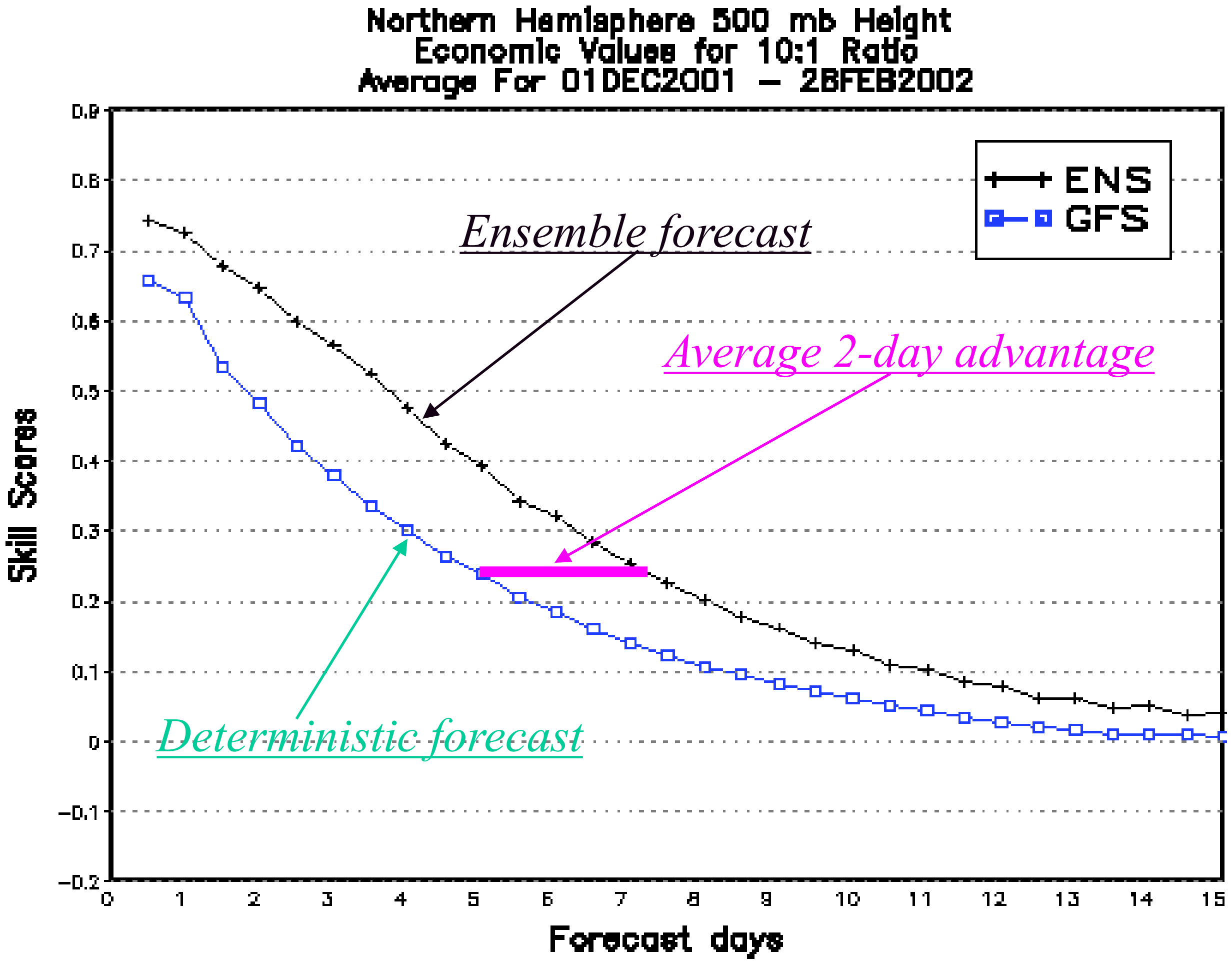
Given a particular forecast, a user either does or does not take action



Economic Value (cost-loss analysis)

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

.. Economic Value (EV) of forecasts



Decision Theory Example

Critical Event: sfc winds > 50kt

Cost (of protecting): $150K

Loss (if damage): $1M

		Forecast? YES	Forecast? NO
Observed?	YES	*Hit* $150K	***Miss*** **$1000K**
	NO	*False Alarm* $150K	*Correct Rejection* $0K

Case	Deterministic Forecast (kt)	Observation (kt)	Cost ($K)	Probabilistic Forecast	Cost ($K) by Threshold for Protective Action 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%

http://www.emc.ncep.noaa.gov/gmb/yzhu/html/opr/naefs.html

NOAA/NWS/EMC/GMB Ensemble Evaluation Home Page - Mozilla Firefox

http://www.emc.ncep.noaa.gov/gmb/yzhu/html/opr/naefs.html

NAEFS Verification Home Page of

Developed by Yuejian Zhu

Please view this disclaimer

Ensemble Verification (short presentation)

Predictability Metting Presentation

Period	NCEP .vs NCEPb	CMC .vs CMCb	NCEP .vs CMC	NCEPb .vs CMCb	NAEFS
spr2011	Yes	Yes	Yes	Yes	Yes
win1011	Yes	Yes	Yes	Yes	Yes
fal2010	Yes	Yes	Yes	Yes	Yes
sum2010	Yes	Yes	Yes	Yes	Yes
spr2010	Yes	Yes	Yes	Yes	Yes
win0910	Yes	Yes	Yes	Yes	Yes
fal2009	Yes	Yes	Yes	Yes	Yes
sum2009	Yes	Yes	Yes	Yes	Yes
spr2009	Yes	Yes	Yes	Yes	Yes
win0809	Yes	Yes	Yes	Yes	Yes
fal2008	Yes	Yes	Yes	Yes	Yes
sum2008	Yes	Yes	Yes	Yes	Yes
spr2008	Yes	Yes	Yes	Yes	Yes
win0708	Yes	Yes	Yes	Yes	Yes
fal2007	Yes	Yes	Yes	Yes	Yes
sum2007	Yes	Yes	Yes	Yes	Yes
spr2007	Yes	Yes	Yes	Yes	Yes
win0607	Yes	Yes	Yes	Yes	Yes
fal2006	Yes	Yes	Yes	Yes	Yes
sum2006	Yes	Yes	Yes	Yes	Yes

NCEP .vs NCEPb - Mozilla Firefox

http://www.emc.ncep.noaa.gov/gmb/yzhu/html/opr/naefs/NCEP_NCEPb_spr2011.html

500 hPa Height Scores NCEP .vs NCEPb

NH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
SH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
TROP	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM

1000 hPa Height Scores (NCEP .vs NCEPb)

NH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
SH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
TROP	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM

850 hPa Temperature Scores(NCEP .vs NCEPb)

NH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
SH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
TROP	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM

2 Meters Temperature Scores(NCEP .vs NCEPb)

NH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
SH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
TROP	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM

10 Meters Wind (U) Scores(NCEP .vs NCEPb)

NH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
SH	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM
TROP	ROC	EV	RPSS	BSS	CRP	CRPS	RMS/SPRD	ERR/ABSE	HISTOGRAM

Back ground !!!