Do Key Performance Targets Work ?
(Or How valid is administrative verification ?)

Clive Wilson

4th International Verification Methods Workshop, Helsinki 8-10 June 2009

Contents

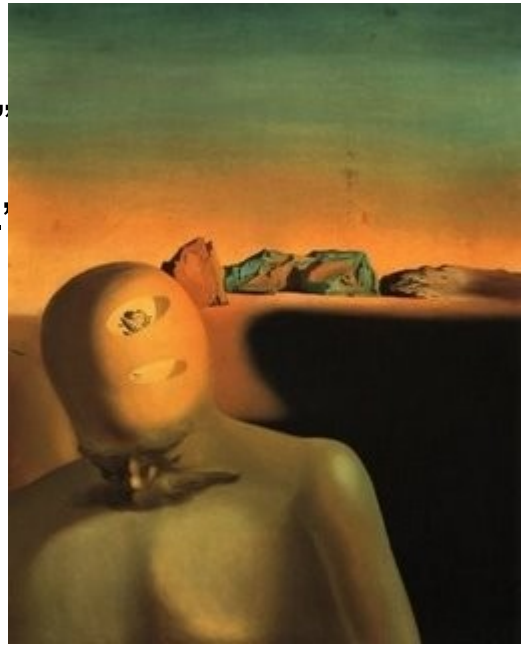
- Administrative & General public verification
- Business speak – Key Performance Indicators and targets (KPI, KPT)
- Composite scores – NWP Index
- Setting targets
- Experience at Met Office
- Conclusions & recommendations

Administrative & General public verification

How well do forecasting systems perform overall ?

Why do we need an overall measure ?

- To manage
- Demonstrate good “value for money”
- Judge if changes improve forecasts-”
- Evaluate investment/cost
- Justify funding



The Average Bureaucrat, Salvador Dali, 1930

Desirable:

- Stable objective process over years
- Small number of summarising scores
- Composite or representative

Cautions – Murphy(1991,1993), Stanski et al (1989)

- Forecast quality is multifaceted – ~~single number~~
- Pitfalls of extreme summarising
  - Single number – tremendous pressure on design of verification system:
    - Does chosen score reliably measure what is desired
    - Components treated fairly in compositing ?
    - How to fairly weight the components in the composite ?

Key Performance Indicators & Targets

KPIs=Metrics used to quantify objectives to reflect strategic performance of an organization

Targets

- Quantifiable
- benchmark
- Time frame
- Eg increase Profit/turnover (NWP index) from xxx at end of March to yyy by next March

Influence behaviour towards strategic aim

- Corporate bonus – encourage effort

Met Office NWP index

- Started in 1990s
- Originally only global forecasts & RMSE-based
  - Too sensitive to anomalous regimes
- Changed to MSE skill-score against persistence
- UK forecasts added
- Combined global and UK since 2001
- Originally annual means
  - Unmatched to development timescales
  - Large interannual variability for UK region
- Multi-annual means (3y)

Global Index

- Parameters-MSLP,H500,W250,W850
- Areas-NH,SH,Tropics(20°N to 20°S)
- MSE Skill v persisted analysis,

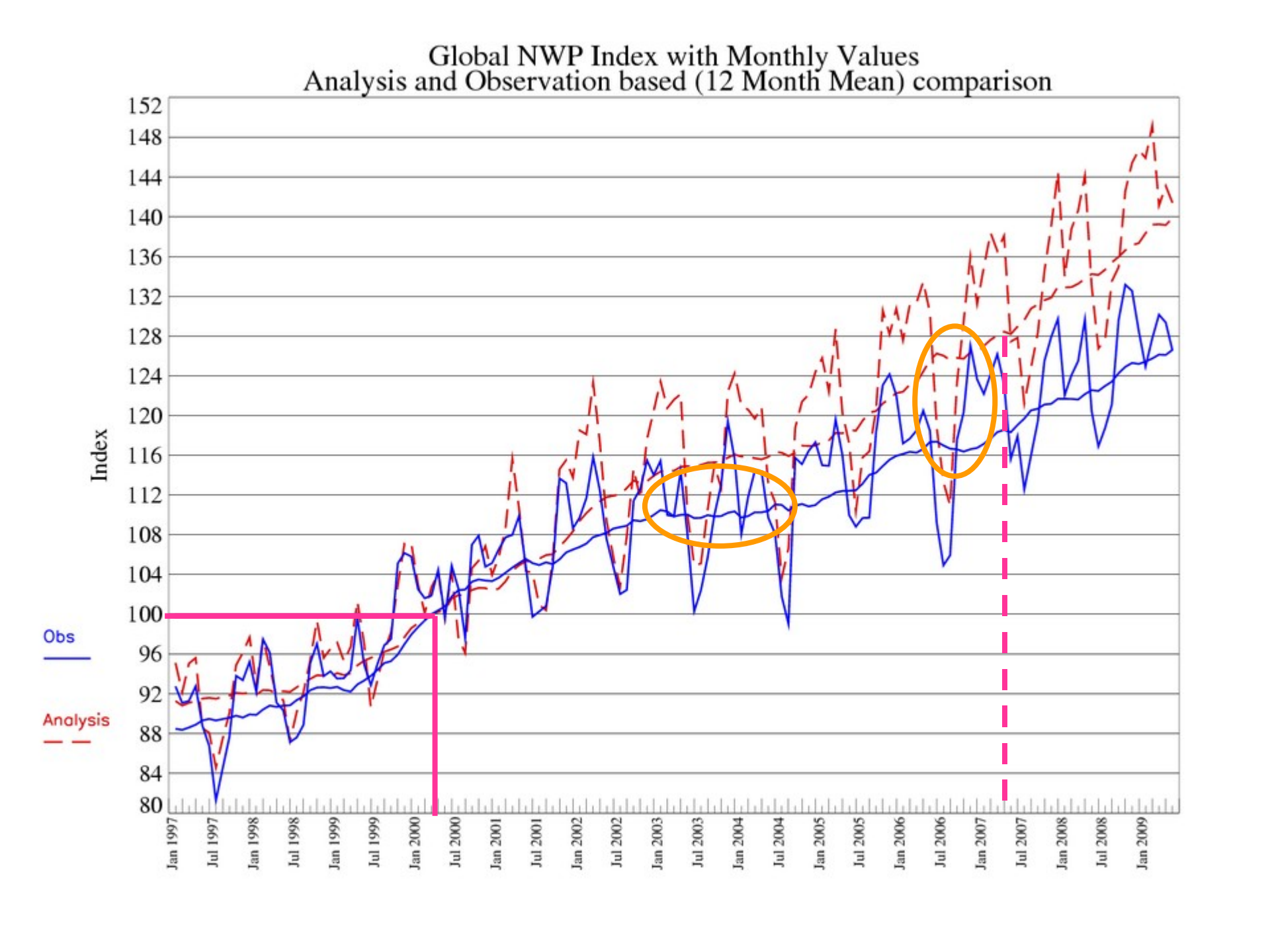
  $S=1-r^2/r^2_p$
- Against observations & analyses
- T+24h to T+120h, by 24h
- Weights reflect main customers/products

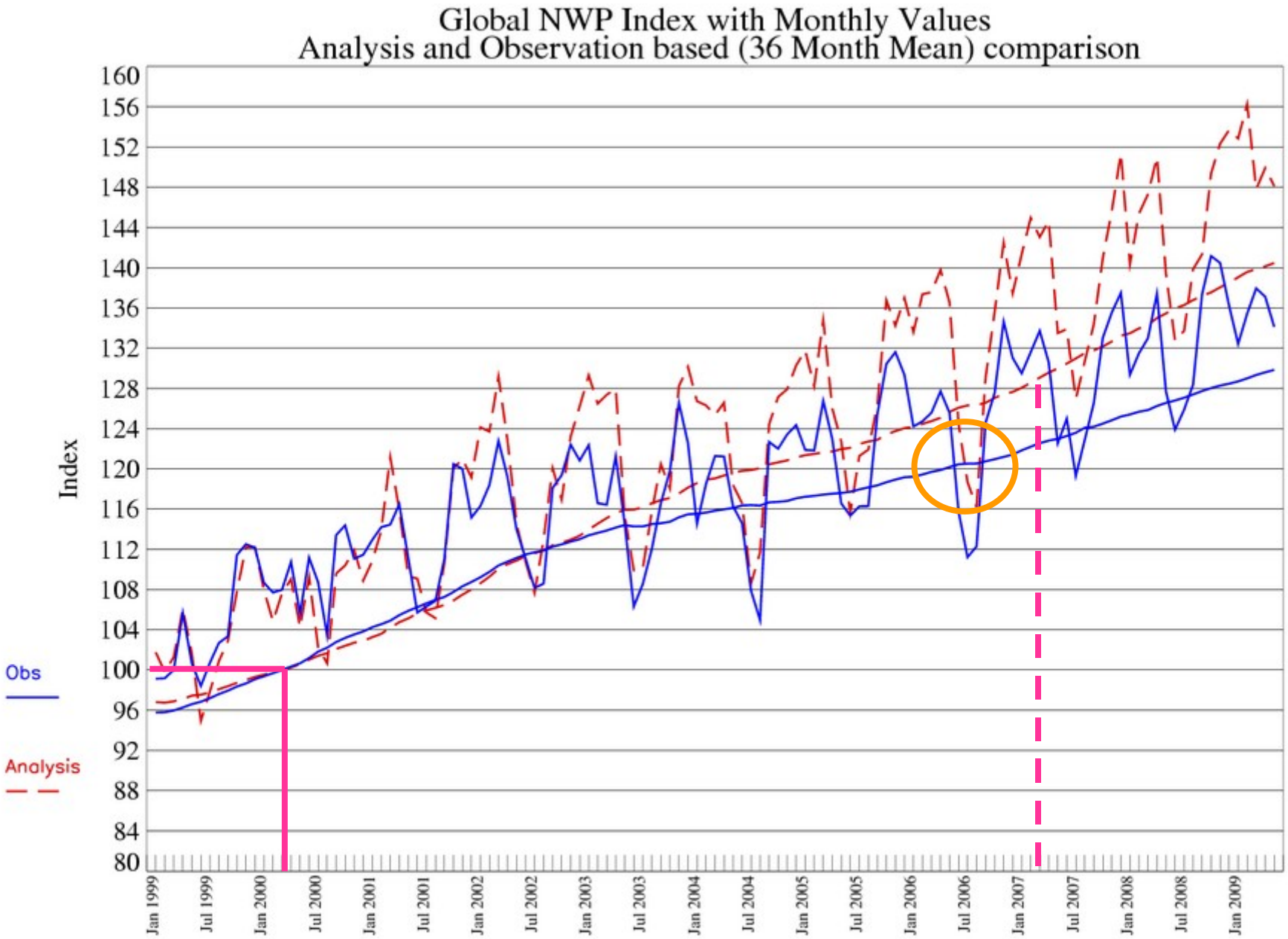
  $N=1/\sqrt{(1-S)}$
- 36-month running mean

  $I=100*N/N_o$,

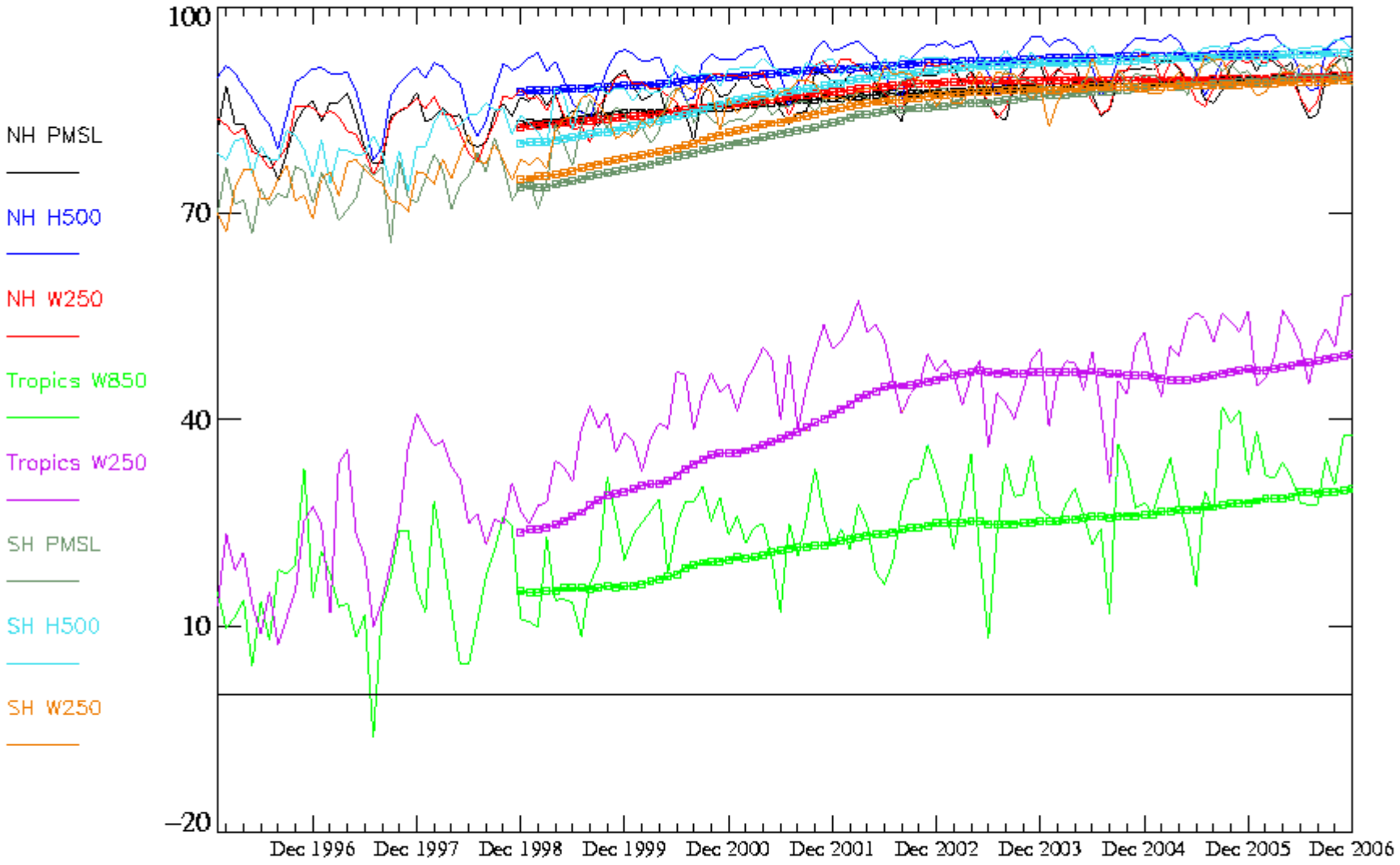
, N_o=value March 2000

I		Forecast Period				
		T+24	T+48	T+72	T+96	T+120
NH	PMSL	10	8	6	4	4
	H500	6	4	2	-	-
	W250	12	-	-	-	-
Tropics	W850	5	3	2	-	-
	W250	6	-	-	-	-
SH	PMSL	5	4	3	2	2
	H500	3	2	1	-	-
	W250	6	-	-	-	-

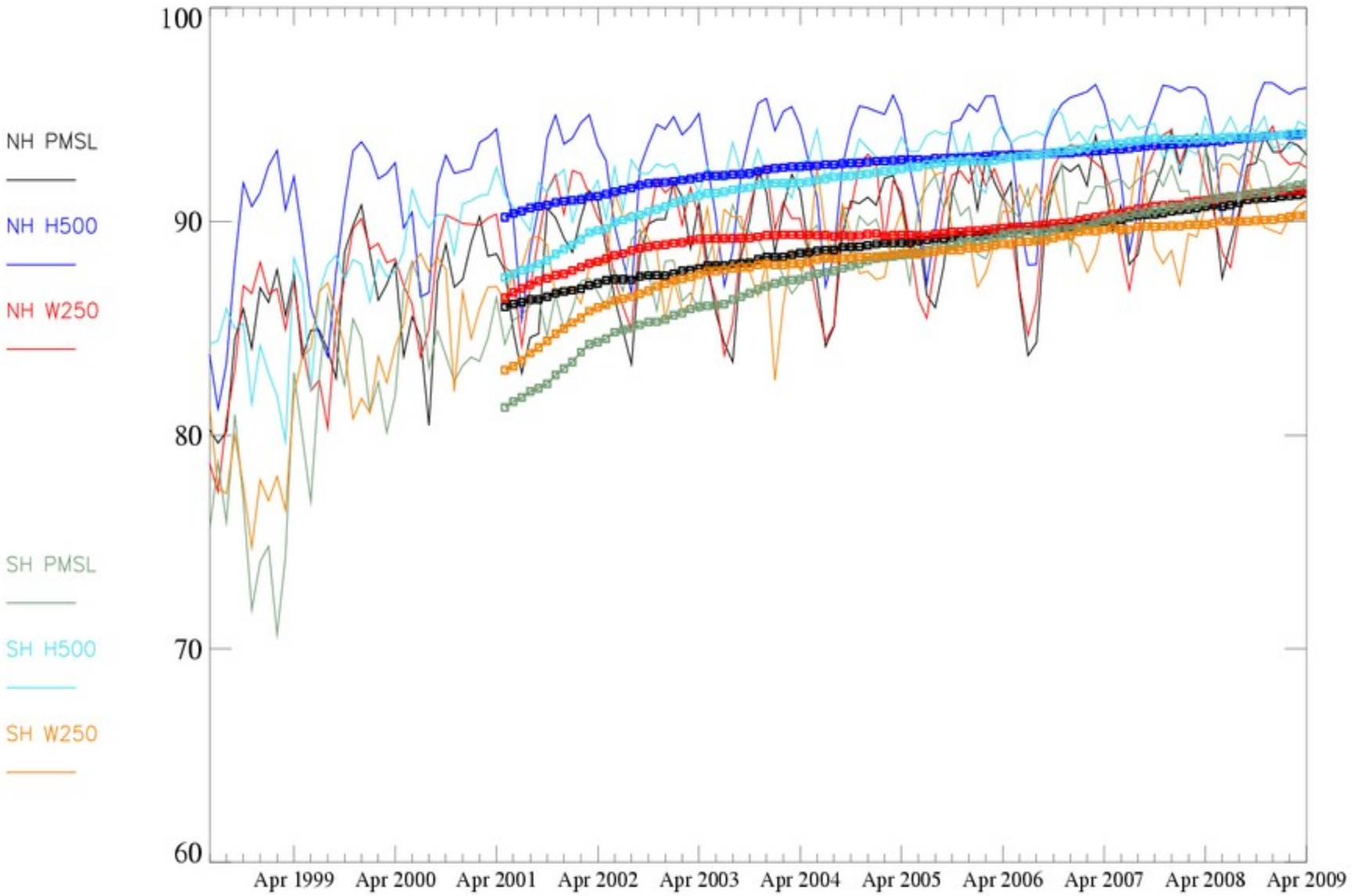

Global NWP Index with Monthly Values
Analysis and Observation based (12 Month Mean) comparison
Index
152
148
144
140
136
132
128
124
120
116
112
108
104
100
96
92
88
84
80
Obs
Analysis
Jan 1997
Jul 1997
Jan 1998
Jul 1998
Jan 1999
Jul 1999
Jan 2000
Jul 2000
Jan 2001
Jul 2001
Jan 2002
Jul 2002
Jan 2003
Jul 2003
Jan 2004
Jul 2004
Jan 2005
Jul 2005
Jan 2006
Jul 2006
Jan 2007
Jul 2007
Jan 2008
Jul 2008
Jan 2009


Global NWP Index with Monthly Values
Analysis and Observation based (36 Month Mean) comparison
Index
160
156
152
148
144
140
136
132
128
124
120
116
112
108
104
100
96
92
88
84
80
Jan 1999
Jul 1999
Jan 2000
Jul 2000
Jan 2001
Jul 2001
Jan 2002
Jul 2002
Jan 2003
Jul 2003
Jan 2004
Jul 2004
Jan 2005
Jul 2005
Jan 2006
Jul 2006
Jan 2007
Jul 2007
Jan 2008
Jul 2008
Jan 2009
Obs
Analysis

Skill Score components of Global NWP Index, Month and 36–Month Values, Observation based.



Skill Score components of Global NWP Index, Month and 36−Month Values, Observation based.

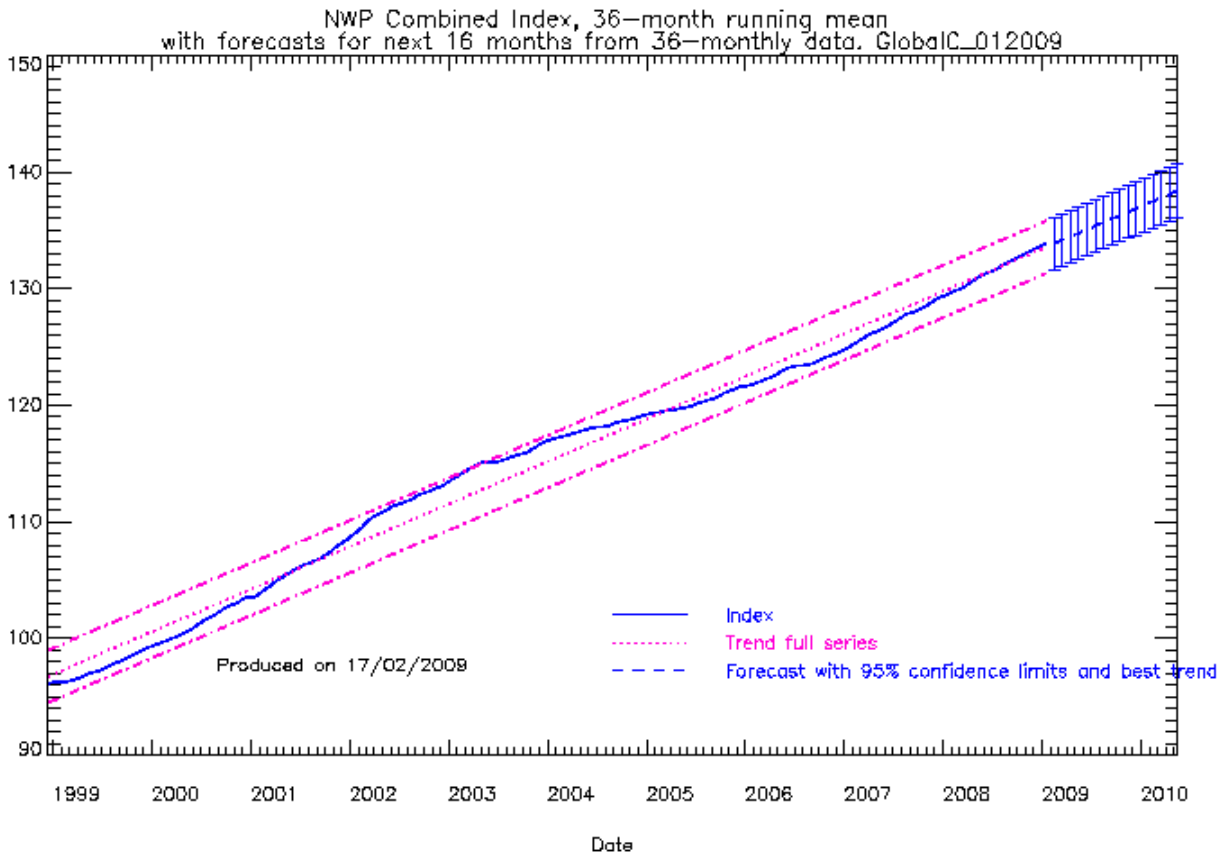


Global Index

Overall improvement modulated by :

- Inter and intra-annual variability
- Different rates of improvement against observations and analyses
  - Changes/few observations esp. tropics
  - Sea/land
- Some years flat – major upgrades >1y
  - Relocation of Met Office
- 12-month mean sometimes declines due to major differences in seasonal performance eg summer 2005 to summer 2006

Extrapolated trends for target setting -Global



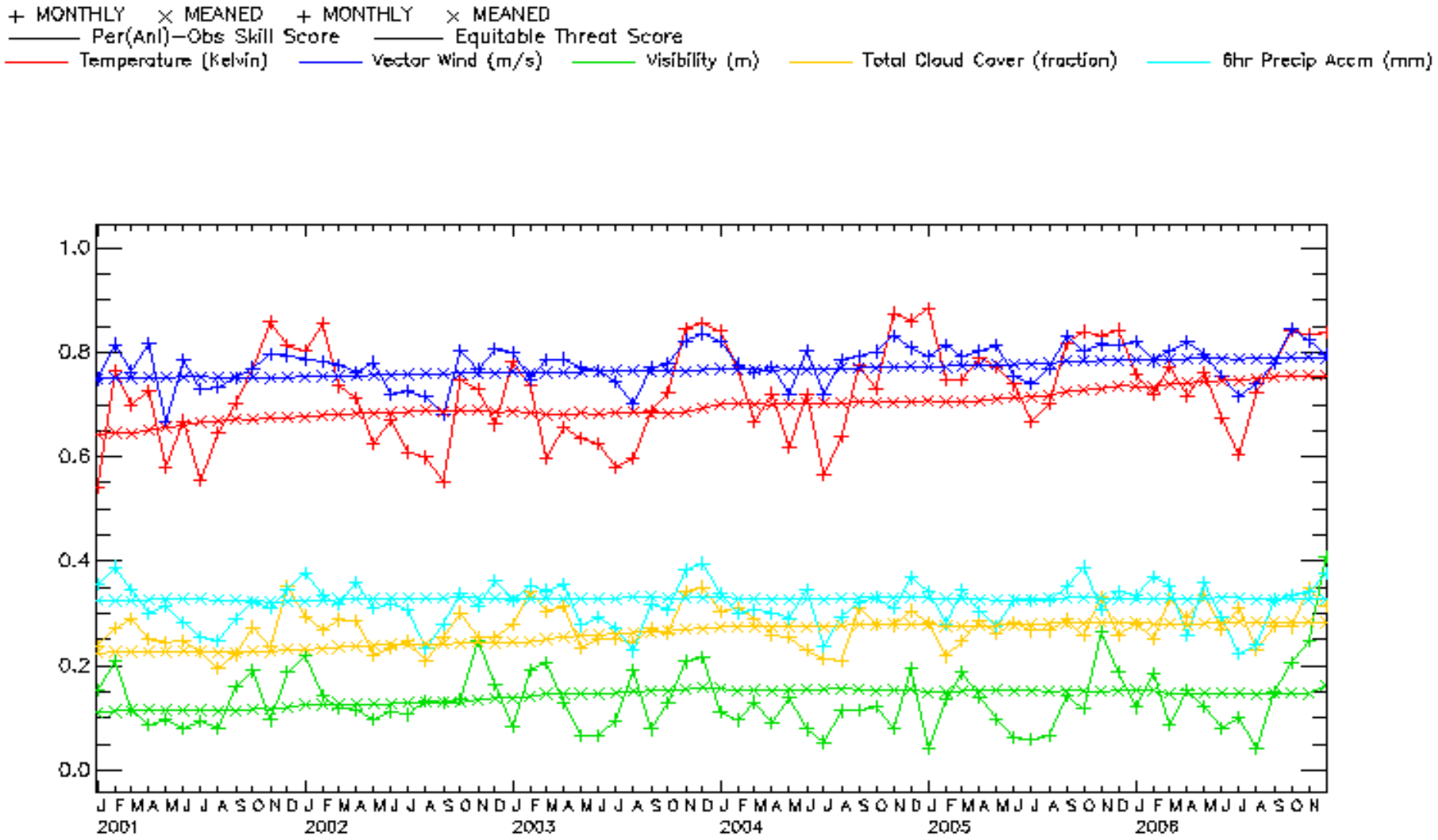
UK Index

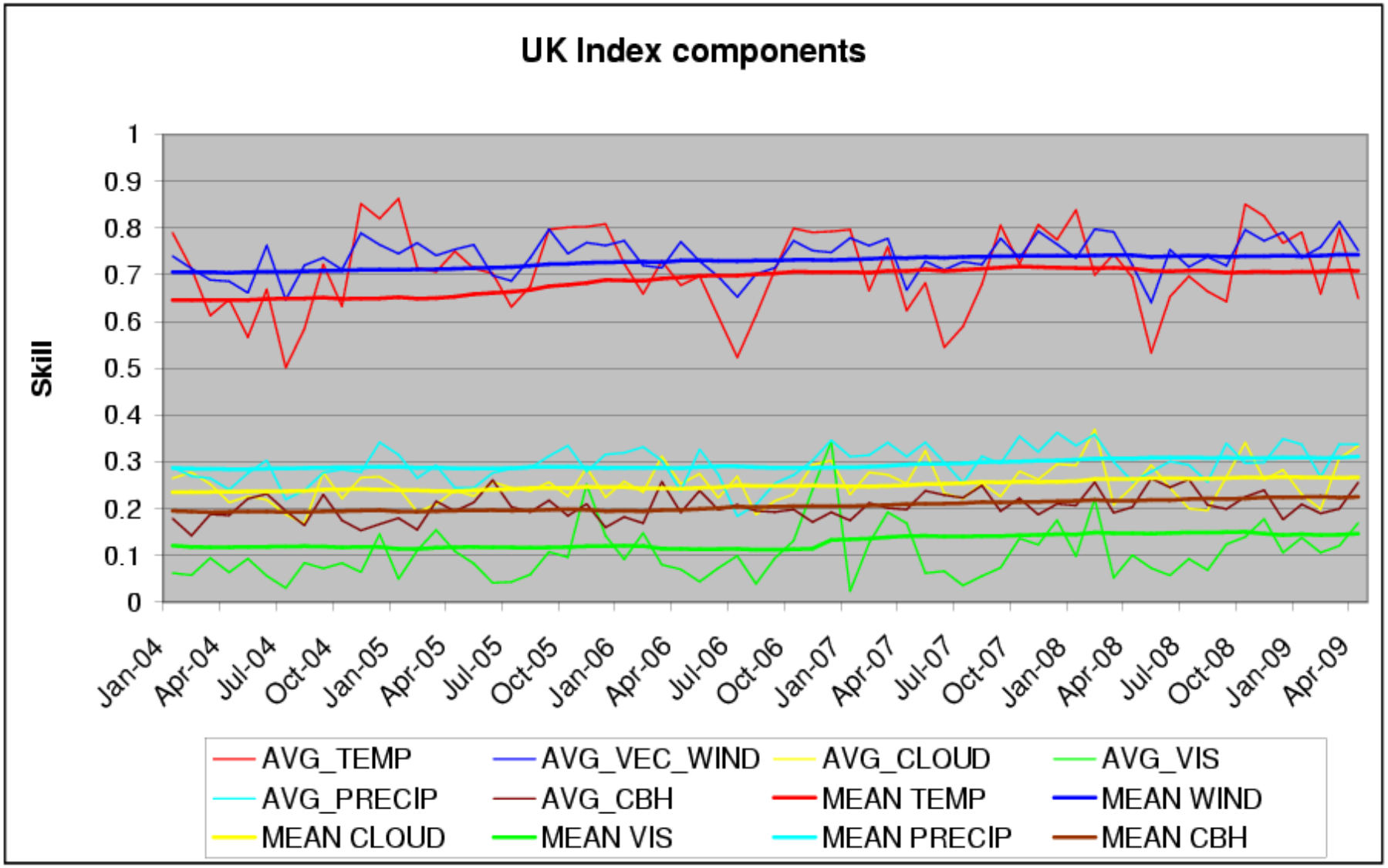
- Screen level Temperature, 10m wind
  - MSE skill
- 6h precipitation, visibility, cloud cover
  - Equitable threat score, 3 thresholds
  - 0.2,1.0,4.0 mm/6h; <5km,<1km,<200m; >2.5,4.5,6.5okta

  42 stations – now all in WMO block3 – Republic of Ireland
- Equal weights to T+6,12,18,24 (now to +48)
- Equal weights to each parameter
- 36 month contingency tables, running means
- $I=100*S/S_o$, S_o=value at March 2000

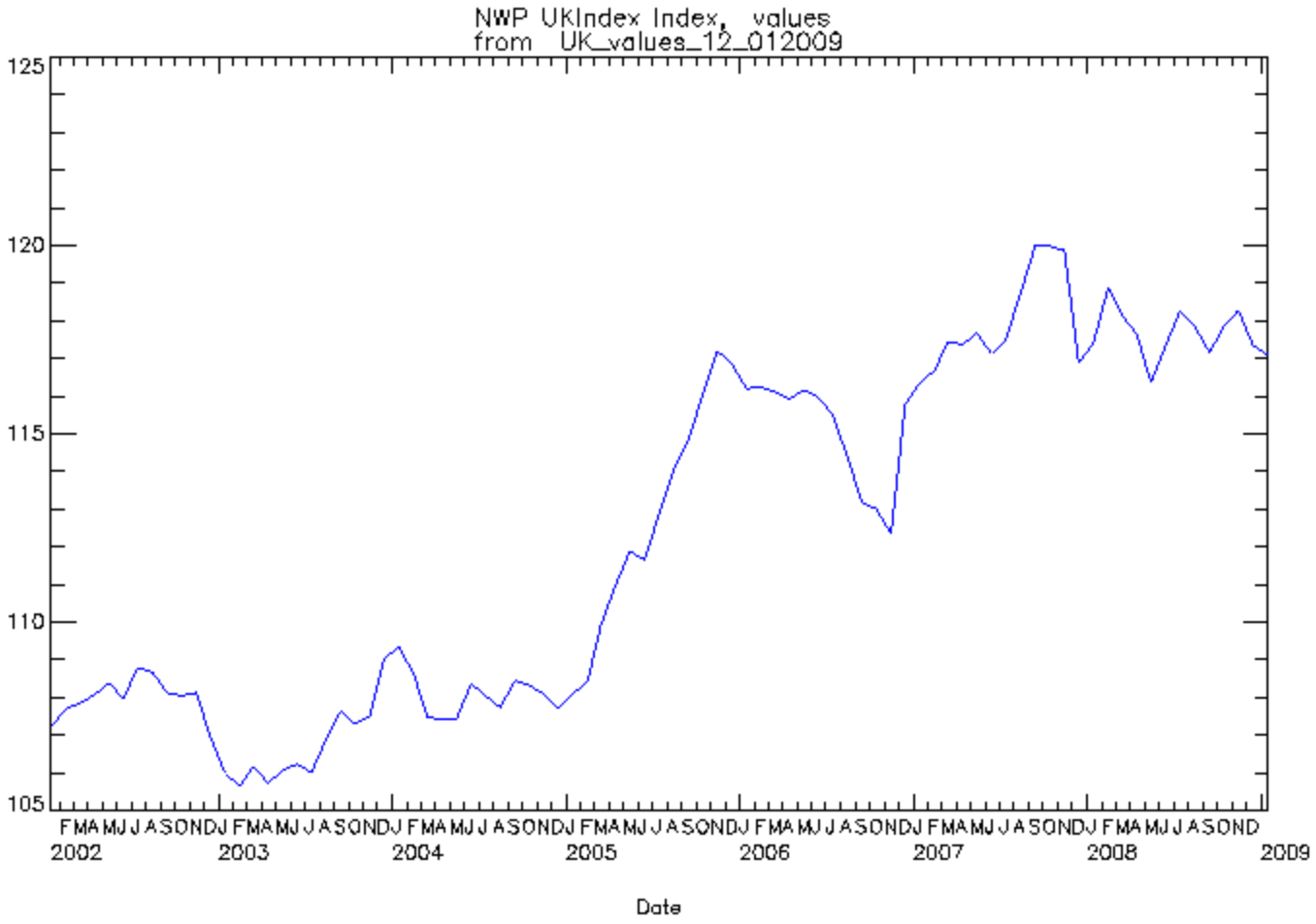
- NWP Index=0.5*(Global +UK)

UK: Combined times: Averaged forecast ranges and thresholds: Surface Obs

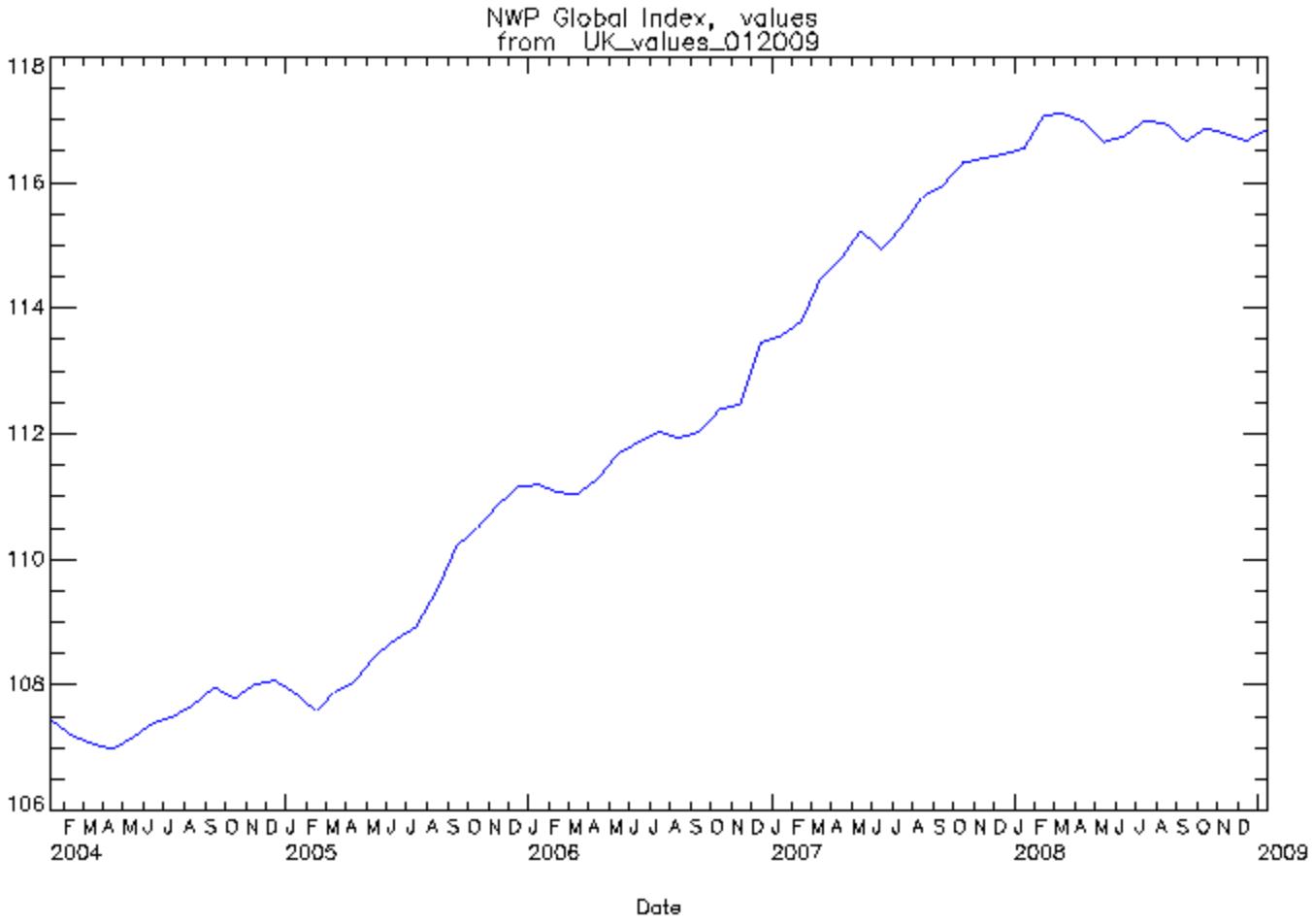




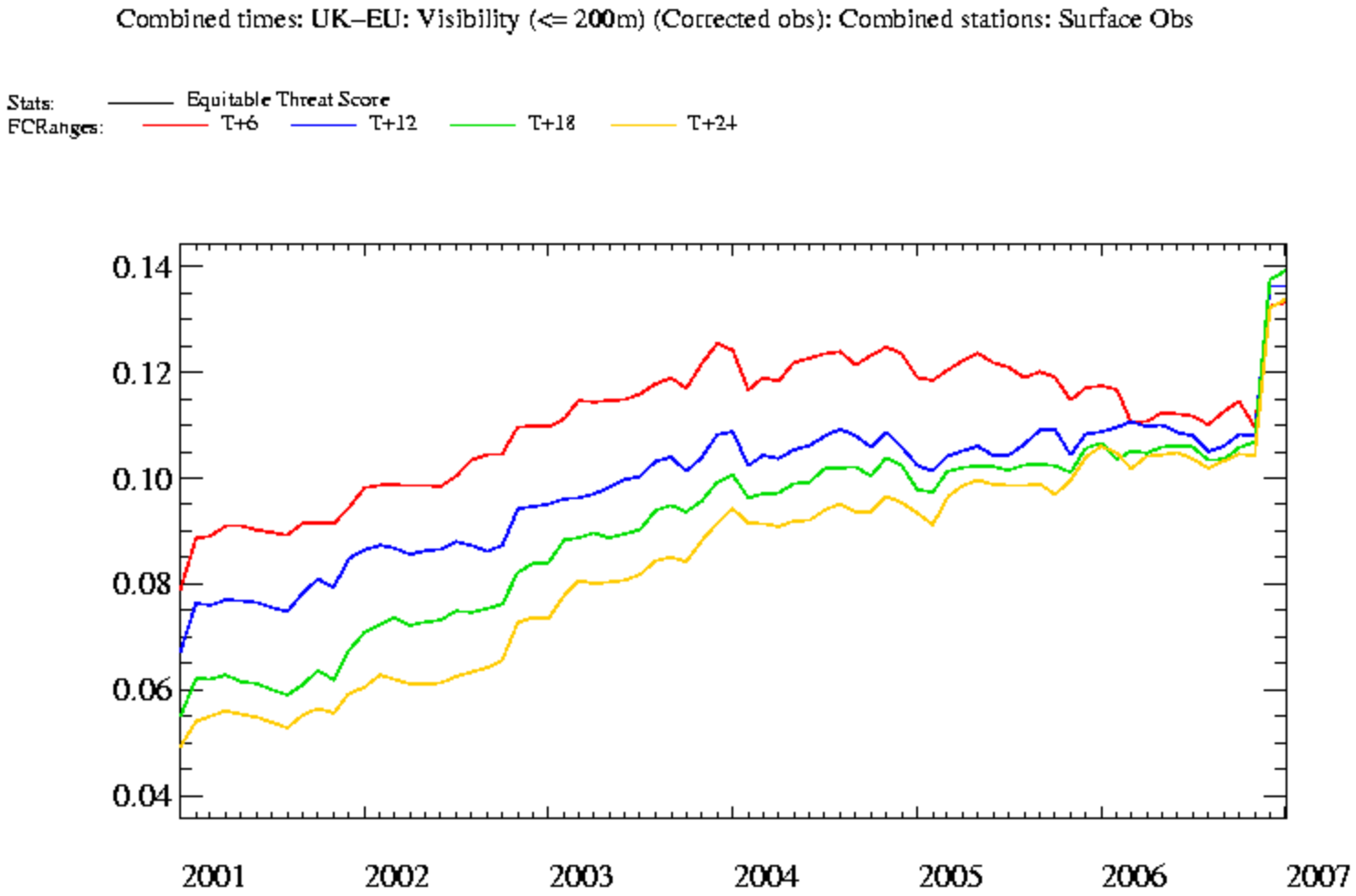
12-month UK index



36-month UK index



One good month !

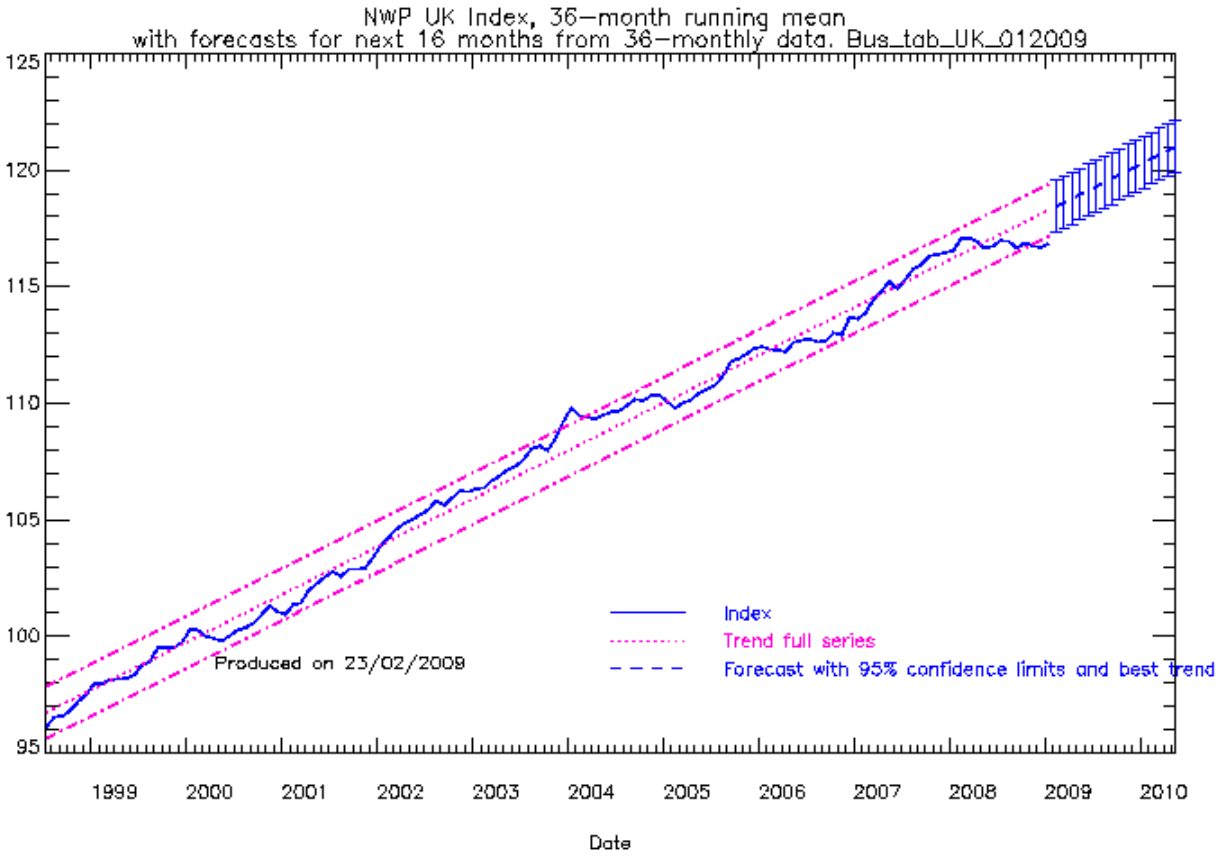


Station Based Verification System

UK Index

- Greater variability – smaller region, regime influence more important
- Even 36-month Index has negative trends for some periods
- Single parameters can have greater influence eg visibility
- Equitable threat score depends on base rate
  - Would prefer Odds ratio (benefit)

Extrapolated trends for target setting -UK



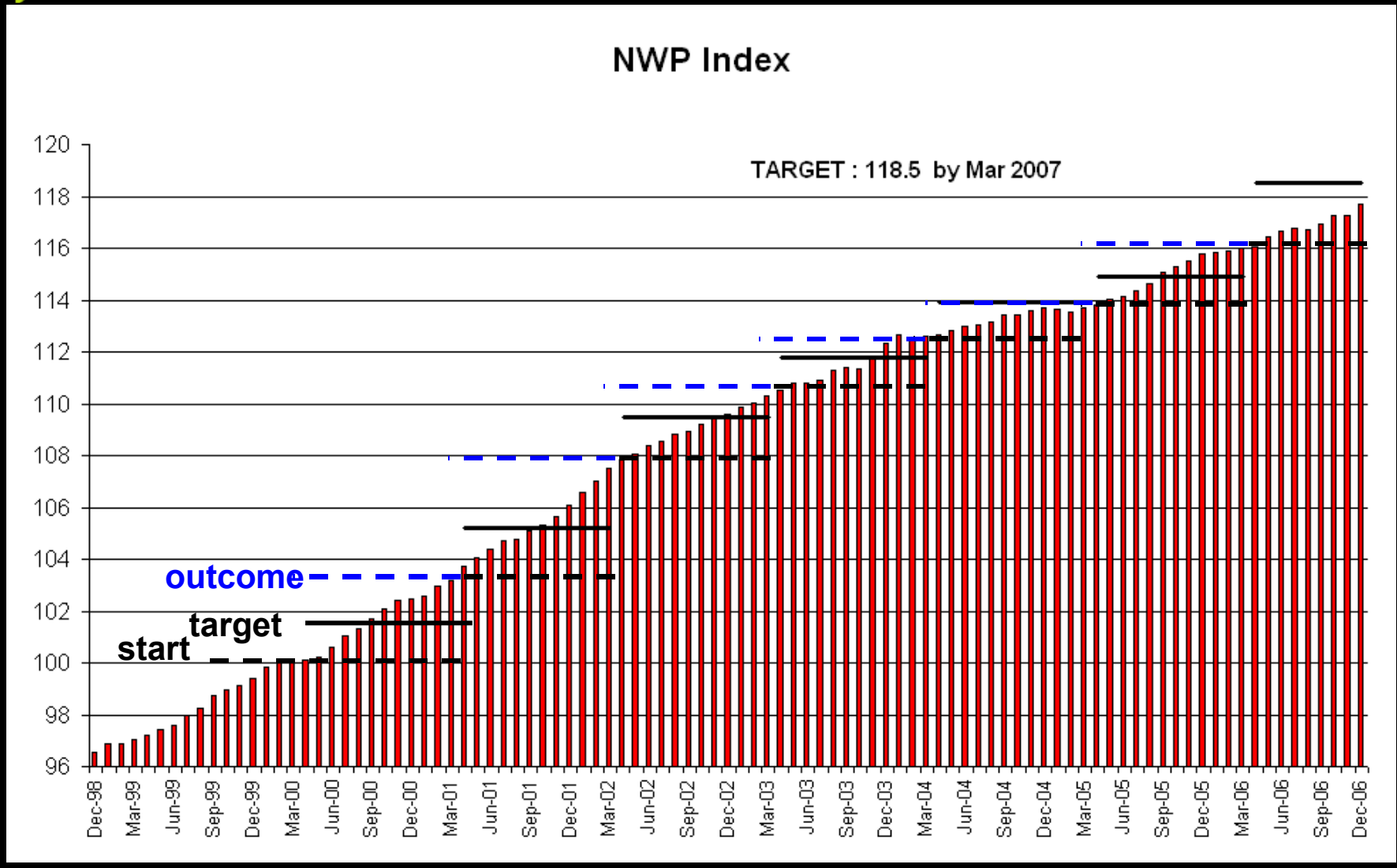
UK targets

- Easy to under/overshoot extrapolation
- Month to month variability may help/hinder achievement of target
- Even less scope to “manage”
  - Hard to combat regime dependence
  - Testing model changes on limited cases/periods may not be representative of overall impact
- Recent years the target has been for combined Global+UK index
  - Stretched based on larger global rate
  - More vulnerable to larger UK variability

Pass or fail assessment
A malign influence

- Nobody likes to be a failure
- If target is too stretching:
  - Demoralising – give up
- Too easy:
  - stakeholders suspicious/ not good value
  - More likely to increase target for following year
- If progress towards target slows or declines:
  - Panic measures to try and put back on course
  - Cherry picking upgrades but limited influence on 3y mean
    - Better to diagnose what is cause – behaviour not shown in (inadequate) testing

How well have we done ?



Past record

Combined Global & UK

date	Target	Outturn	Target increase	Outturn -target
2000	**100.**			
2001	**101.6**	**103.2**	**1.6**	**1.6**
2002	**105.2**	**107.5**	**2.0**	**2.3**
2003	**109.5**	**110.0**	**2.0**	**0.5**
2004	**111.8**	**112.5**	**1.8**	**0.7**
2005	***113.9***	***113.7***	***1.4***	***-0.2***
2006	**114.9**	**116.0**	**1.2**	**1.1**
2007	**118.5**	**120.1**	**2.5**	**1.6**
2008	**122.4**	**123.5**	**2.4**	**1.1**
2009	***125.8***	***125.8***	***2.3***	***0.0***

How to improve use of KPT

- Abandon pass/fail
  - Set interval range and give credit for progress within that
  - Interval accounts for likely impacts and variability
- Annual targets need to be modulated by known risks eg relocation
- If longer term “aspirations” are set , even more important not to use pass/fail
  - Eg improve by [x to y] % over 3years based on past average improvement and uncertainty
- Use comparative measures against other centres to reduce regime influence

Consequences of Indexes

- All upgrades need to have positive or neutral impact – “objective criterion” for model development
  - Sometimes bundle changes
  - But need to look at individual components of upgrade
- Some parameters receive greater attention
  - Tropical winds
- Customers want simple idea of improvement
  - But do not understand skill and compositing
  - Need to look at specific parameters of interest to individual customers

Do they work ?

- As decision criterion for upgrades – yes
- As motivation for united team effort – probably NOT, but people still care about the bonus!
- As management tool for controlling effort – unrealistic expectations, damaging influence, change in priorities - NO
- Demonstrating improved performance/progress – yes,
  - but can encourage gaming/hedging & lack of integrity
  - Eg forecasters trying to maximise PoP Brier score to achieve target

Postscript
Mason & Weigel (2009)

- Generic framework – two alternative forced choice (2AFC)
  - $p_{2afc} > 0.5$, unskilful guess = 0.5
- Dichotomous forecasts
  - Rescaled Peirce SS, $p_{2afc} = ½ (PSS+1)$
- Continuous forecasts
  - Rescaled Kendall's correlation coefficient,
  - $p_{2afc} = ½ (\tau+1)$
- Still need compositing/index for overall performance measure