Second Verification Workshop
CBRFC, 11/20/08

Identifying and reducing bias in real-time ensemble forecasts

James Brown, Dong-Jun Seo

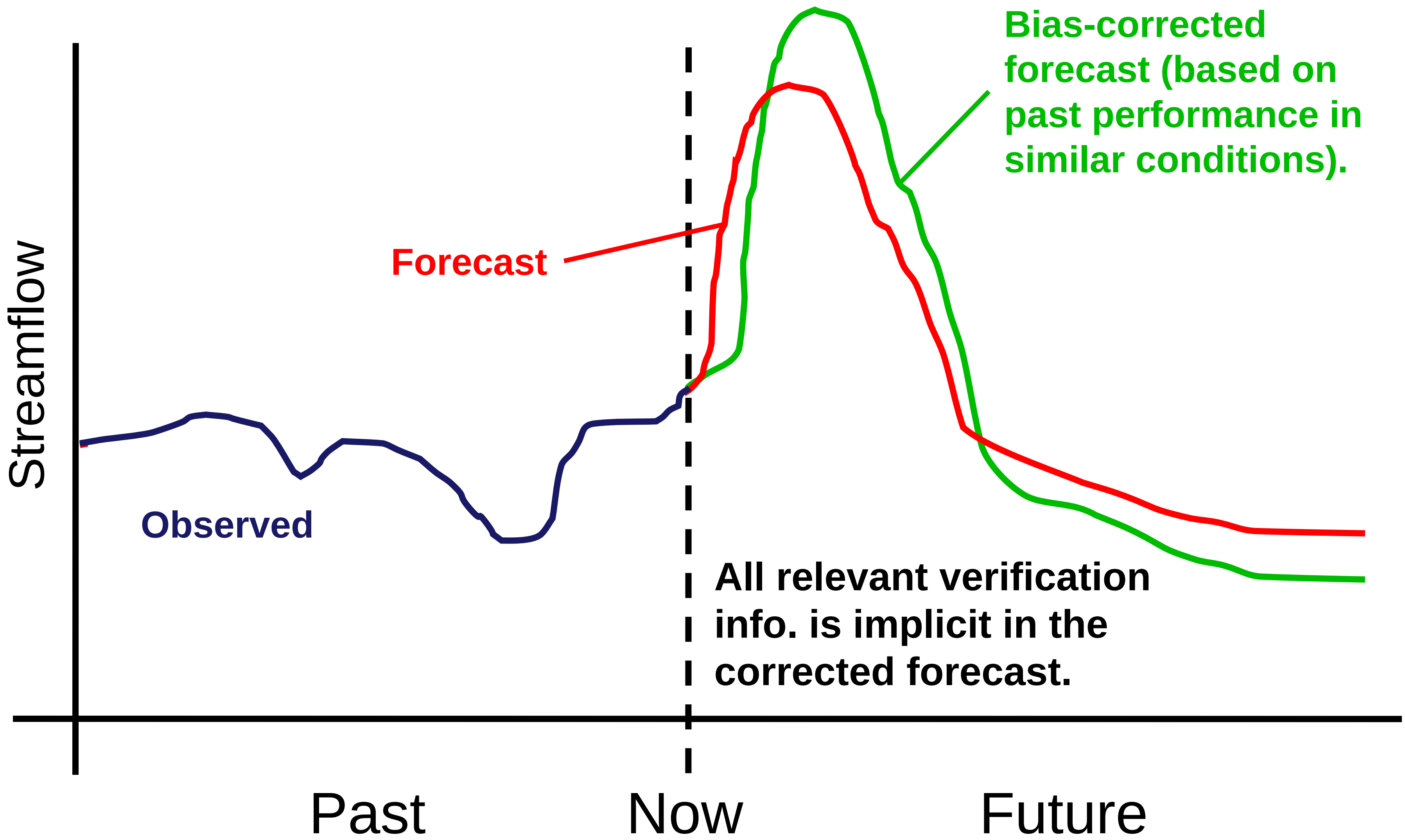
james.d.brown@noaa.gov

Contents

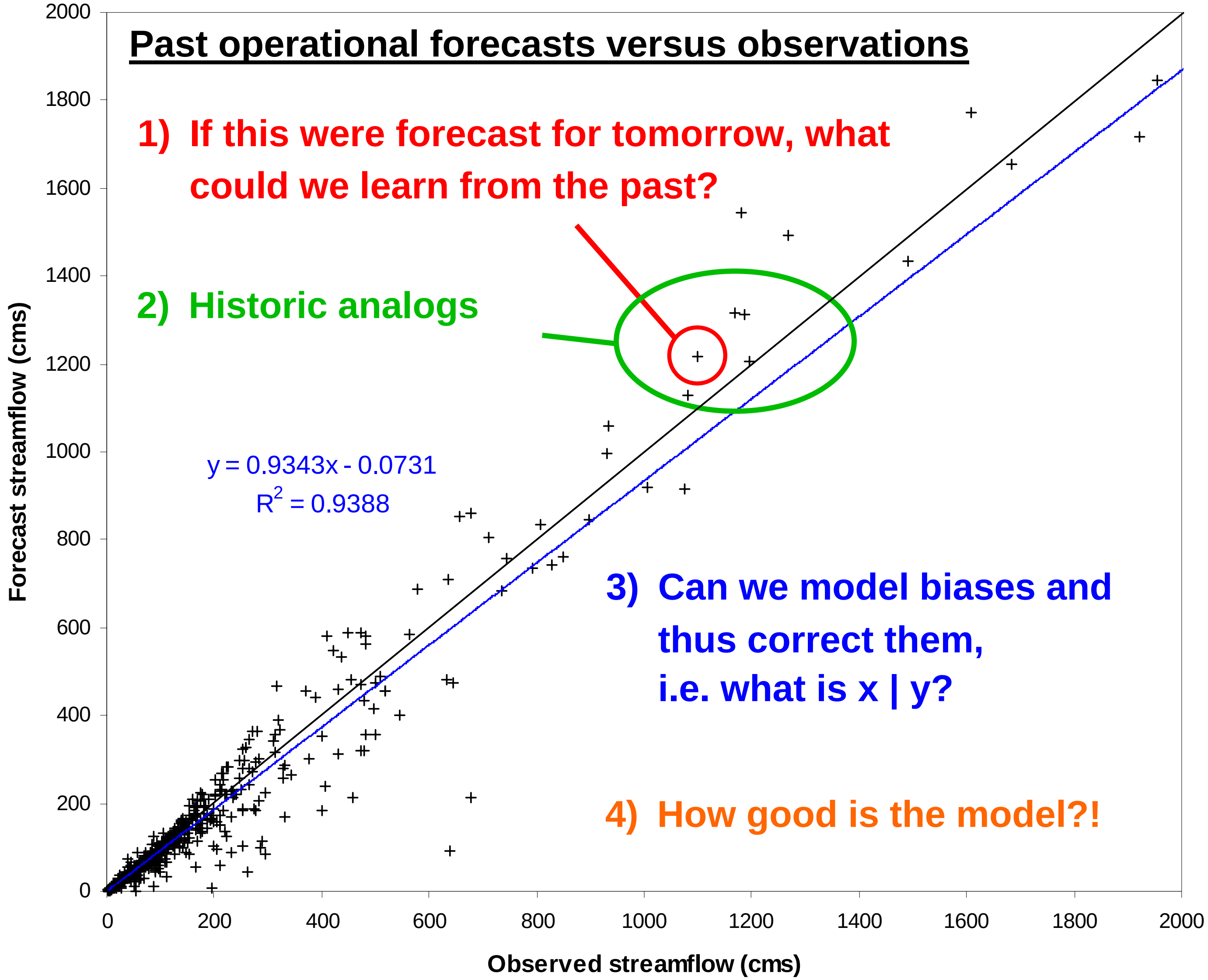
1. **Problem of real-time verification**
   - **Diagnostic metrics too cumbersome….**
   - **….not tailored to live forecast situation**
   - **Biases of historic analogs = a guide to future**
2. **Real-time bias correction technique**
   - **Non-parametric (precipitation, flow)**
3. **Some example results**
   - **GEFS precipitation and ESP streamflow**

1. Problem definition

Single-valued example



Single-valued example



y = forecast
x = observed

***f*(x|y)?**

Linear ***f***:

y = 0.93x - 0.073
x = (y+0.073)/0.93

f(x|y= 1220)=

$$\frac{(1220 + 0.073)}{0.93}$$

=1312 cms

Unknown 'truth'
= 1112 cms)

Two parts to problem

1. **Modeling 'truth' (x) given forecast (y)**

- **How to model `f(x|y)`**
- **Do we need to add conditions, `f(x|y,s)`?**
- **Example: `s` could be ice blocking flow**

2. **Identifying/visualizing historic analogs**

- **How to identify and visualize analogs to `y`?**
- **Important because `f(x|y,s)` is only a model**
- **This is not easy. So far, we focused on (1)…**

How to model f(x|y) if y is an ensemble forecast?

What if y is an ensemble?

Same basic concept:

X = observed (unknown for live forecast)

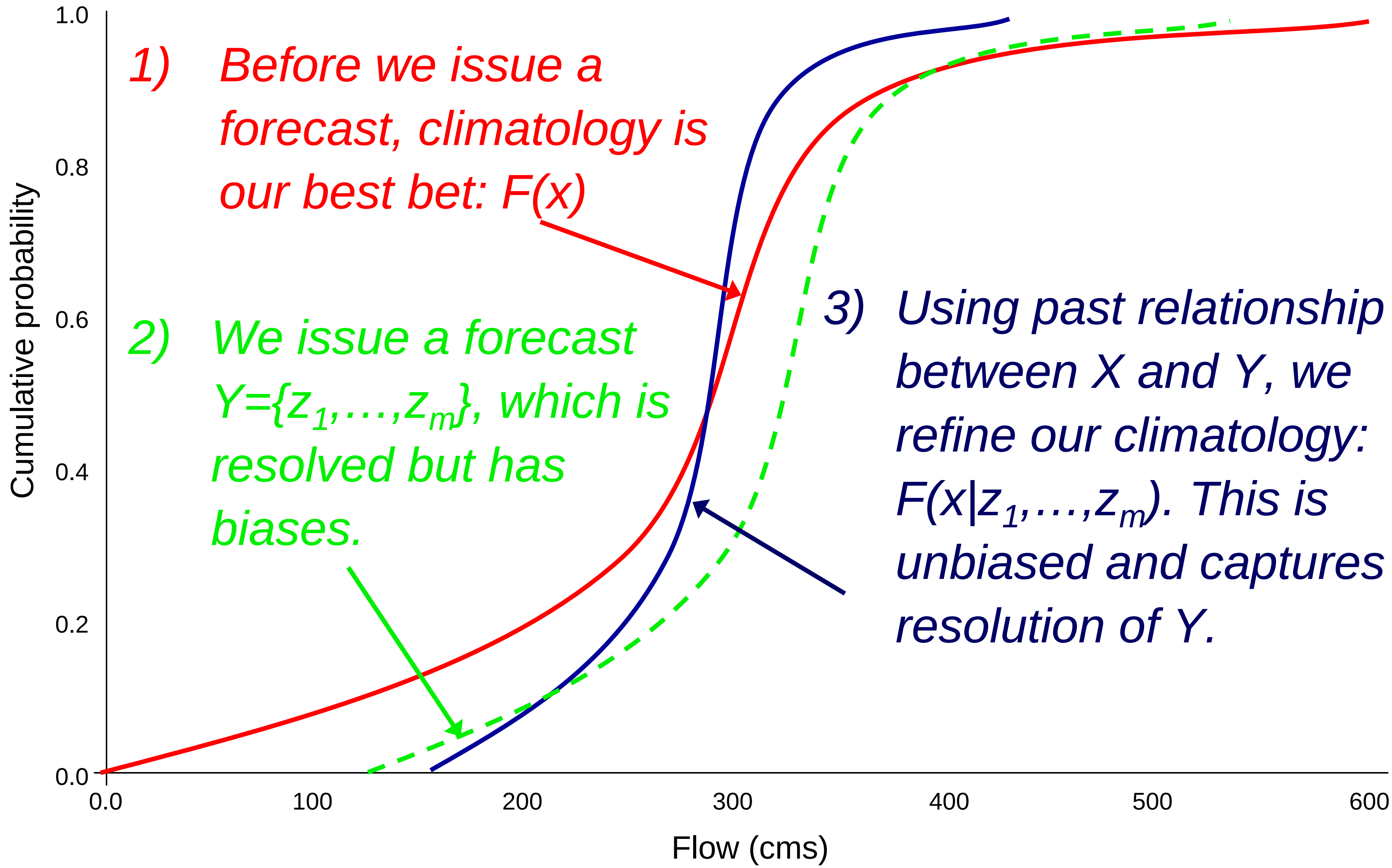
$Y = \{Z_1, \ldots, Z_m\}$ = live ensemble forecast

The aim is to model (from past data):

$$F(x|z_1,\ldots,z_m) = \text{Prob}[X \leq x|z_1,\ldots,z_m] \quad \forall x$$

i.e. what is observed ("true") probability dist. given the real-time forecast (based on past relation between forecast and "truth").

What if y is an ensemble?



How to model?

- We need to model $F(x|z_1,\ldots,z_m)$
- No single 'parametric' model for all forecast types (e.g. joint normal)…
- …data transform (e.g. normal-score transform) is often tricky
- What about a non-parametric model, driven by what the data tell us?

2. Indicator approach

Indicator approach

- What is the probability that a dice throw, X, is $\leq$ 3?
- Take *n* samples of X = {1,2,6,4,2,5,1,3}
- <u>Answer:</u> average no. of times X $\leq$ 3:

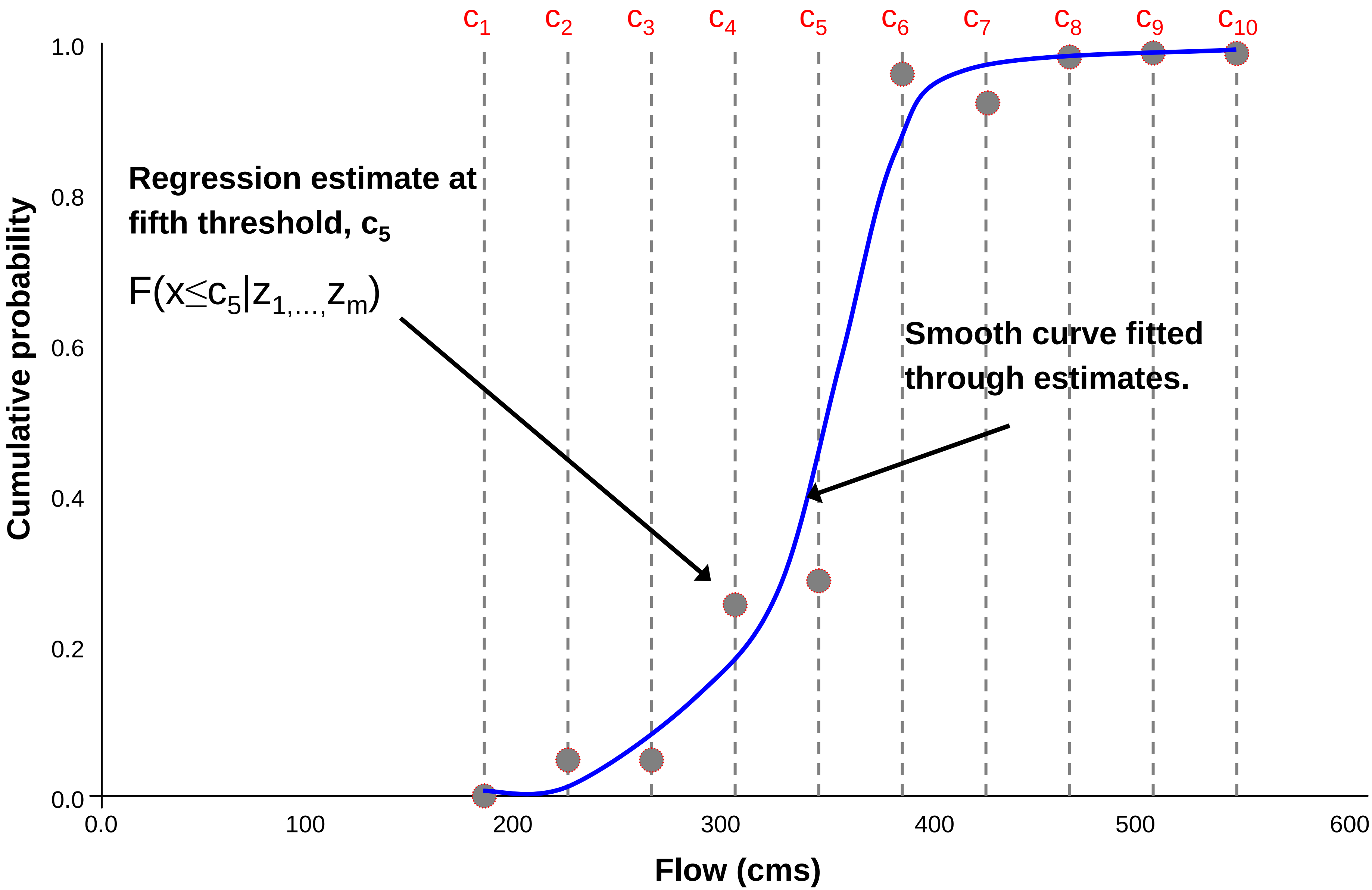
$$\mathrm{Prob}[X \leq 3] \approx \frac{1}{n}\sum_{j=1}^{n} I_X(x_j) \quad \text{where} \quad I_X(x_j) = \begin{cases} 1, x_j \leq 3 \\ 0, \text{otherwise} \end{cases}$$

- Expectation of an indicator function
- Repeat for all possible x, we get full pdf

Indicator approach

- Our problem is much tougher. We cannot simply count samples. We have way too many indicator variables, so many combinations not observed.
- How to fill in the blanks?
- We use multiple indicator (linear) regression.

Indicator approach



3. Results

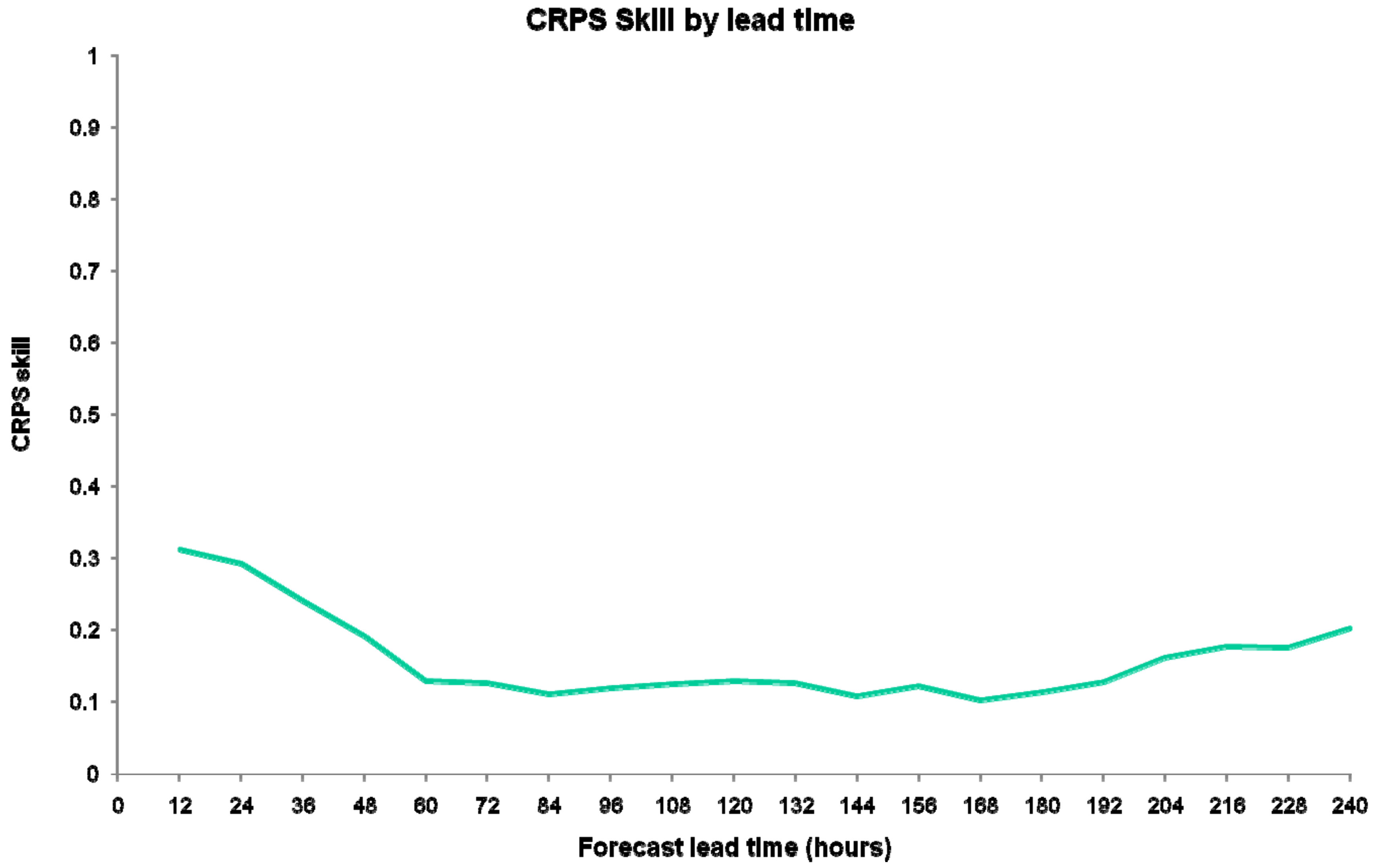
GFS precipitation

- Ensemble precipitation (12-hourly) from operational GEFS, 2000-2005.
- Precipitation is a tough test (intermittent and highly skewed).
- Verified raw GEFS ensembles with indicator-corrected forecasts in Juniata, PA (MAP used as observed).
- Split sample (independent) verification by rotating sample data.

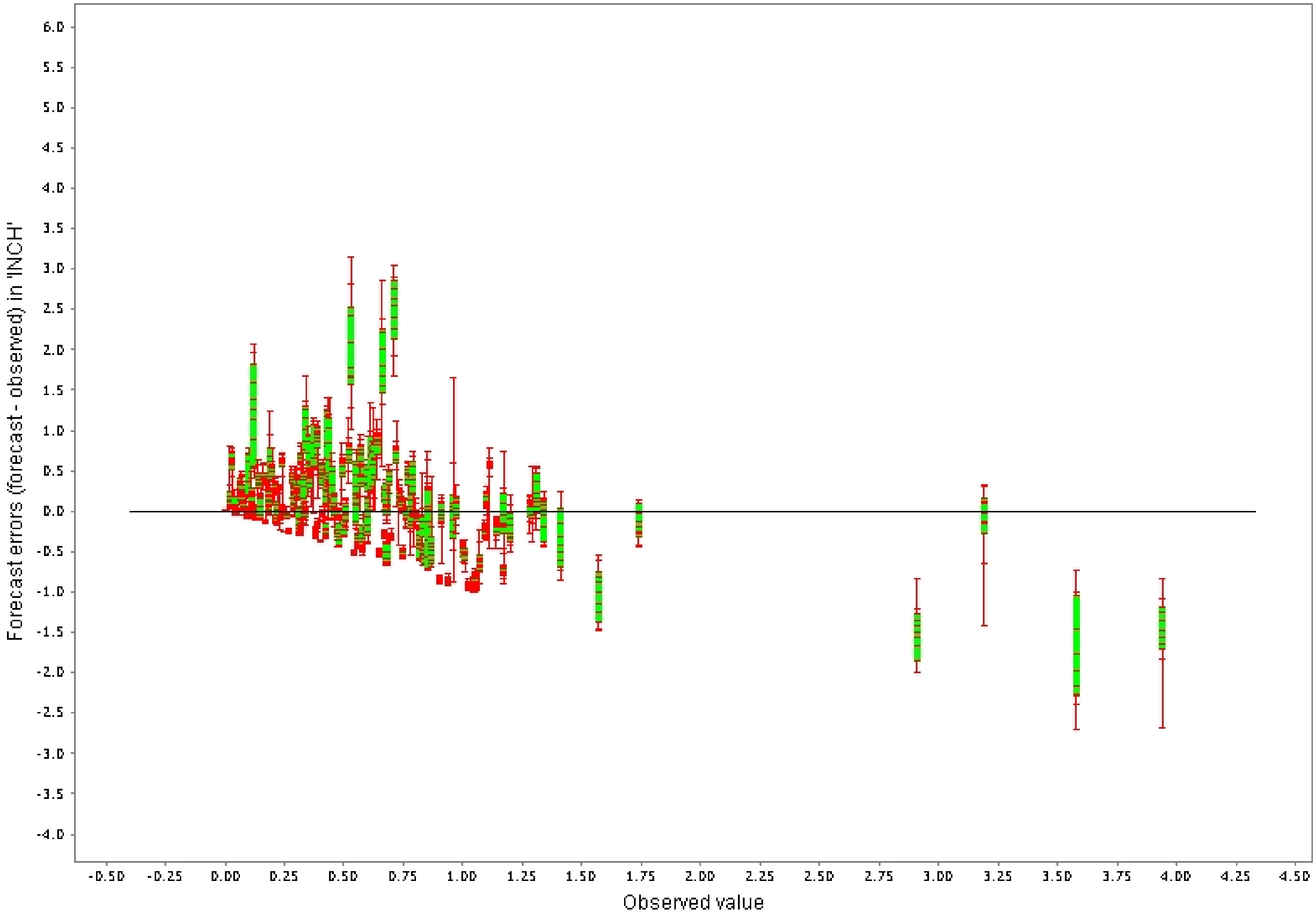
Summary of results

- The raw GEFS ensembles were surprisingly good.
- Indicator-corrected ensembles were ~30% better by CRPS. (The indicator approach explicitly minimizes CRPS).
- The indicator-corrected ensembles were significantly more reliable.
- Very similar quality to EPP for days 1-2, and much better beyond.

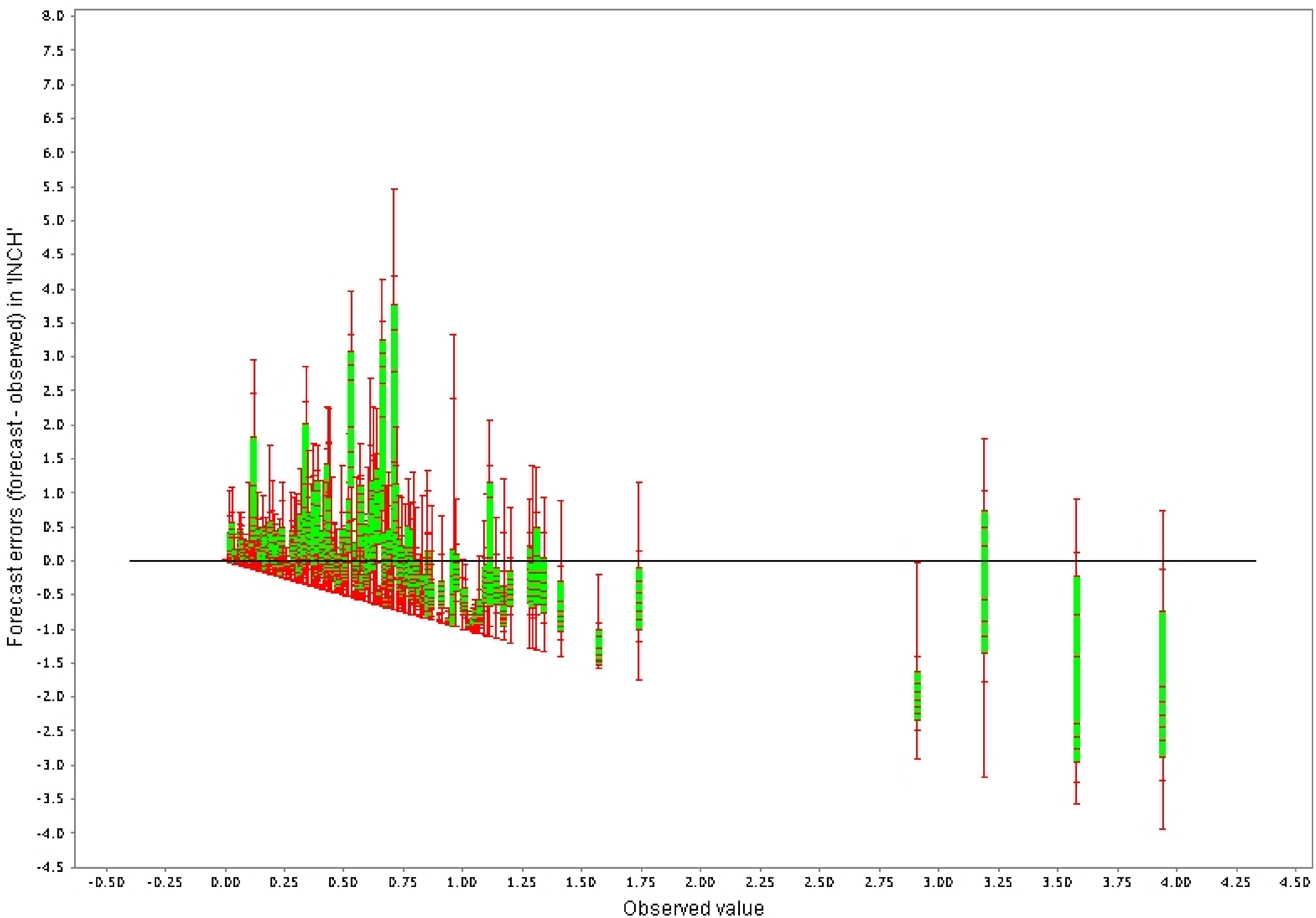
Summary of results



Modified box plot of ensemble forecast errors against observed value.
Real.Time.Verification.GFS_ensembles at lead hour 12



**Modified box plot of ensemble forecast errors against observed value.
Real.Time.Verification.Cond_obs_GFS at lead hour 12**

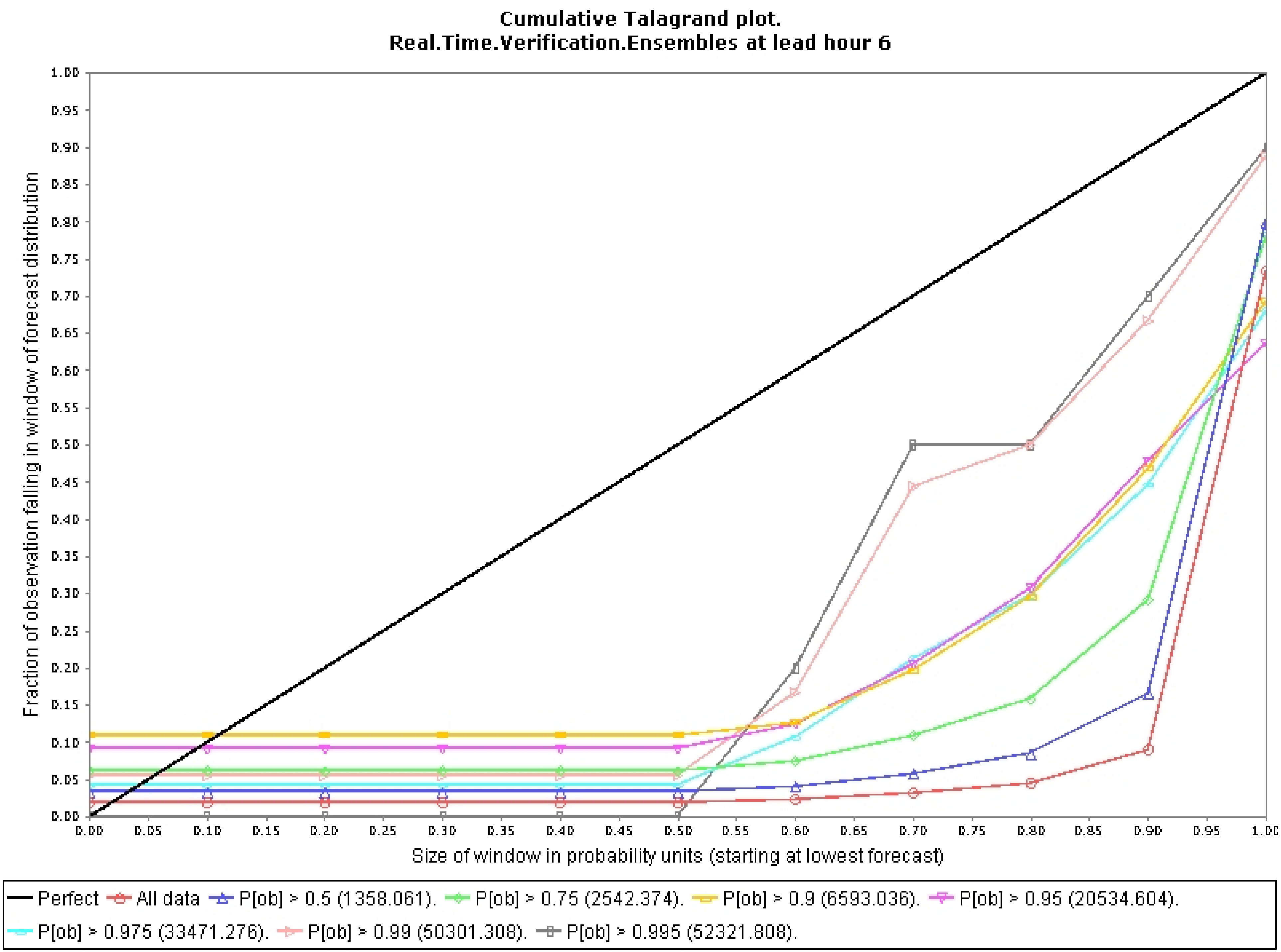


ESP flow

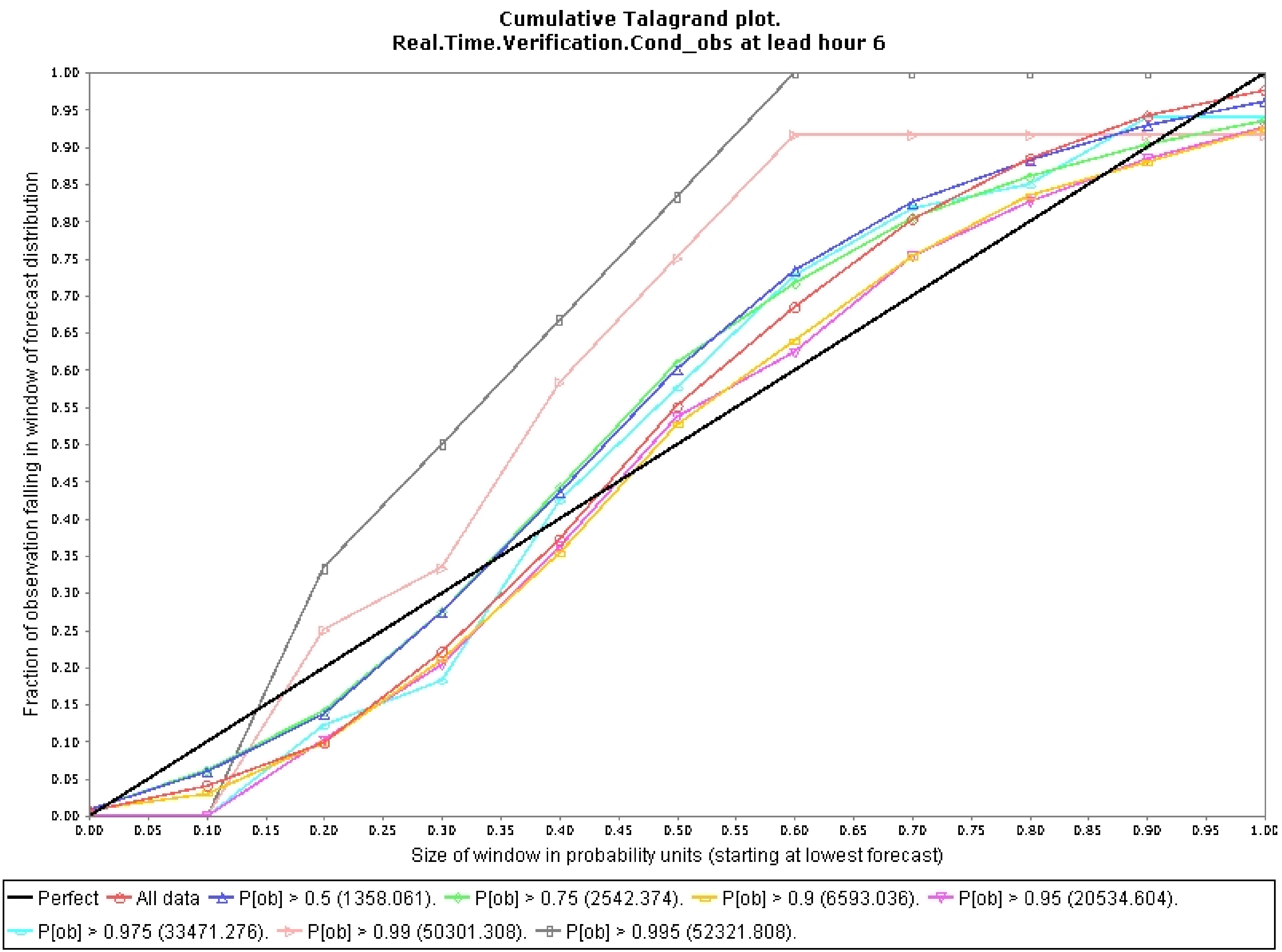
- ESP forecasts from 2003-2008 for QUAO2 in ABRFC.
- Used RFC flow observations for indicator-correction/verification.
- Split sample (independent) verification by rotating sample data.

Summary of results

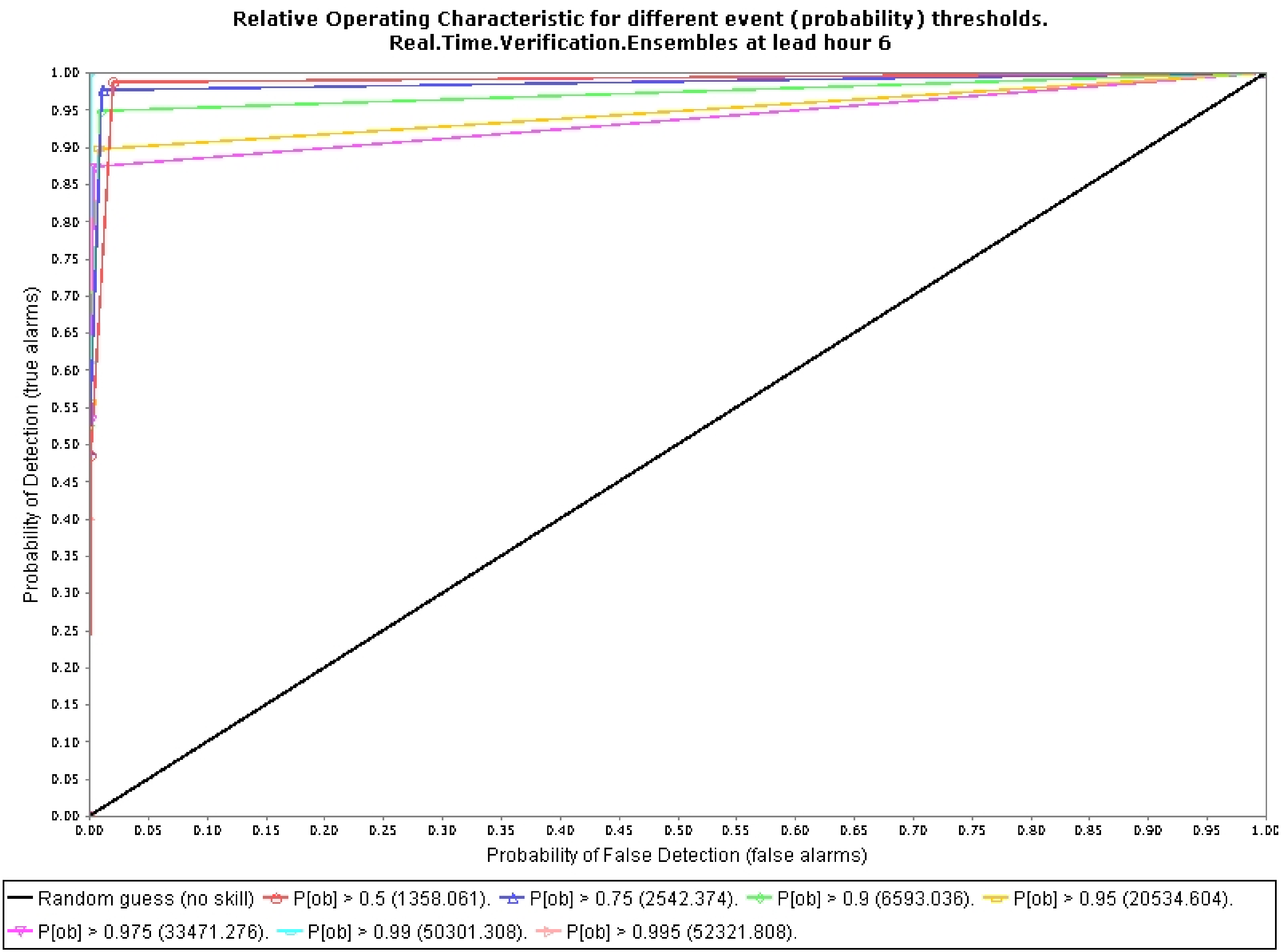
- The raw ESP ensembles were surprisingly bad (lack of hydro-uncertainty).
- Indicator-corrected ensembles were up to ~70% better by CRPS.
- The indicator-corrected ensembles were much more reliable and resolved.

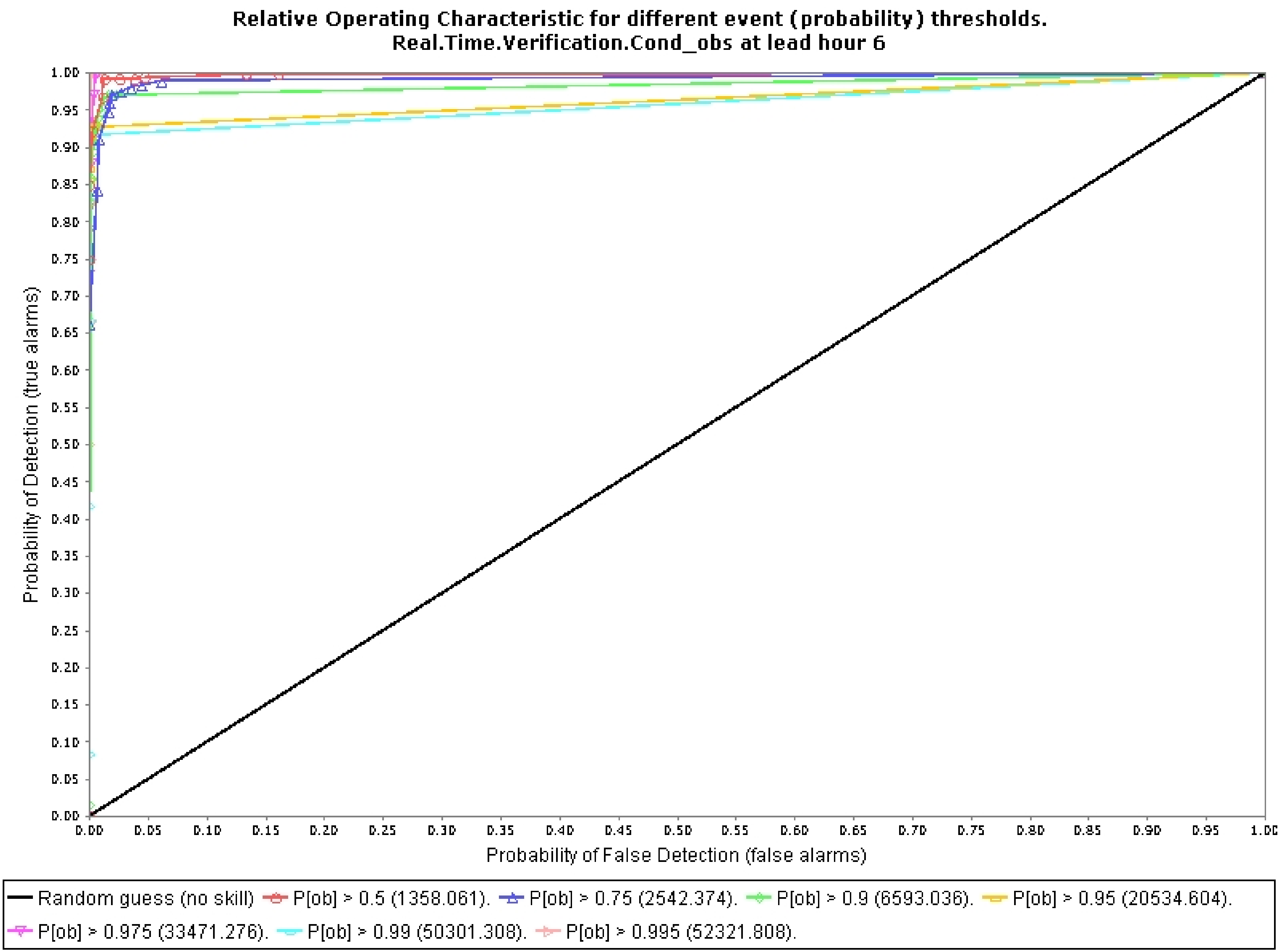


Cumulative Talagrand plot.
Real.Time.Verification.Ensembles at lead hour 6
Fraction of observation falling in window of forecast distribution
Size of window in probability units (starting at lowest forecast)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Perfect
All data
P[ob] > 0.5 (1358.061).
P[ob] > 0.75 (2542.374).
P[ob] > 0.9 (6593.036).
P[ob] > 0.95 (20534.604).
P[ob] > 0.975 (33471.276).
P[ob] > 0.99 (50301.308).
P[ob] > 0.995 (52321.808).



Cumulative Talagrand plot.
Real.Time.Verification.Cond_obs at lead hour 6
Fraction of observation falling in window of forecast distribution
Size of window in probability units (starting at lowest forecast)
1.00
0.95
0.90
0.85
0.80
0.75
0.70
0.65
0.60
0.55
0.50
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.00 0.05 0.10 0.15 0.20 0.25 0.30 0.35 0.40 0.45 0.50 0.55 0.60 0.65 0.70 0.75 0.80 0.85 0.90 0.95 1.00
Perfect
All data
P[ob] > 0.5 (1358.061).
P[ob] > 0.75 (2542.374).
P[ob] > 0.9 (6593.036).
P[ob] > 0.95 (20534.604).
P[ob] > 0.975 (33471.276).
P[ob] > 0.99 (50301.308).
P[ob] > 0.995 (52321.808).


Relative Operating Characteristic for different event (probability) thresholds.
Real.Time.Verification.Ensembles at lead hour 6
Probability of Detection (true alarms)
Probability of False Detection (false alarms)
Random guess (no skill)
P[ob] > 0.5 (1358.061).
P[ob] > 0.75 (2542.374).
P[ob] > 0.9 (6593.036).
P[ob] > 0.95 (20534.604).
P[ob] > 0.975 (33471.276).
P[ob] > 0.99 (50301.308).
P[ob] > 0.995 (52321.808).


Relative Operating Characteristic for different event (probability) thresholds.
Real.Time.Verification.Cond_obs at lead hour 6
Probability of Detection (true alarms)
Probability of False Detection (false alarms)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
0.45
0.50
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
0.95
1.00
Random guess (no skill)
P[ob] > 0.5 (1358.061).
P[ob] > 0.75 (2542.374).
P[ob] > 0.9 (6593.036).
P[ob] > 0.95 (20534.604).
P[ob] > 0.975 (33471.276).
P[ob] > 0.99 (50301.308).
P[ob] > 0.995 (52321.808).

Conclusions and next steps

Indicator approach shows promise

- \+ It explicitly minimizes the MSE of the observed probabilities, i.e. CRPS. (an important verification statistic).
- \+ Leads to significant gains in CRPS and other verification statistics.
- \+ Good for cases where parametric assumptions are unrealistic (e.g. precip.).
- \- High-dimensional technique, i.e. it follows the data, so it requires good hindcasting

Conclusions and next steps

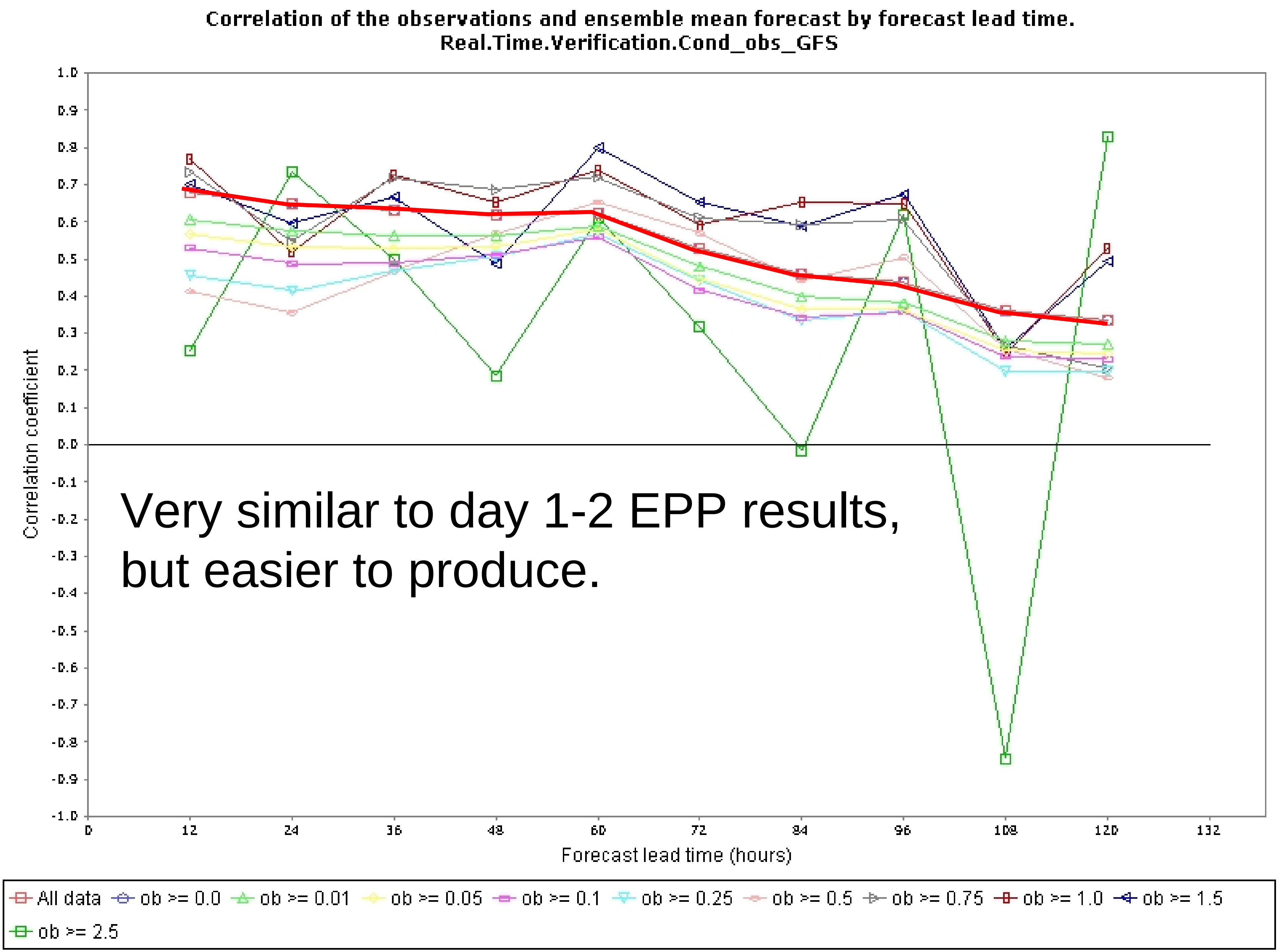
Need to test at an RFC

- Internal testing complete by FY09 Q2
- Need a candidate RFC to field-test
- Envisage testing similar to HMOS (ABRFC)

Work on visualizing analog forecasts

- Visualizing analogs remains important
- Particularly important for “unusual” cases
- Need a tool to identofy/visualize analogs

Additional slides


Correlation of the observations and ensemble mean forecast by forecast lead time.
Real.Time.Verification.Cond_obs_GFS
Correlation coefficient
Forecast lead time (hours)
All data
ob >= 0.0
ob >= 0.01
ob >= 0.05
ob >= 0.1
ob >= 0.25
ob >= 0.5
ob >= 0.75
ob >= 1.0
ob >= 1.5
ob >= 2.5
Very similar to day 1-2 EPP results, but easier to produce.