MARFC EVS Verification Case Study

Andrew Philpott, Bill Marosi (MARFC)

Julie Demargne, James Brown (OHD)

NWS Verification Team Meeting
11/10/2008

Ensemble Streamflow Prediction

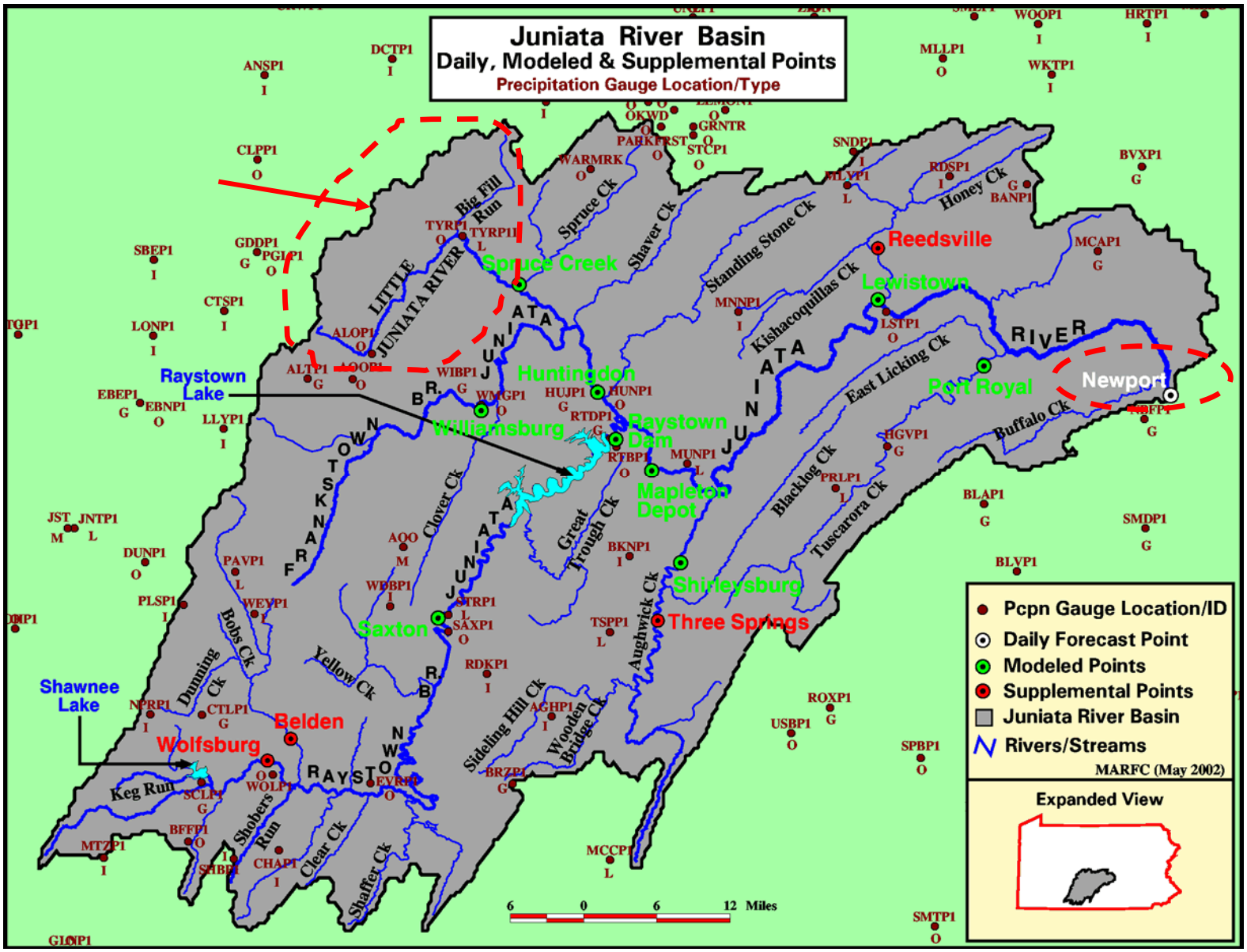
- A 6-hourly probabilistic streamflow forecast is generated each morning out to lead time 7 days, for all 10 modeled points in the Juniata River basin.
- The 50 streamflow traces are generated by 50 PQPF traces and 50 temperature forecast traces, starting from one carryover state, and using the Continuous API rainfall/runoff model with all operational mods.

ESP Forcings: Ensemble Pre-Processor (EPP2)

- The first 48 hours of 6-hourly PQPF traces are generated from the HAS deterministic QPF using statistical methods
- Days 3-7 PQPF are generated from a smoothed 50 year climatology.
- 5 days of temperature traces are generated from deterministic forecast, with days 6-7 from smoothed climatology

This Case Study

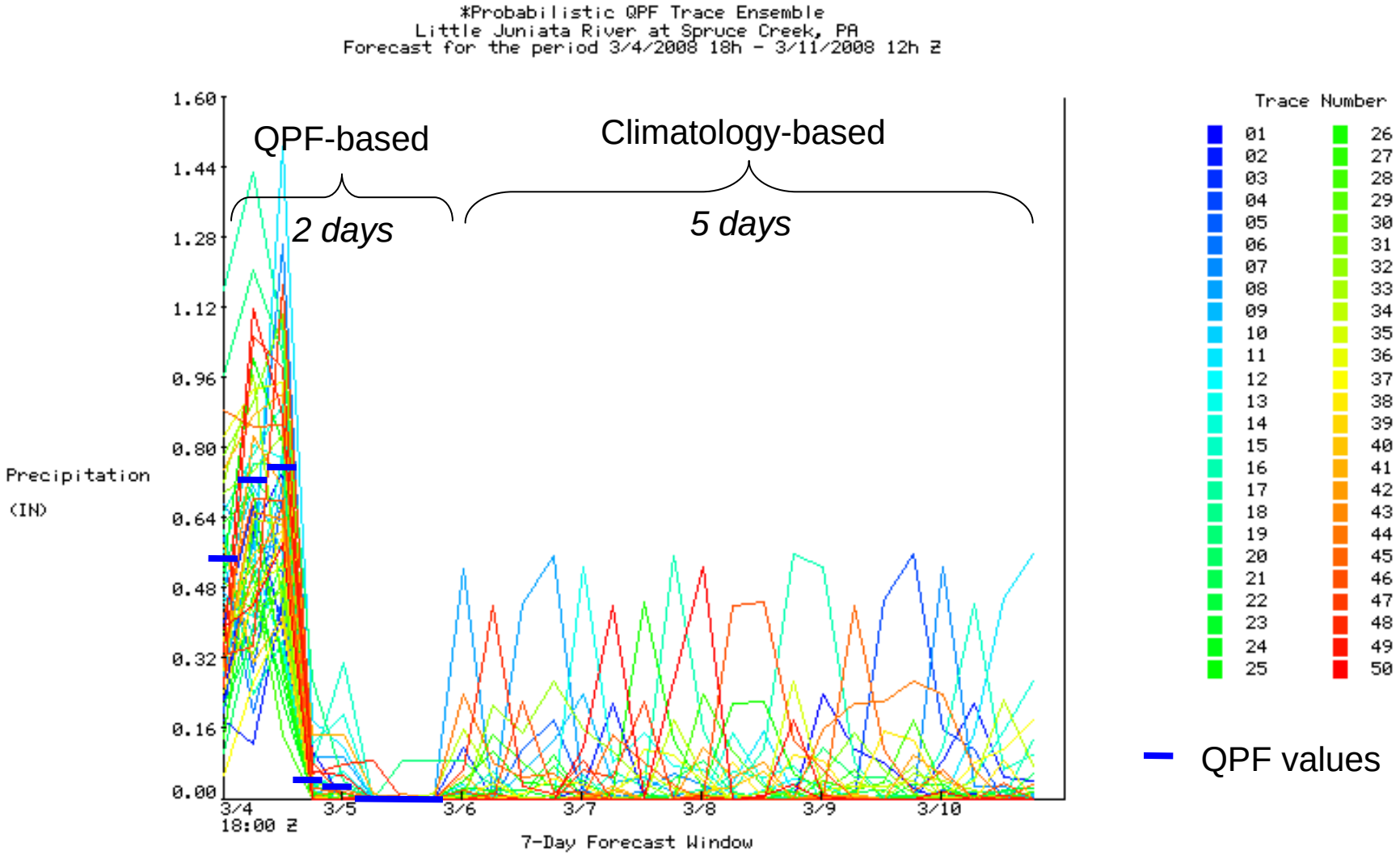
- We compared ESP forecasts to USGS streamflow data, and EPP2 precipitation and temperature traces to observed MAP, MAT for 2/2006 through 6/2008
- We will look at Spruce Creek (headwater point, peaks 6 hrs after rainfall) and Newport (downstream gage, crests about 2.5 days later)
- During this period, there was one near-flood event (March 2008) at Spruce Creek.



Map by David Solano, senior HAS

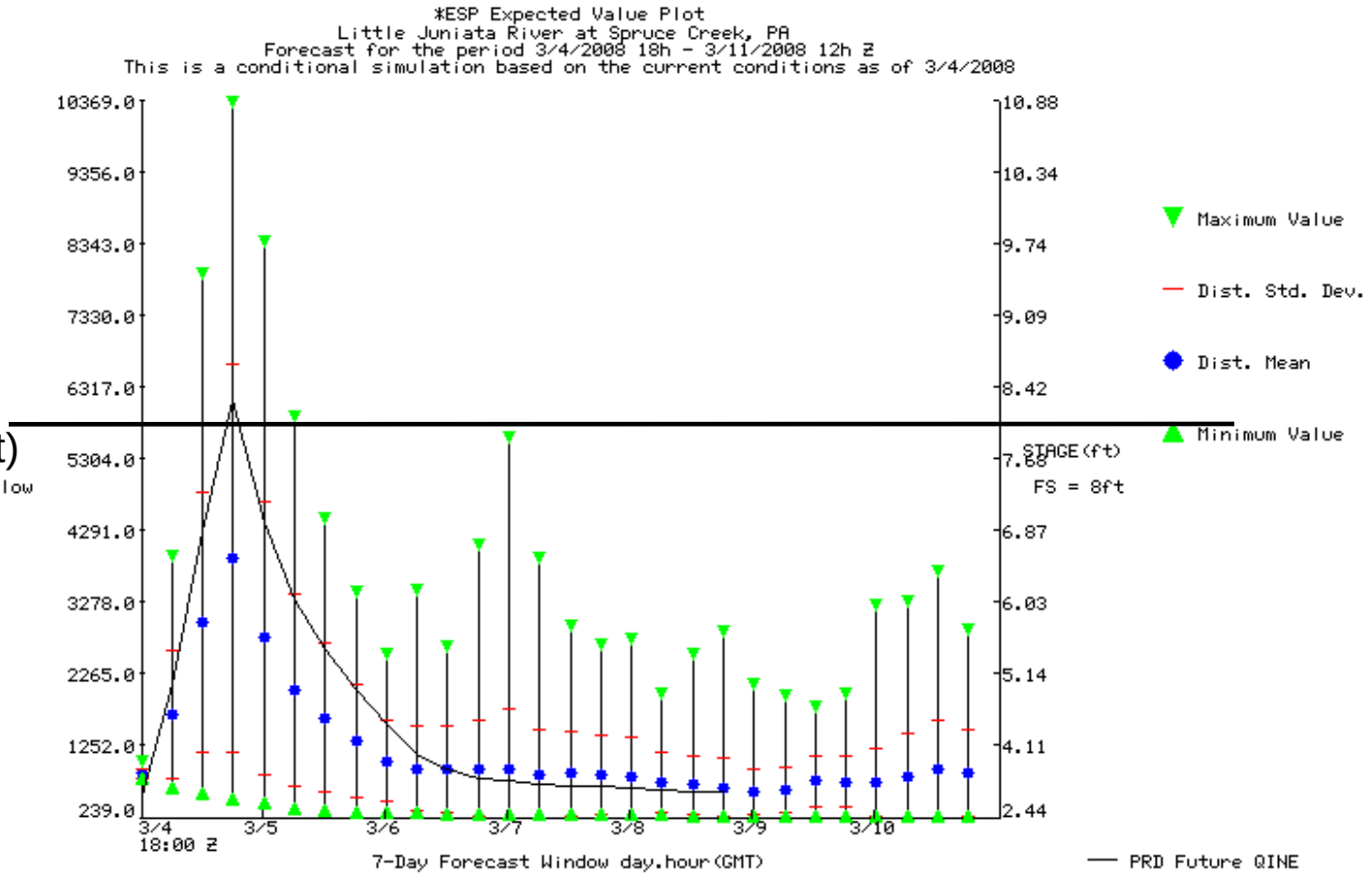
We issue 7-day ESP forecasts each day for 10 simulated points.

EPP2 generated precipitation forcings for the 3/4/2008 forecast, Spruce Creek



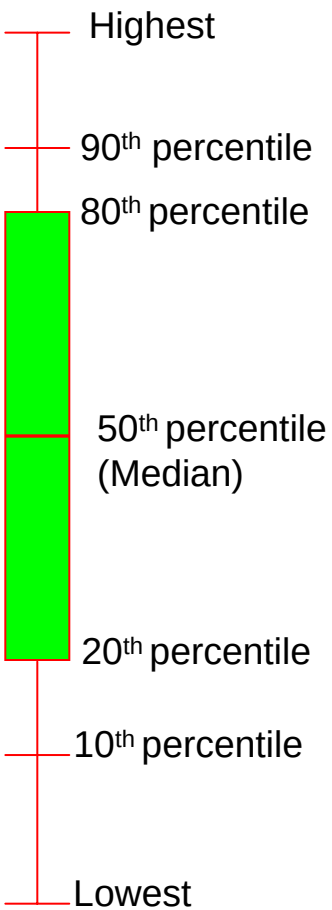
Deterministic QPF (for lead hours 6-48): 0.55, 0.72, 0.76, 0.04, 0.03, 0.00, 0.00, 0.00 in

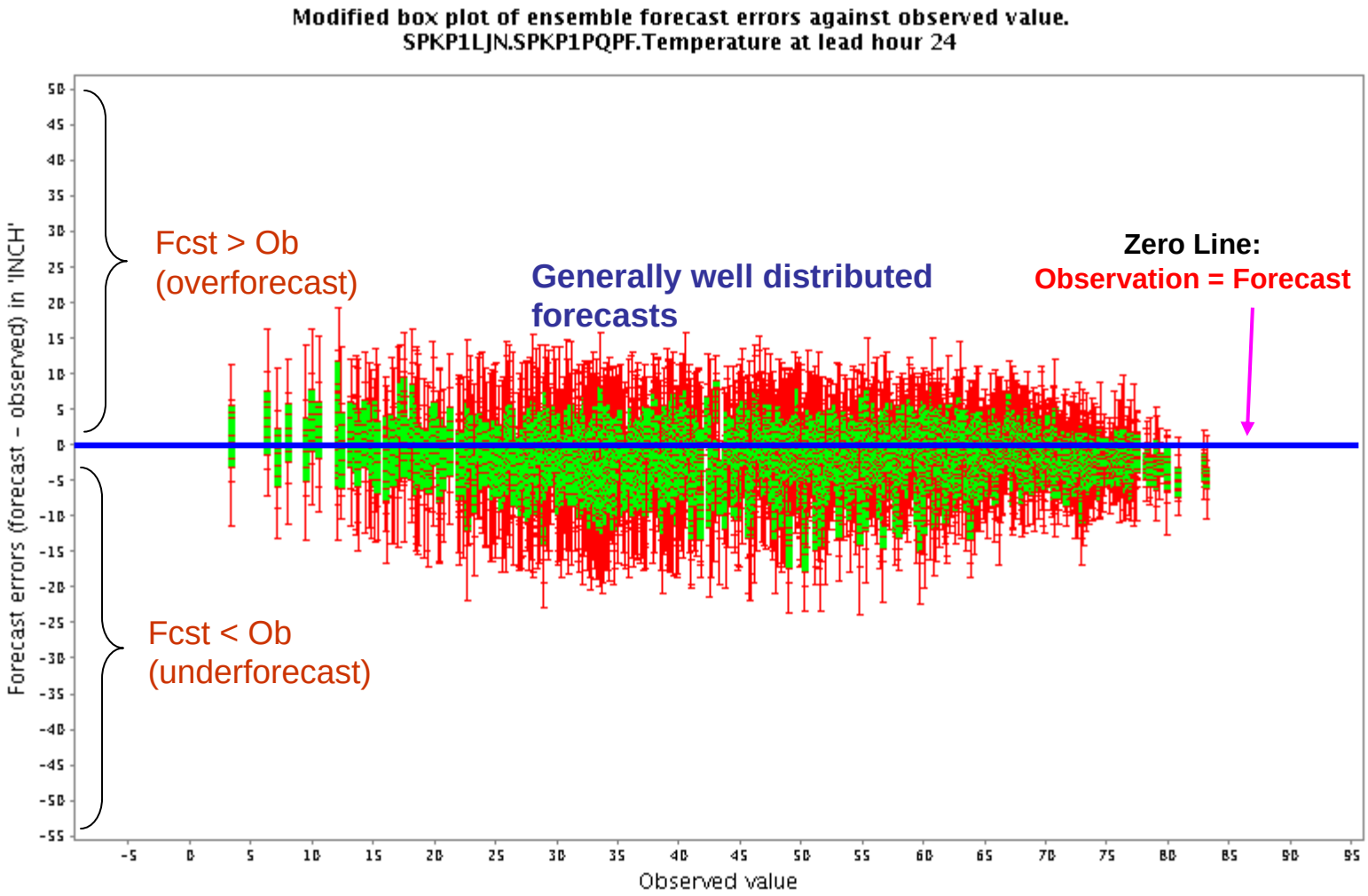
Expected Values for Streamflow 3/4/2008 Spruce Creek



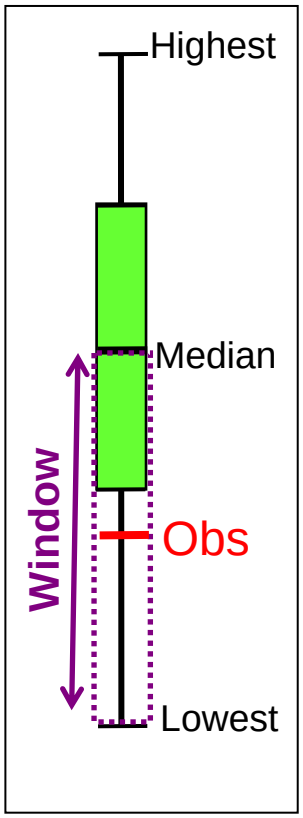
Flood Stage (8ft)

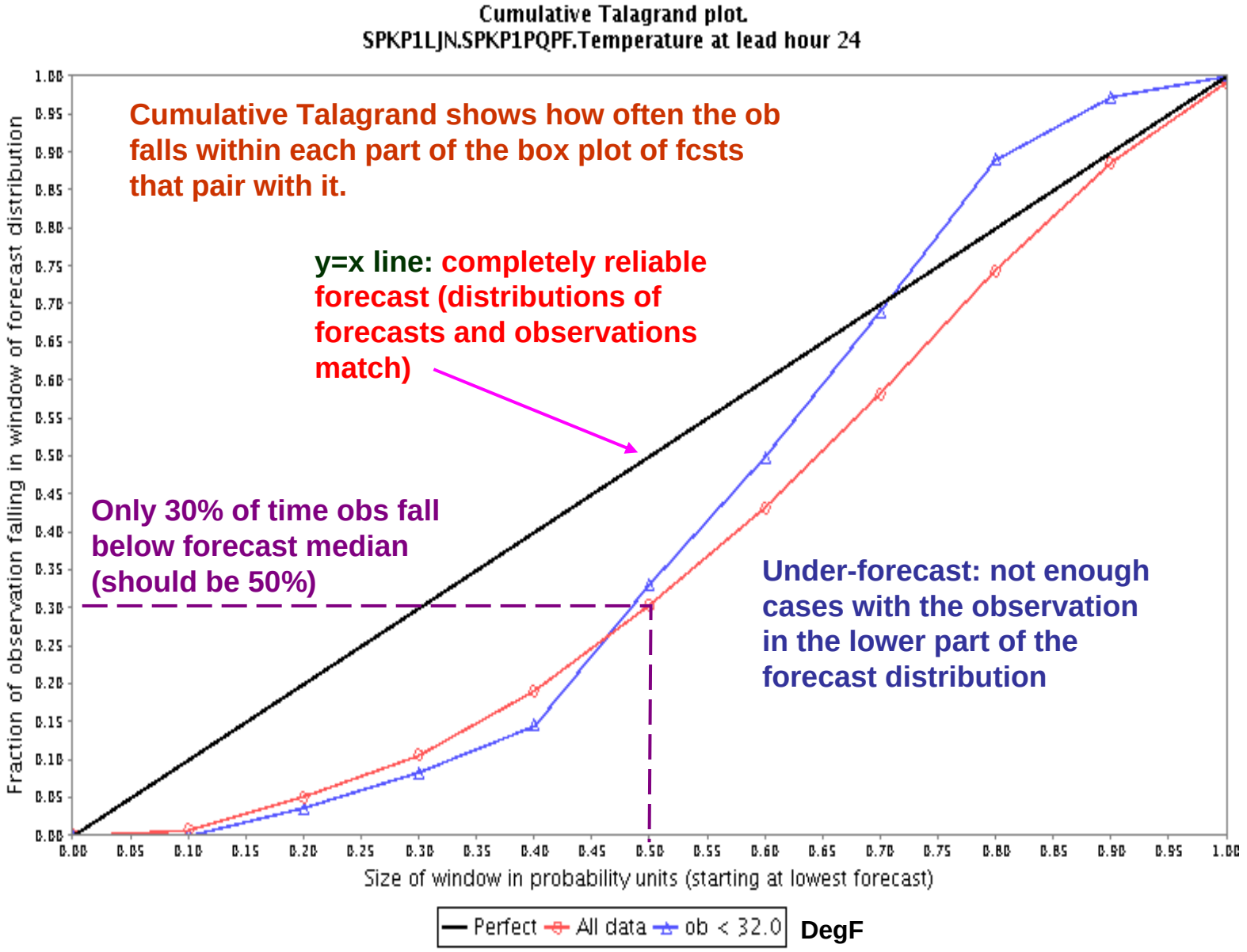
Daily average temperature: Plot of ensemble trace error versus observed value



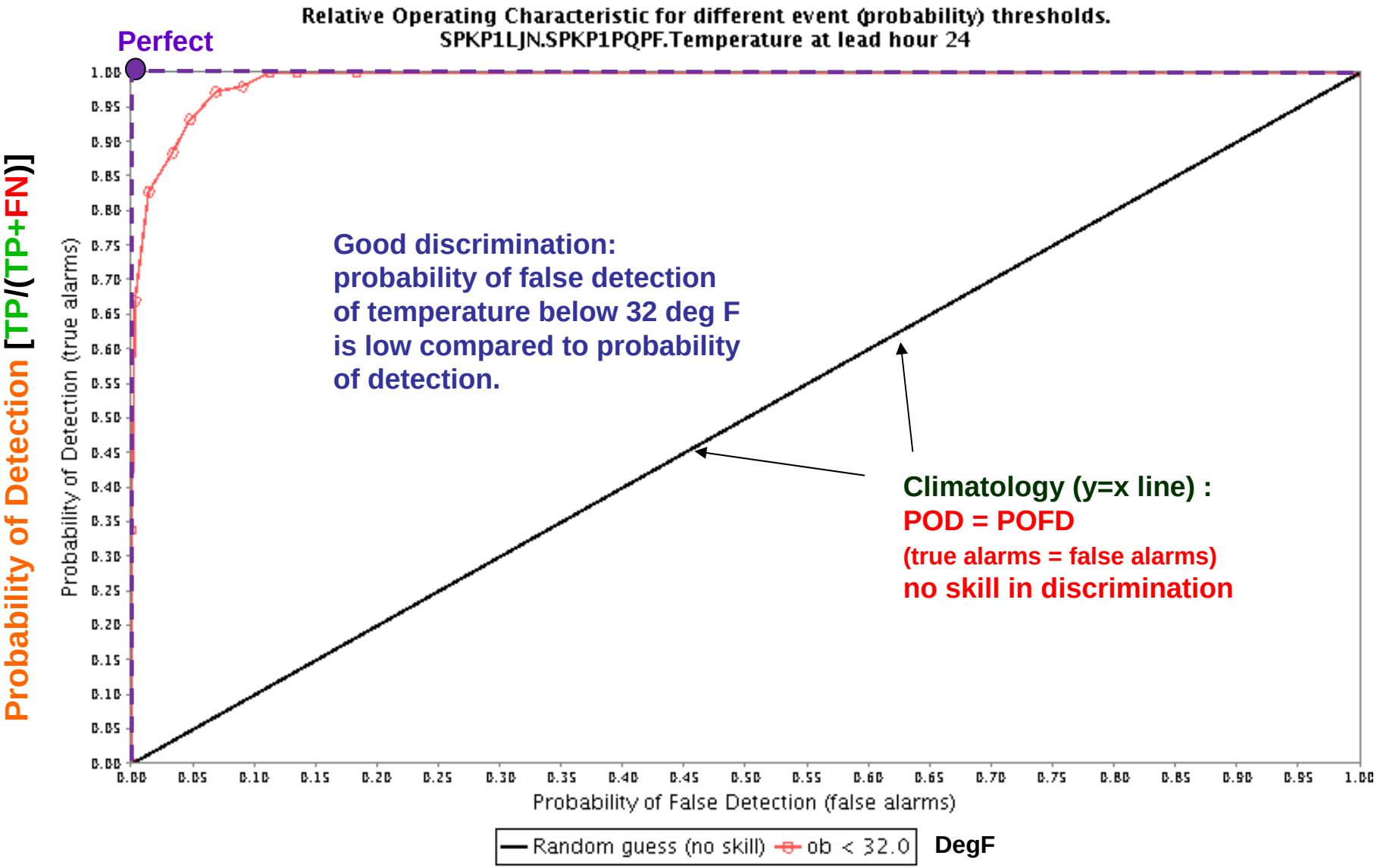


Reliability: Talagrand Plot





Discrimination: ROC Curve



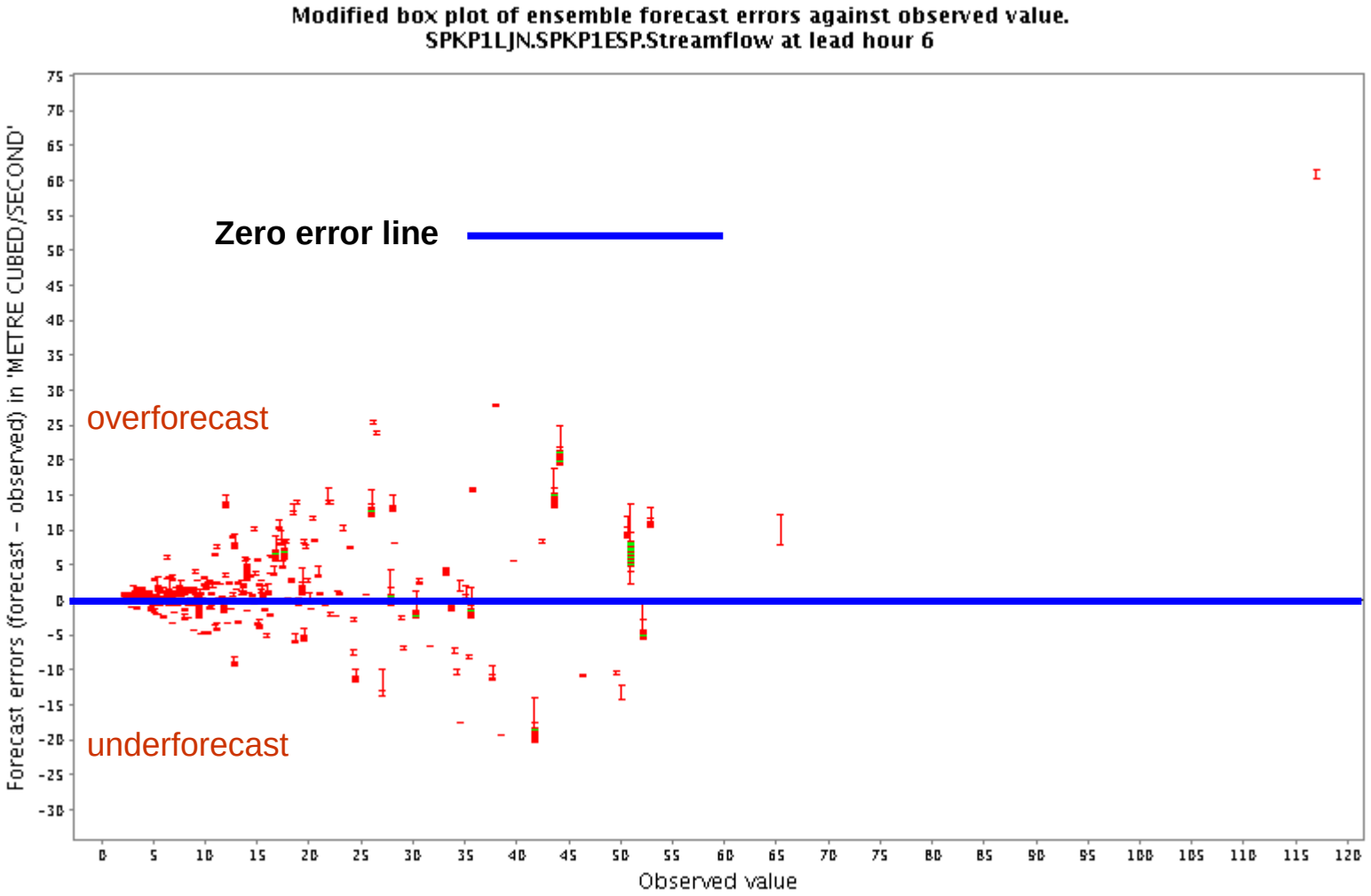
Probability of Detection [TP/(TP+FN)]

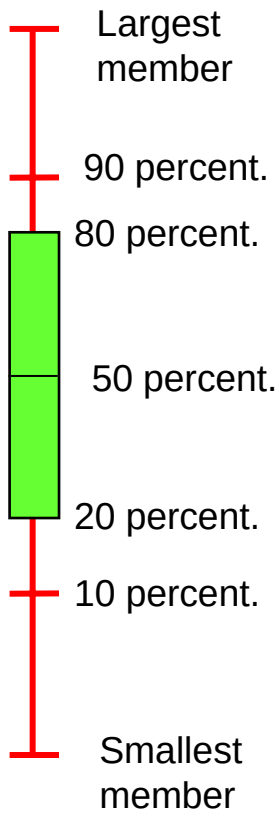
Good discrimination: probability of false detection of temperature below 32 deg F is low compared to probability of detection.

Climatology (y=x line) :
POD = POFD
(true alarms = false alarms)
no skill in discrimination

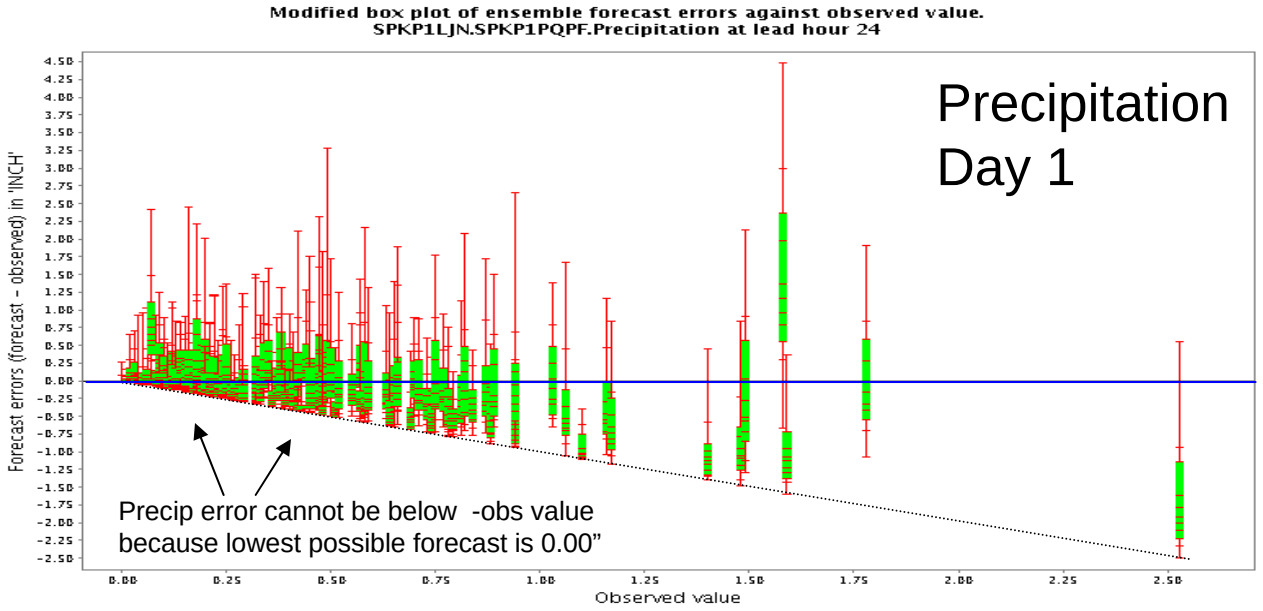
Probability of False Detection [FP/(FP+TN)]

Spruce Creek Streamflow
Box Plots: Error versus Obs Value

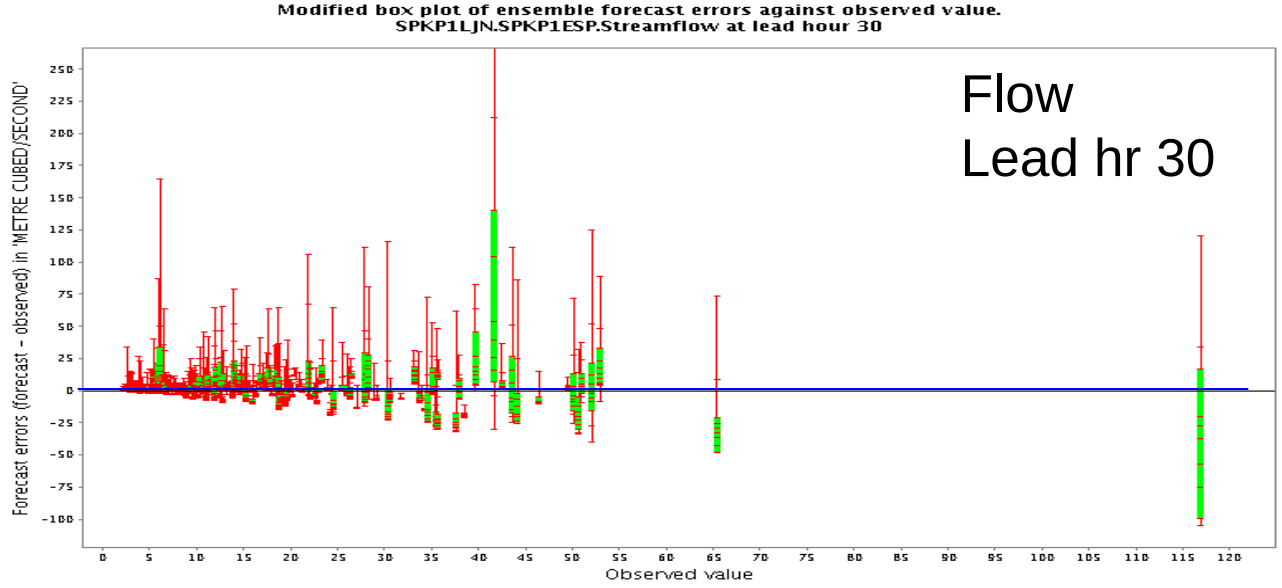




Spruce Creek
Day 1 Precipitation, Hour 30 Streamflow (QPF)

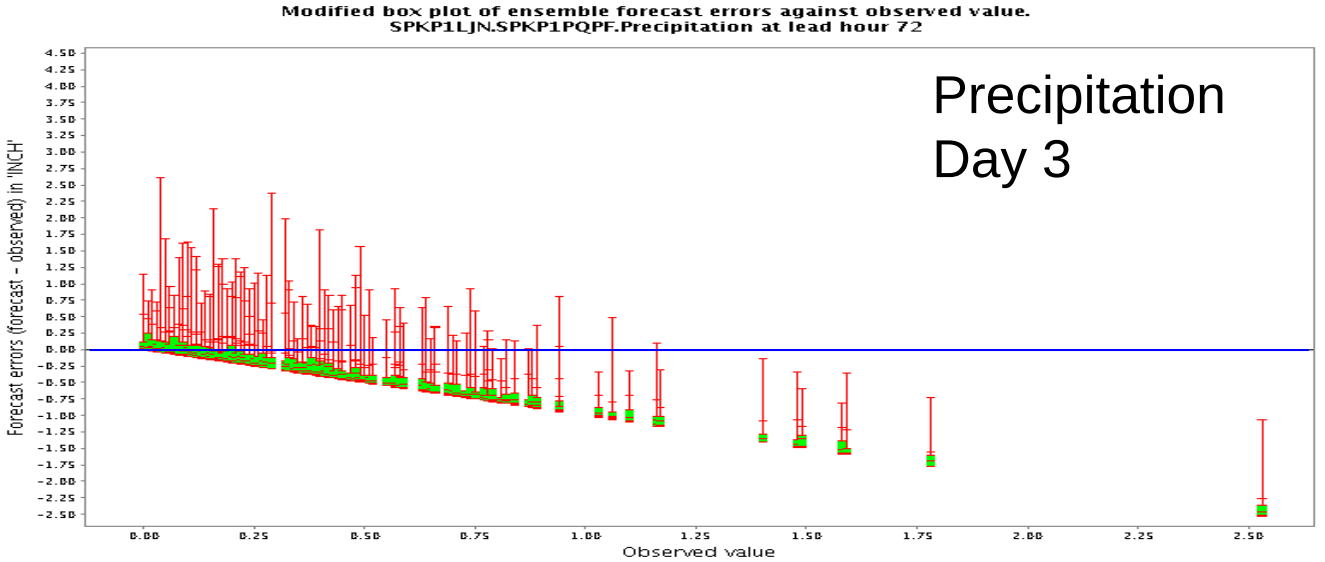


Zero error line

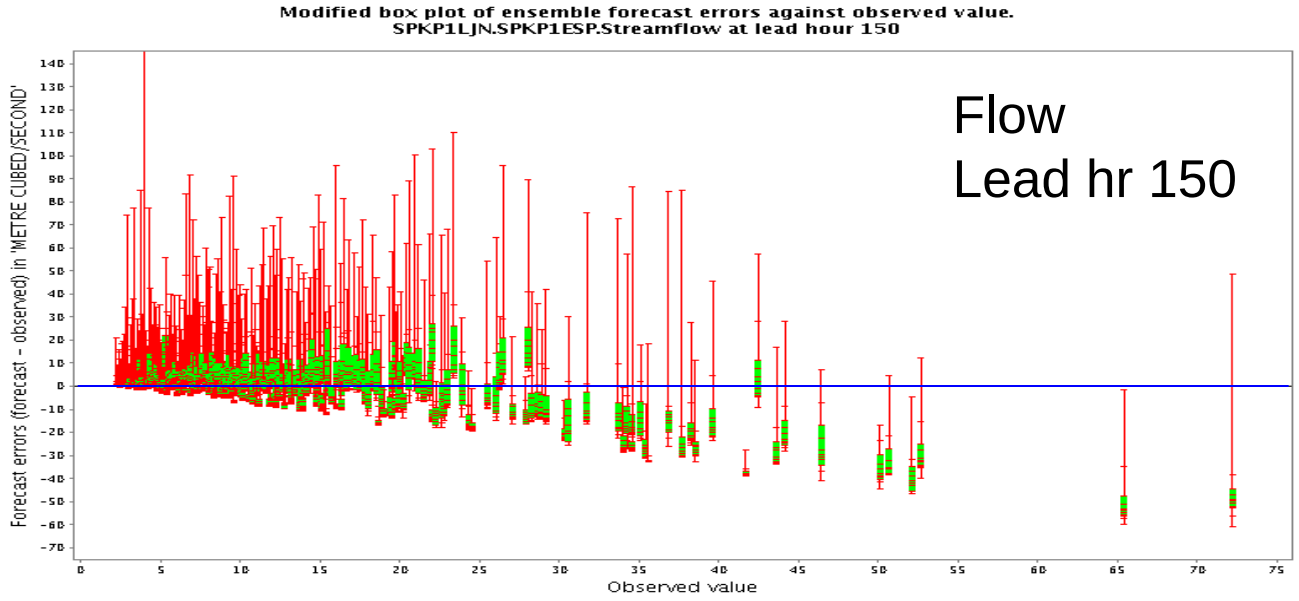


Spruce Creek

Day 3 Precipitation, Hour 150 Streamflow (Climo)

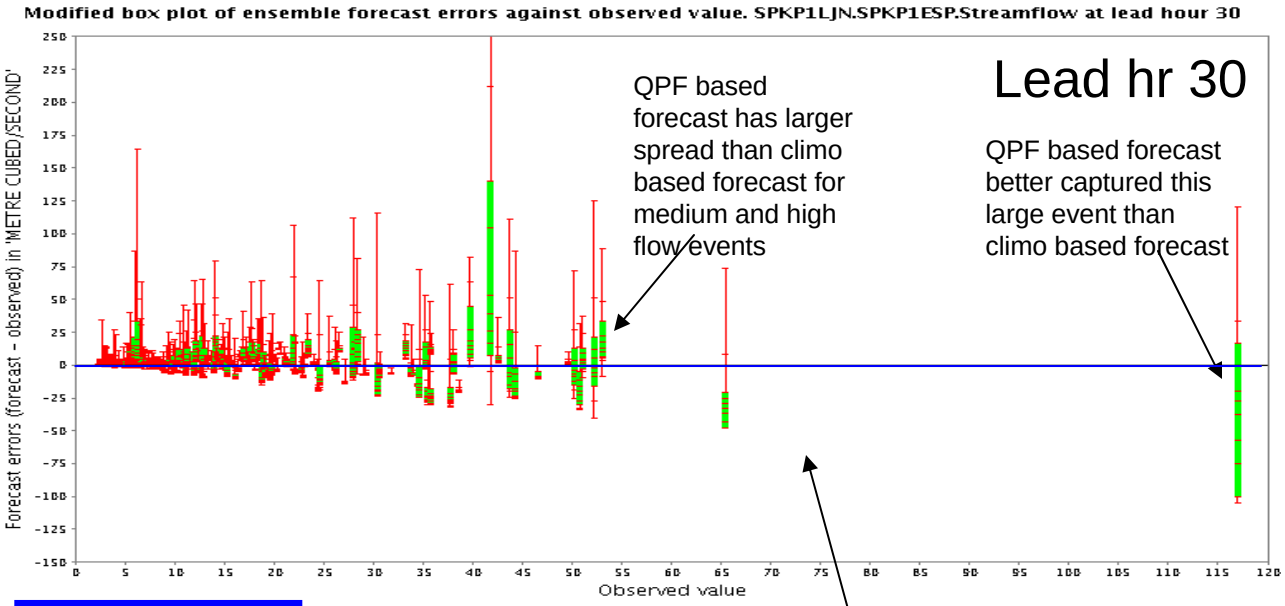


Zero error line



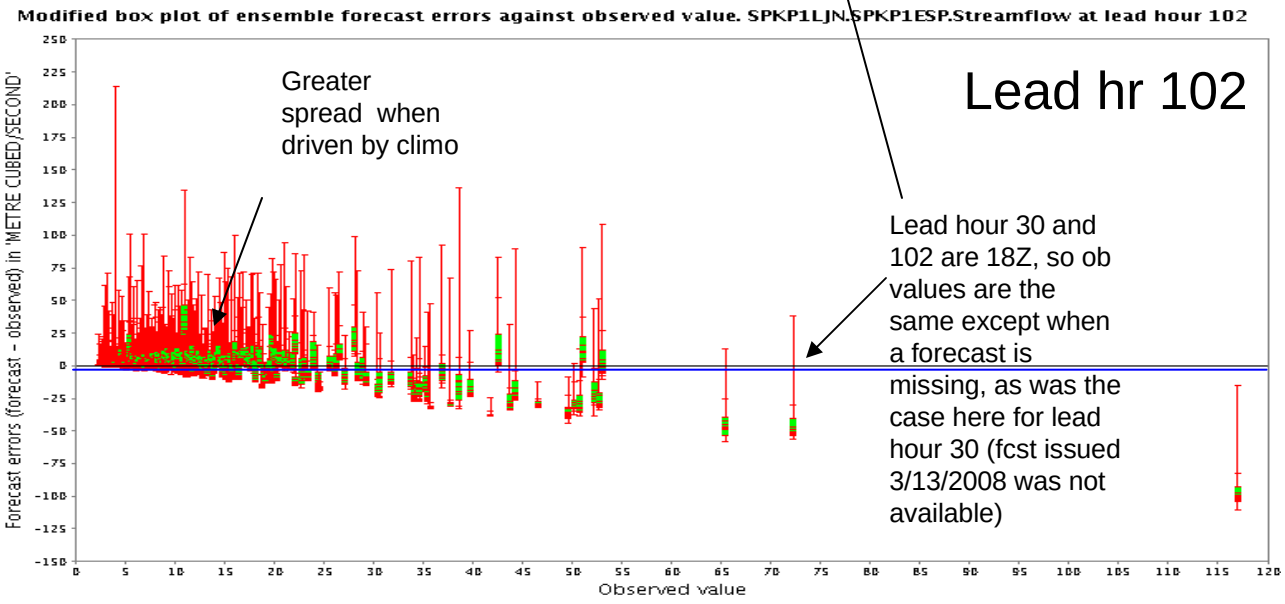
Spruce Creek Streamflow
QPF-based versus climatology-based

(note – vertical and horizontal scales identical in the two graphs)



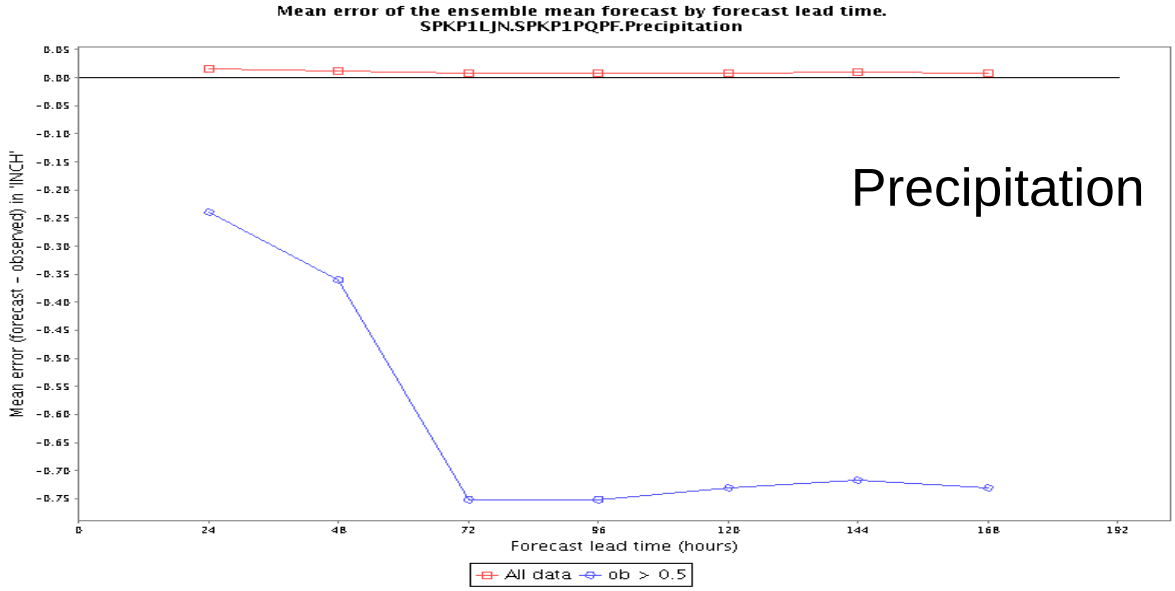
At Lead hour 30, QPF based PQPF is the primary forcing

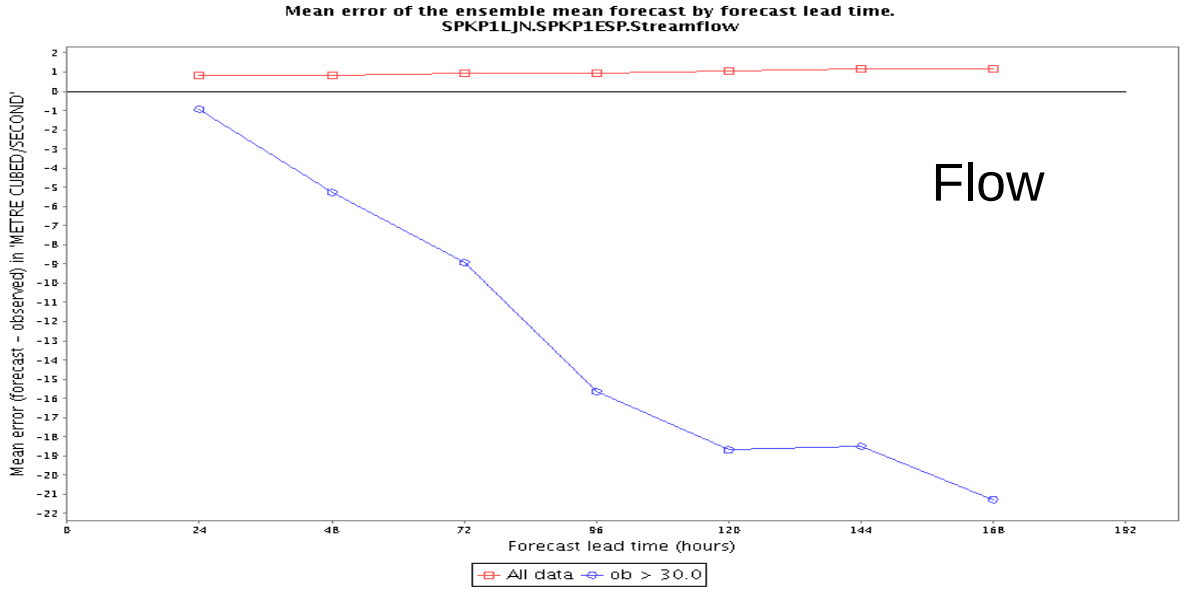
Zero error line



Runoff at hour 102 is from climatology based PQPF

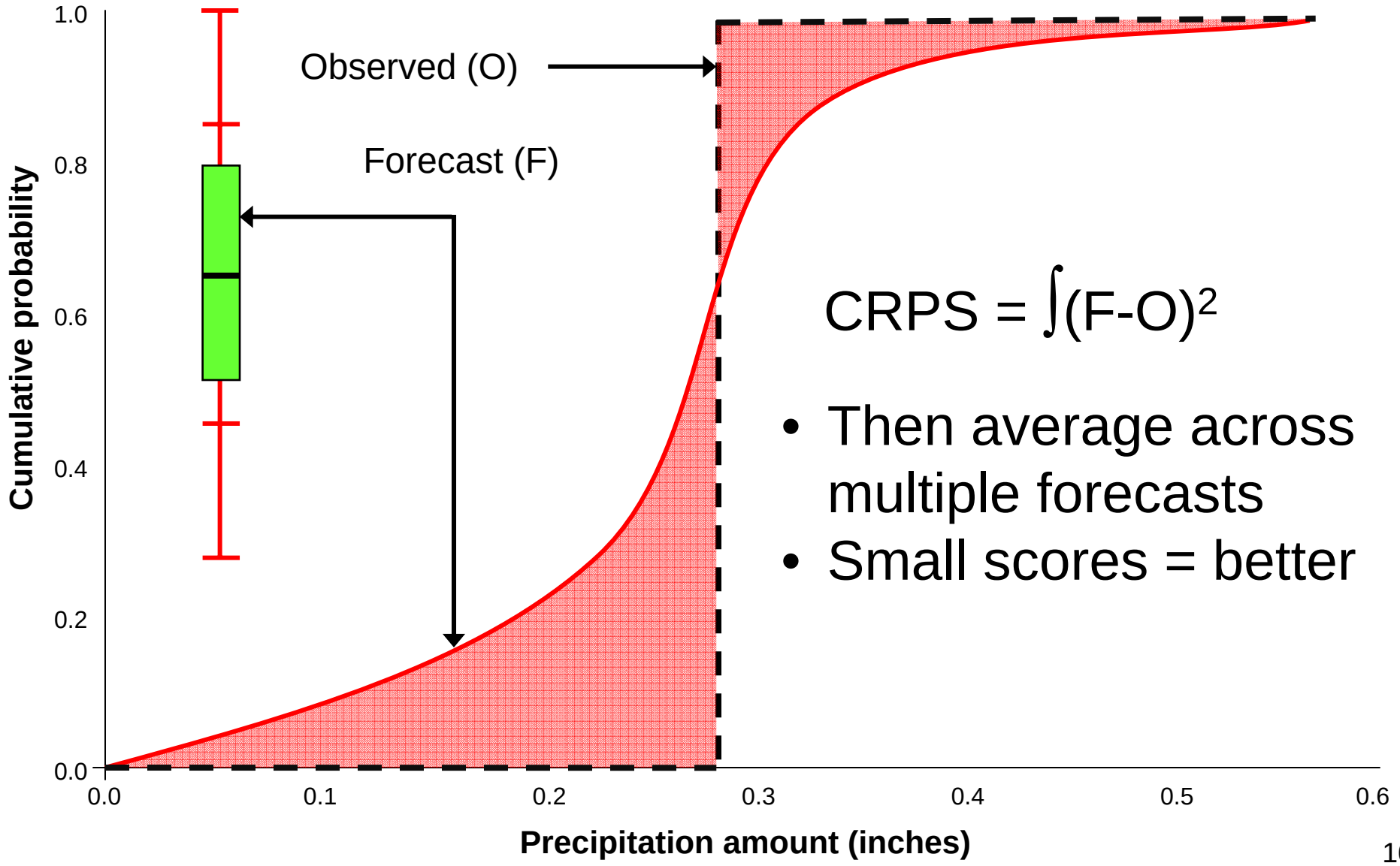
Mean Error (bias) in Ens Mean versus Observation





Precipitation and streamflow show little bias overall, but an underforecast bias (that worsens with lead time) for higher rain and higher water events.

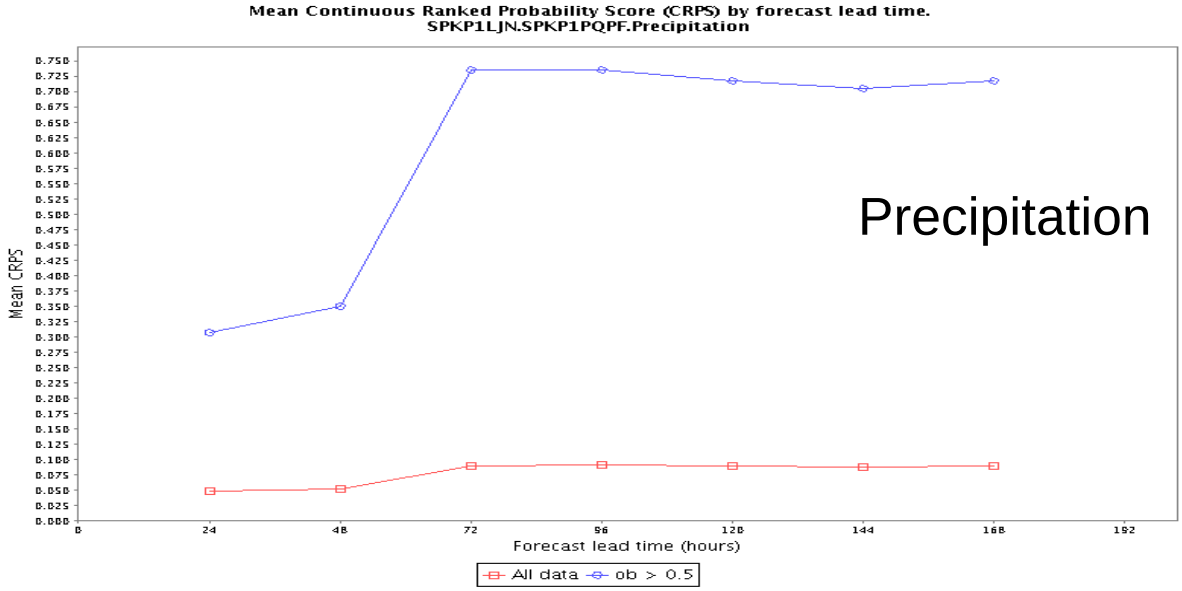
CRPS (slide by James Brown)



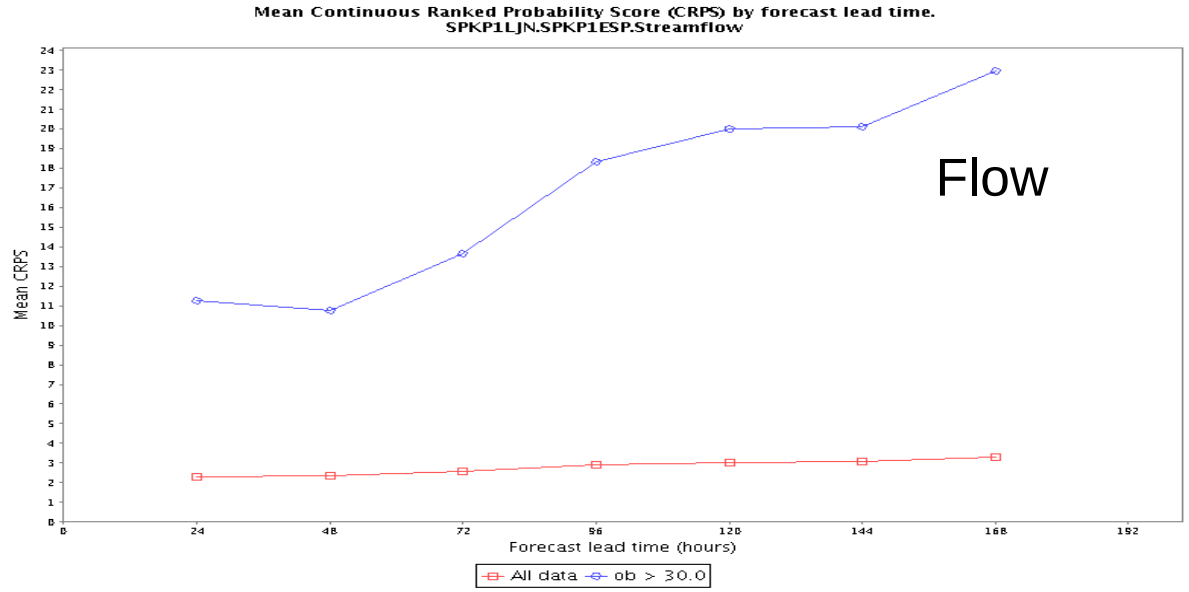
$$CRPS = \int (F-O)^2$$

- Then average across multiple forecasts
- Small scores = better

MCRPS for Spruce Creek



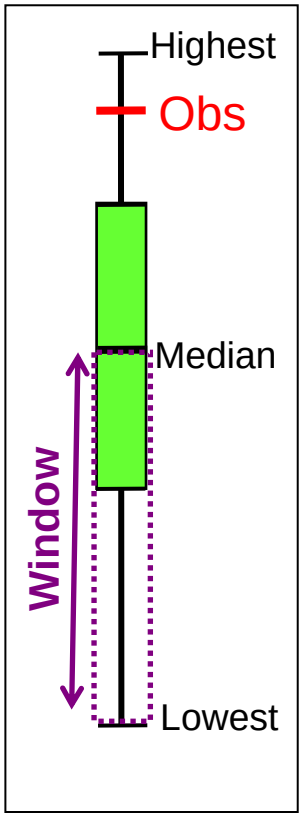
Precipitation

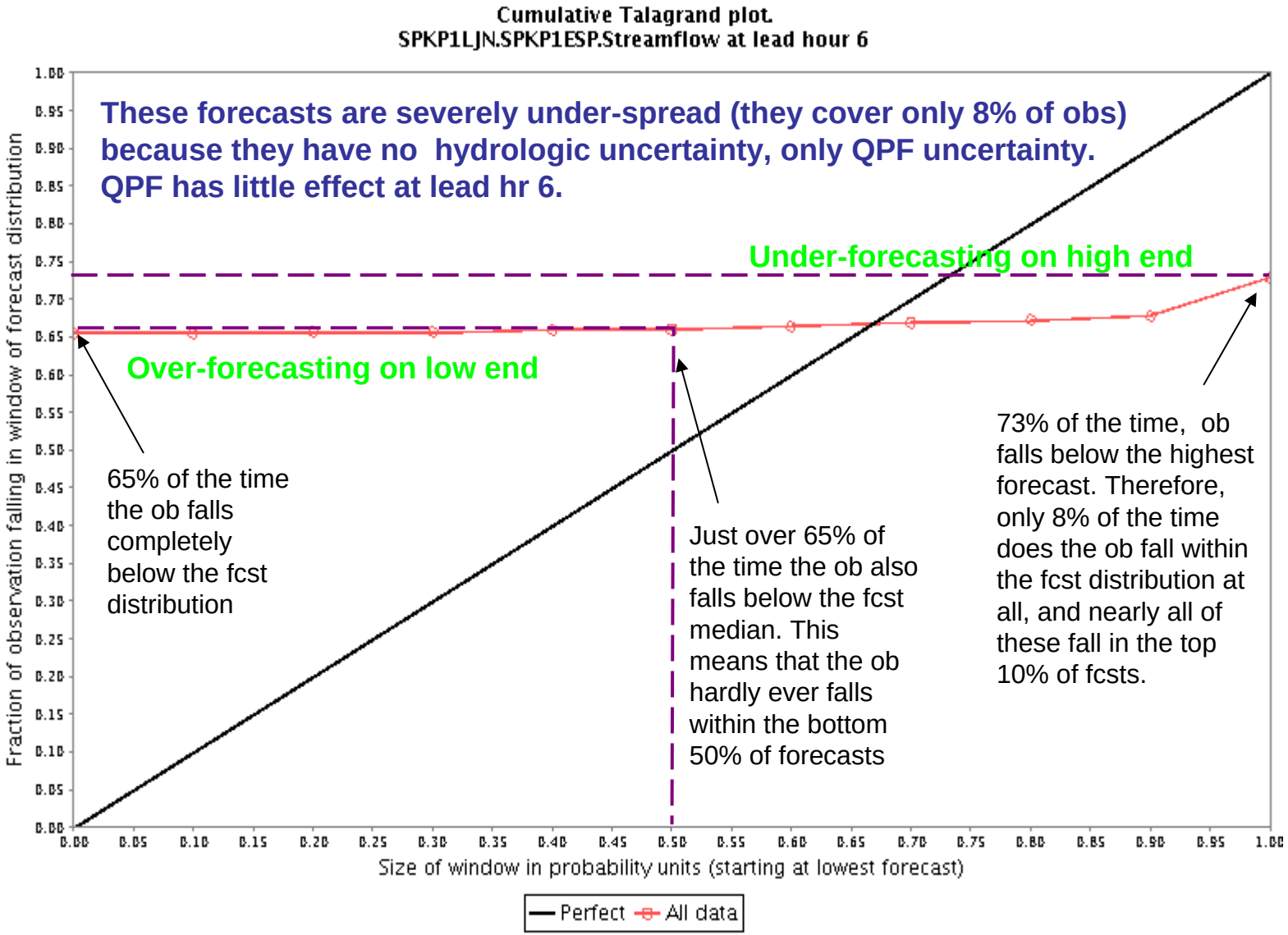


Flow

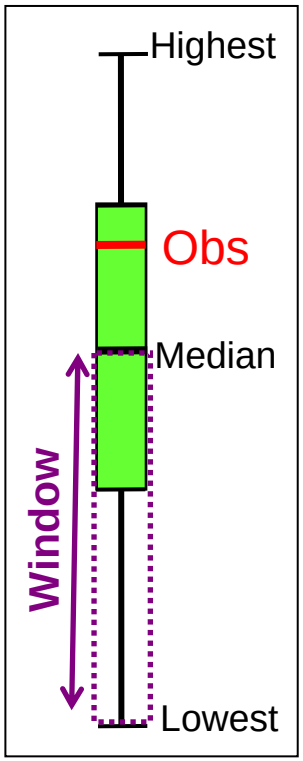
Deteriorating forecast accuracy with lead time, particularly for higher flow and higher precip events. Precip error increases abruptly in the transition from QPF based to climo based PQPF.

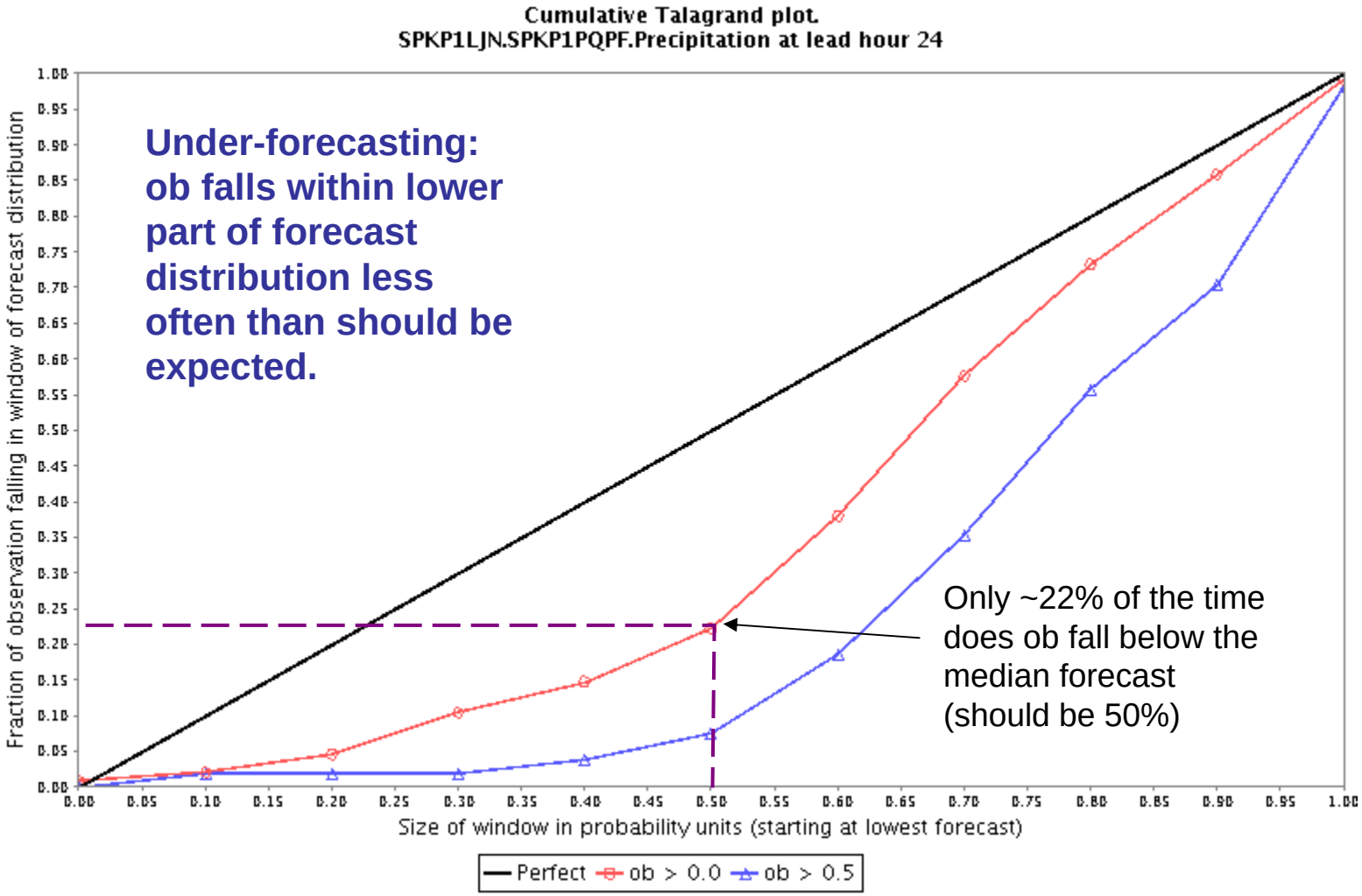
Reliability: Talagrand at 6hrs
Spruce Creek Streamflow



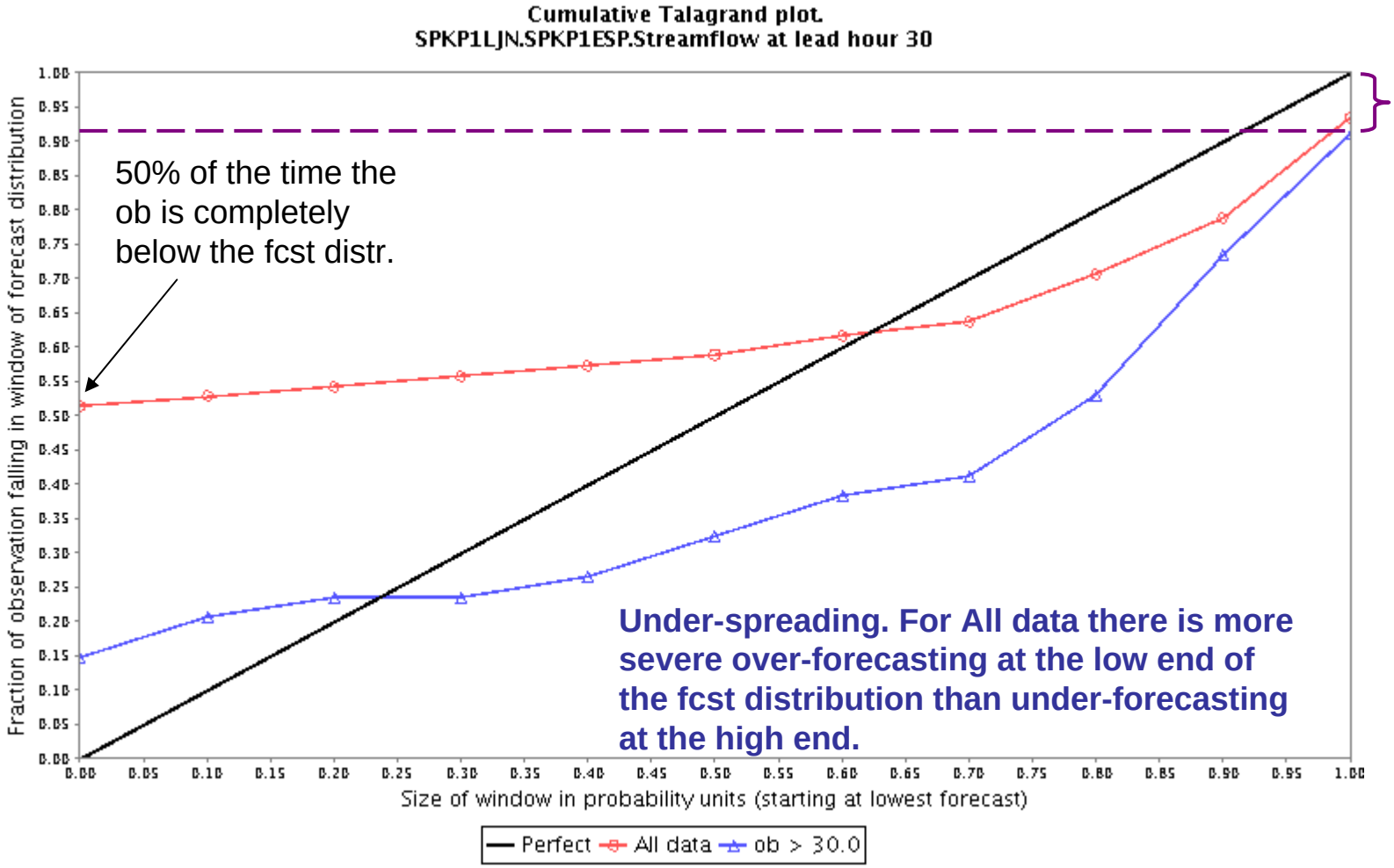


Reliability: Talagrand for Day 1 Total Precipitation





Reliability: Talagrand at 30 hrs
Spruce Creek Streamflow



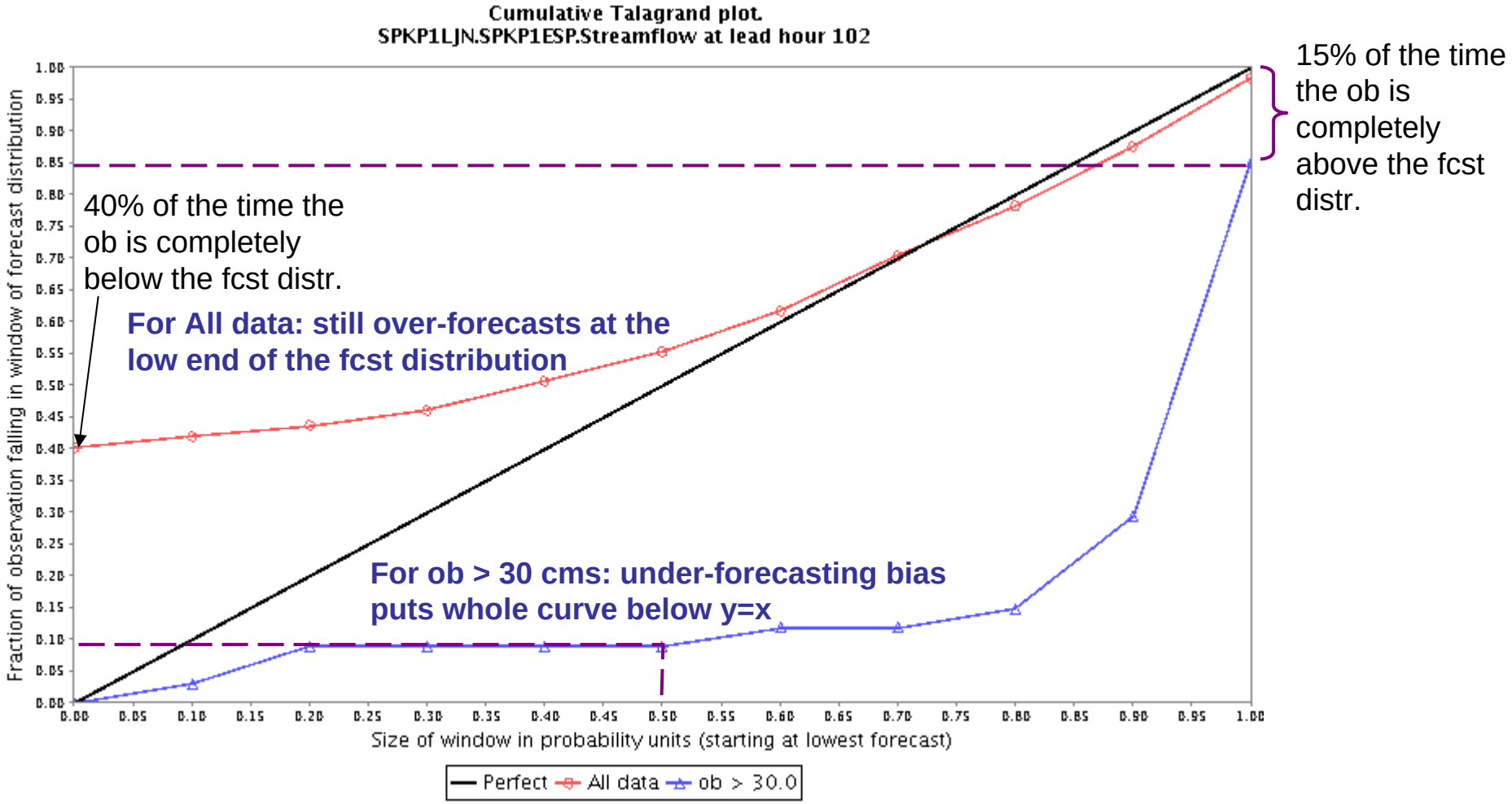
50% of the time the ob is completely below the fcst distr.

~8% of the time the ob is completely above the fcst distr.

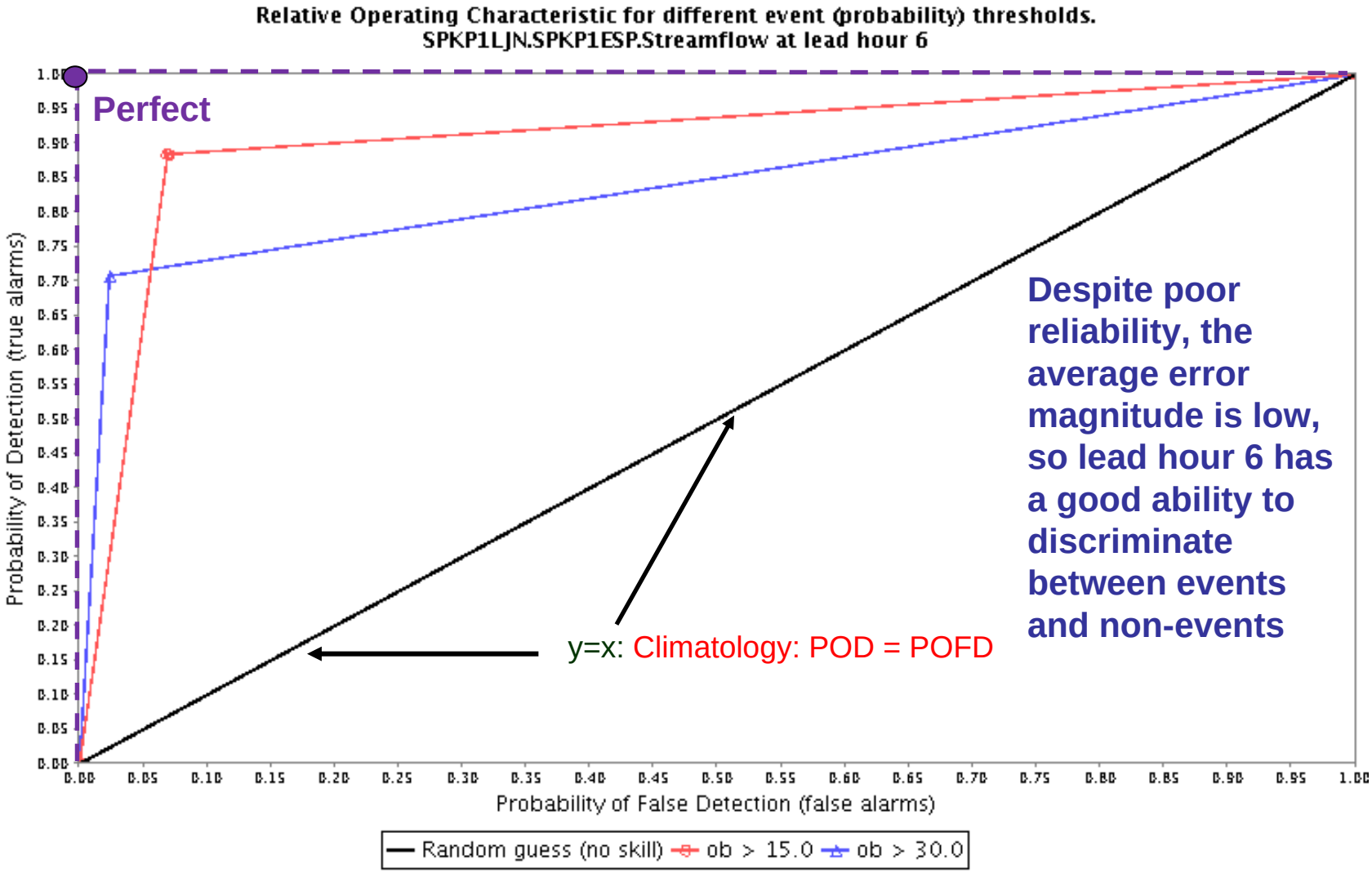
Under-spreading. For All data there is more severe over-forecasting at the low end of the fcst distribution than under-forecasting at the high end.

Reliability: Talagrand at 102 hrs
Spruce Creek Streamflow

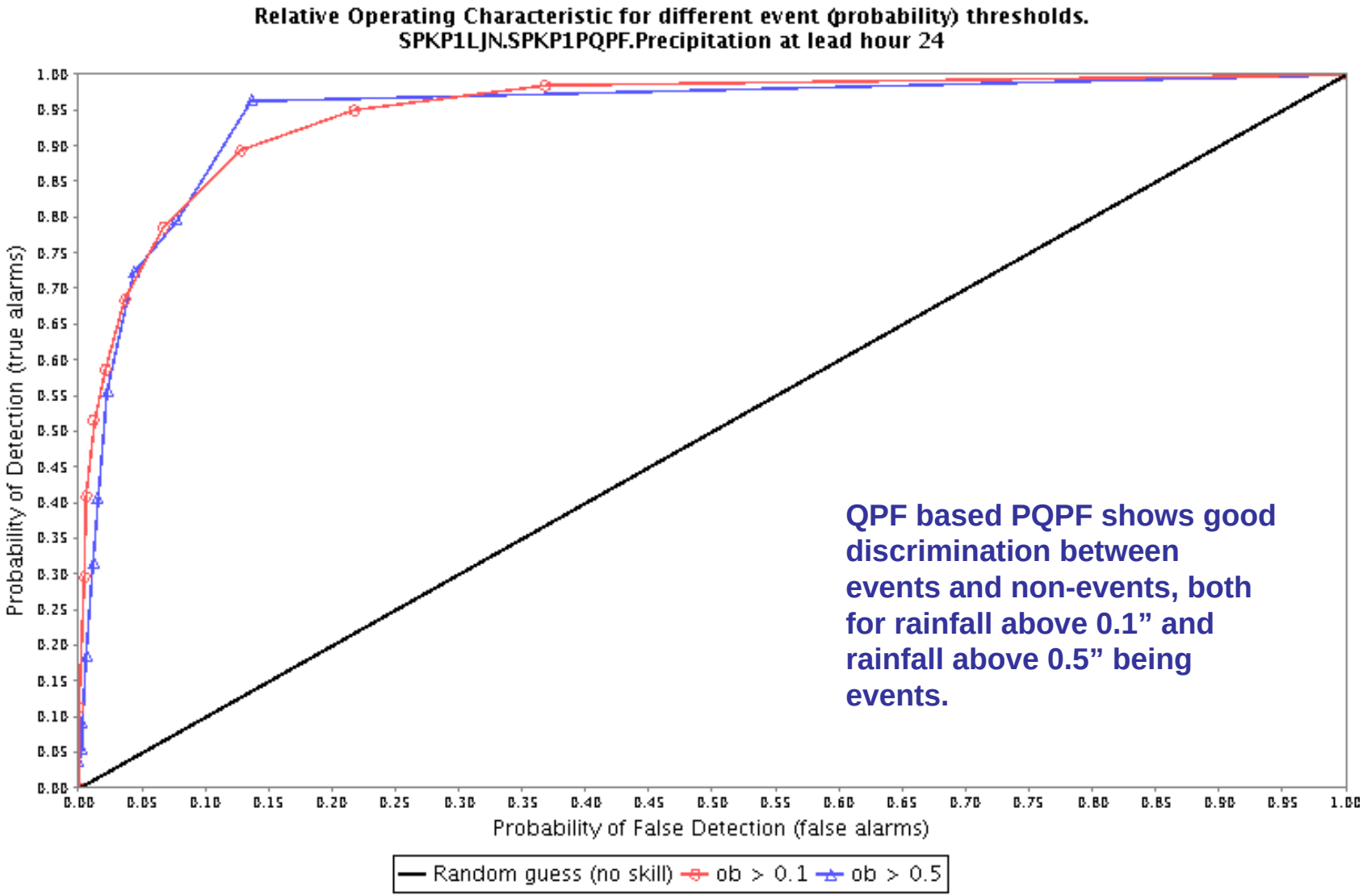
Larger spread of low flow forecasts using climatological forcing improves overall reliability, but high flows become less reliable because of an under-forecasting bias.



Discrimination: ROC at 6 hrs
Spruce Creek Streamflow

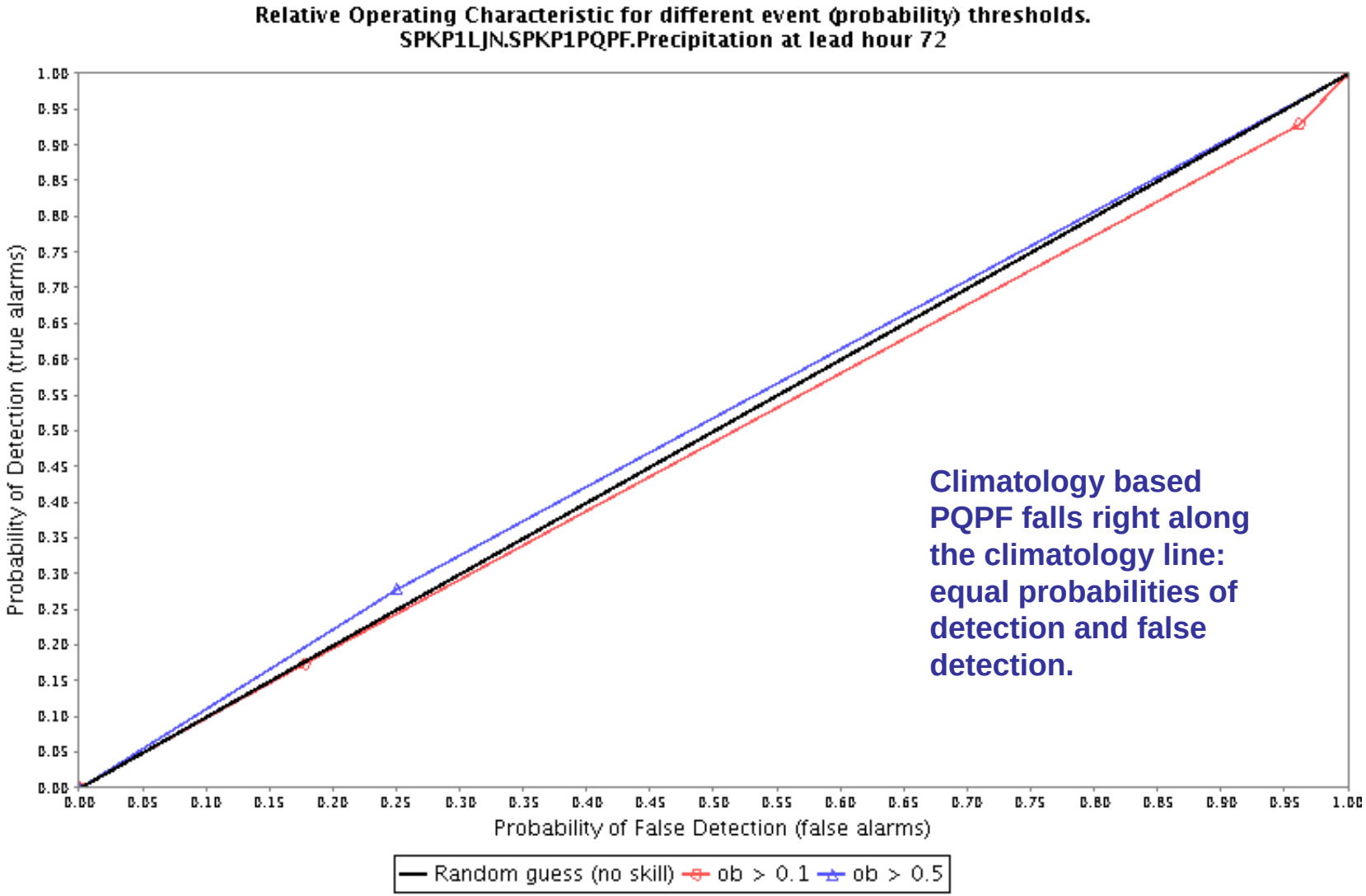


Discrimination: ROC for Day 1
Spruce Creek Precipitation



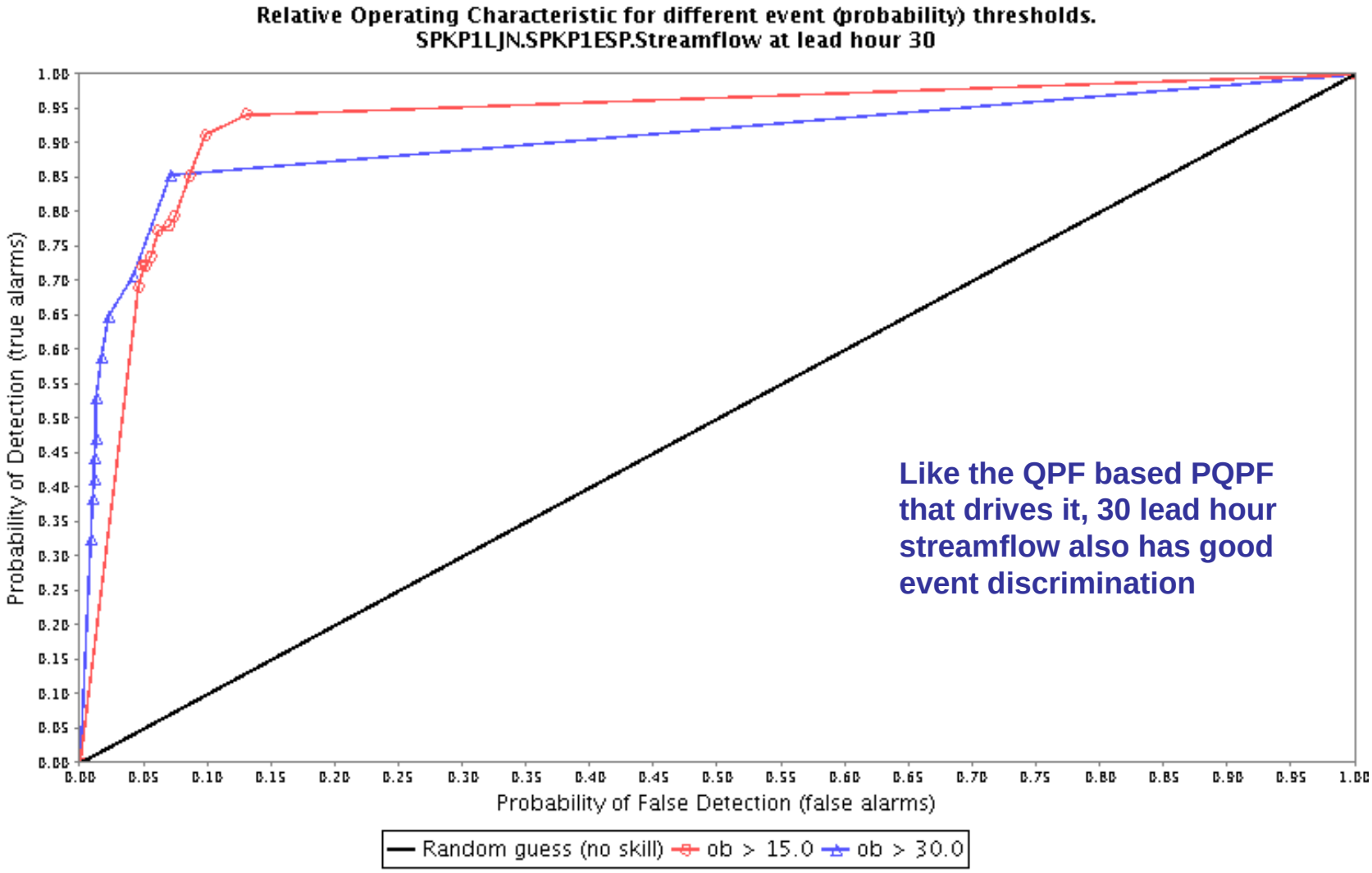
QPF based PQPF shows good discrimination between events and non-events, both for rainfall above 0.1” and rainfall above 0.5” being events.

Discrimination: ROC for Day 3
Spruce Creek Precipitation



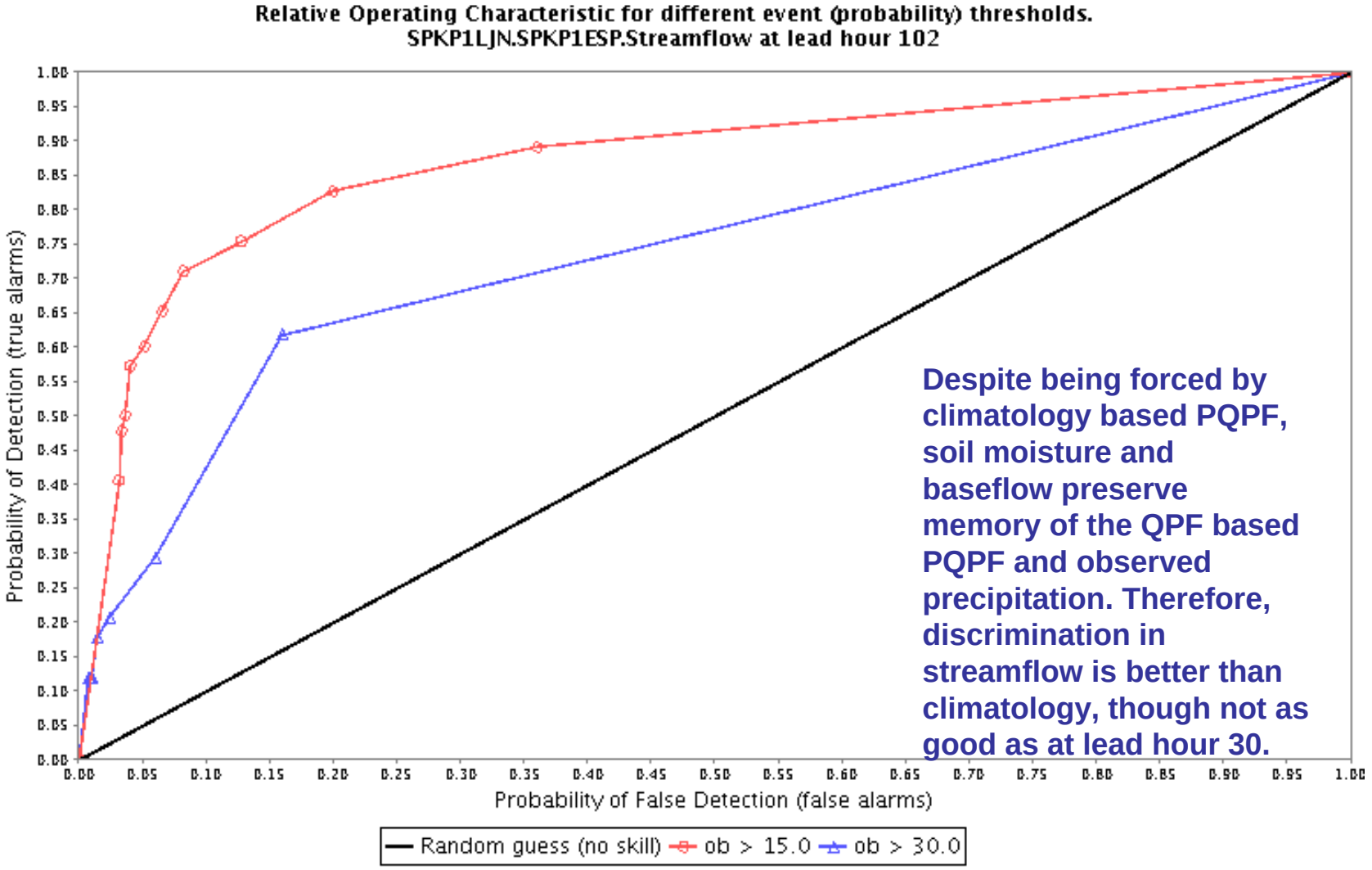
Climatology based PQPF falls right along the climatology line: equal probabilities of detection and false detection.

Discrimination: ROC at 30 hrs
Spruce Creek Streamflow



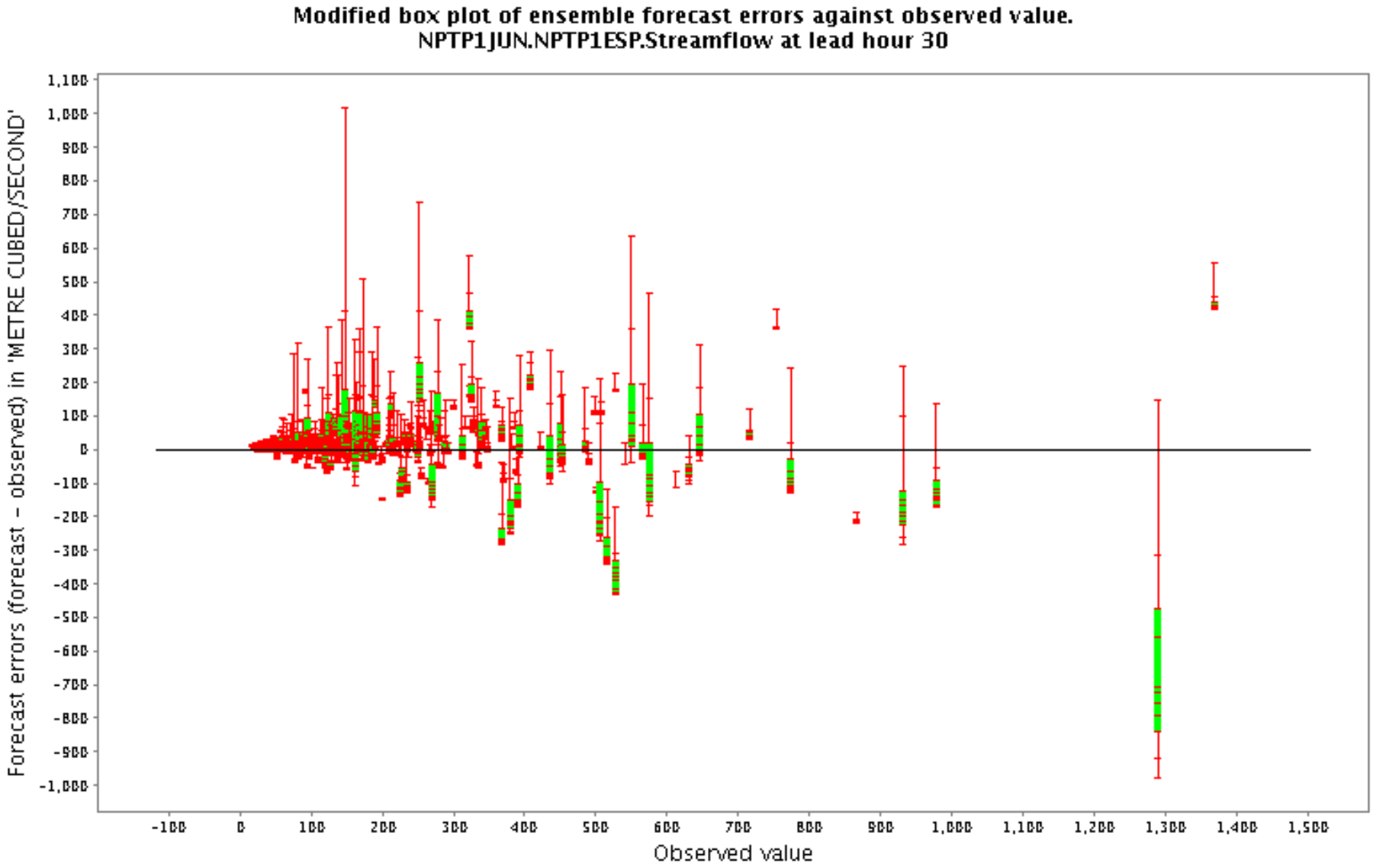
Like the QPF based PQPF that drives it, 30 lead hour streamflow also has good event discrimination

Discrimination: ROC at 102 hrs
Spruce Creek Streamflow

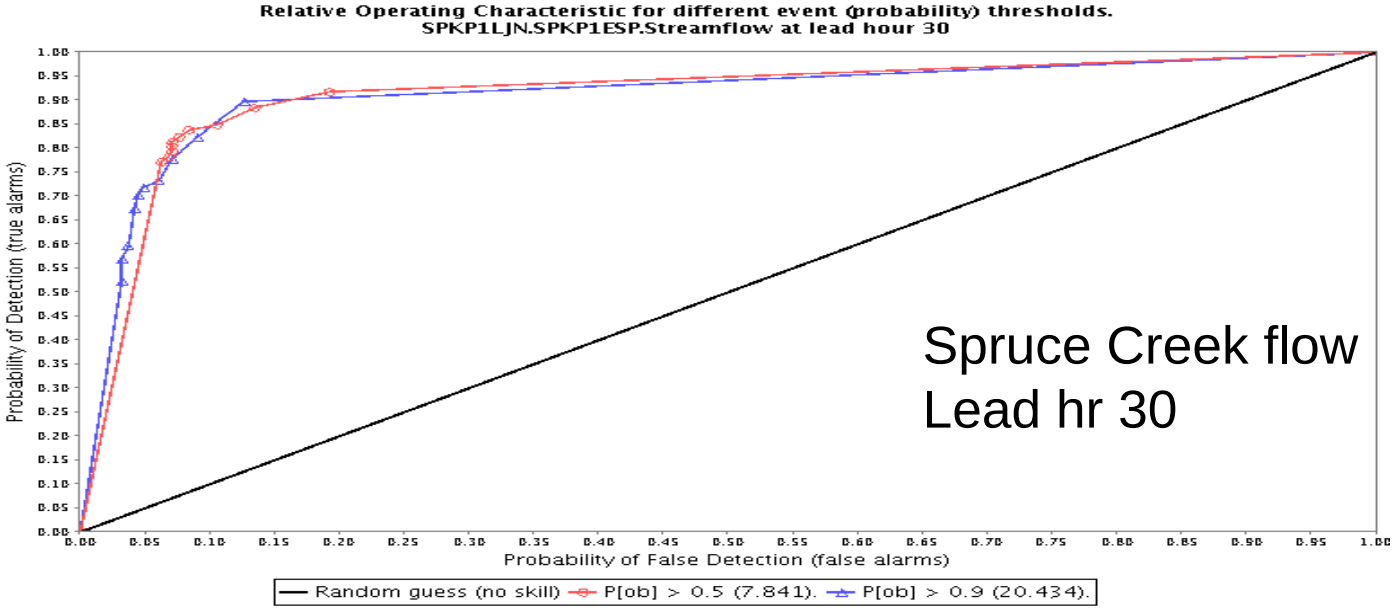


Despite being forced by climatology based PQPF, soil moisture and baseflow preserve memory of the QPF based PQPF and observed precipitation. Therefore, discrimination in streamflow is better than climatology, though not as good as at lead hour 30.

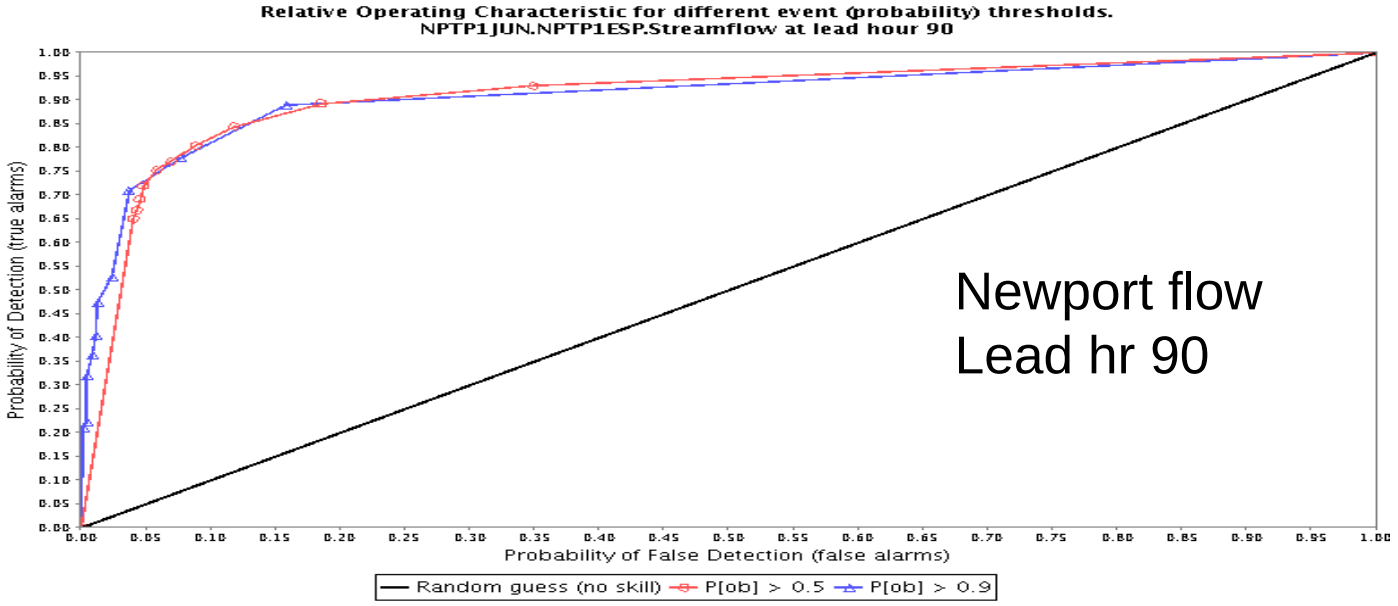
Newport Streamflow
Box Plots: Error versus Obs Value



ROC Curve: Newport at 90 hours compared to Spruce Creek at 30 hours --- quite similar



Spruce Creek flow
Lead hr 30



Newport flow
Lead hr 90

Main Conclusions

- The temperature ensembles have relatively high reliability and discrimination.

- HAS QPF based PQPF in the first 48 hours discriminates well between events and non-events. It tends to underforecast the high events, however.
- Smoothed climatology based PQPF cannot discriminate between events and non-events, and has a larger underforecasting bias for high events than the QPF based PQPF.

- Streamflow forecasts at early lead hours have very little spread because the ensembles do not incorporate hydrologic and initial state uncertainties, only uncertainties in the QPF and temperature forecast. They are unreliable, but generally low error magnitude yields good discrimination between events and non-events.

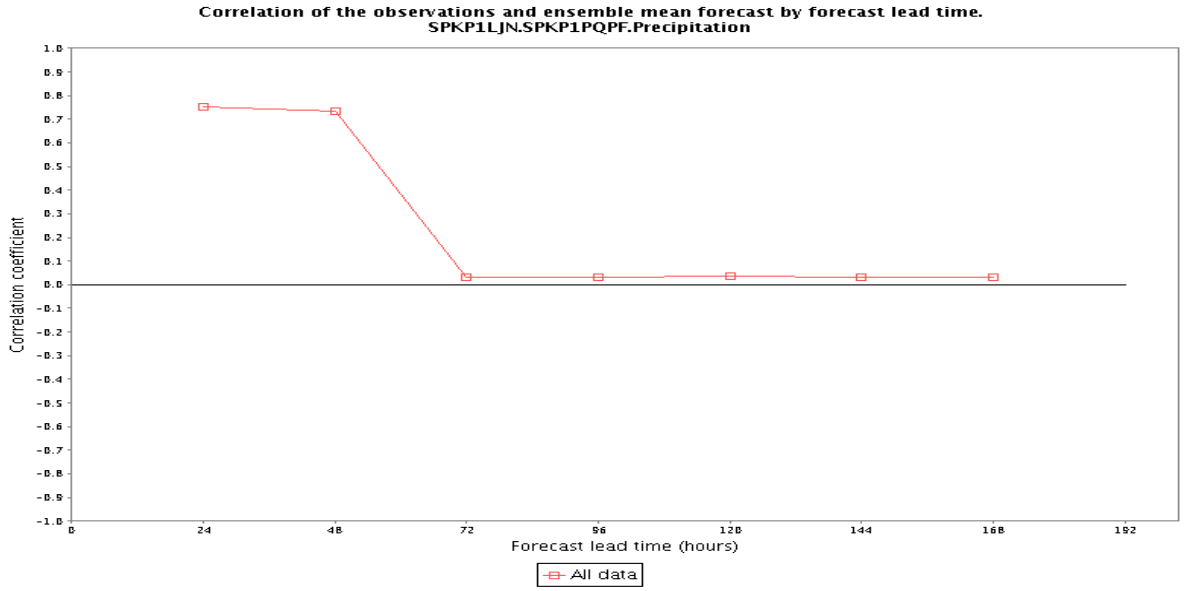
Main Conclusions (cont)

- At later lead times, when the first 48 hours of PQPF are the dominant forcing, spread increases and reliability improves somewhat. The forecasts are still underspread due to hydrologic uncertainty, particularly on the low end. Discrimination of events versus non-events remains high.
- The longest lead times incorporate climatological precipitation forcings. Reliability increases slightly, but discrimination decreases drastically and higher flows are more severely underforecast. Overall, the earlier lead hour forecasts are better quality, particularly for higher flows.

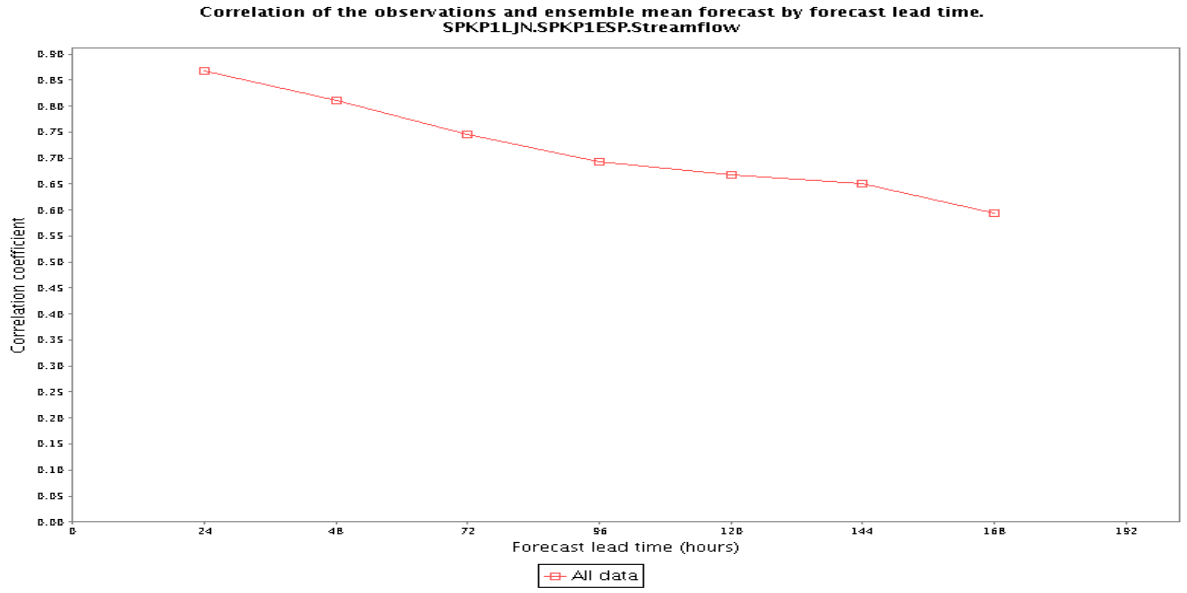
- Points downstream respond more slowly to basin rainfall, and therefore the transitions between no spread to QPF-forced to climatology-forced take more lead hours to develop.

EXTRA SLIDES

Linear Correlation of Ensemble Mean to Observation

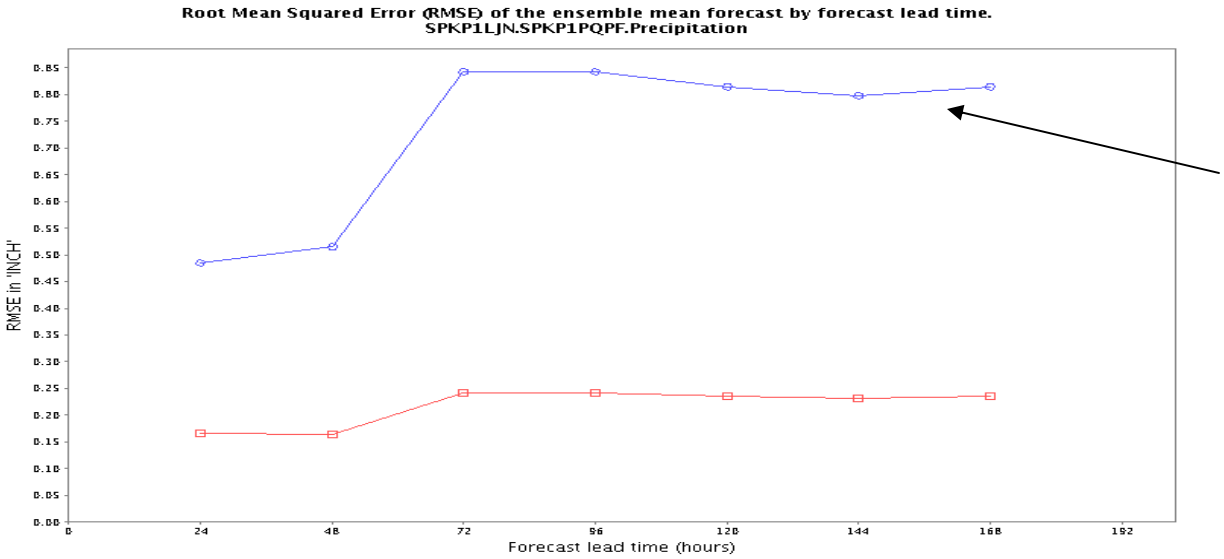


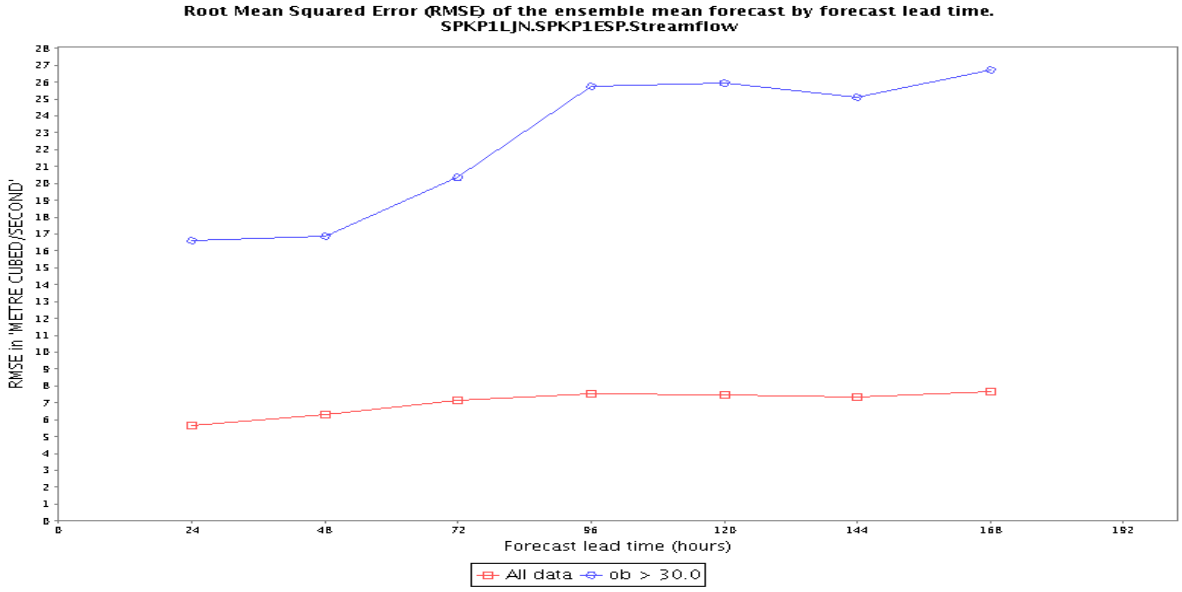
Daily Total Precipitation



Daily Average Streamflow

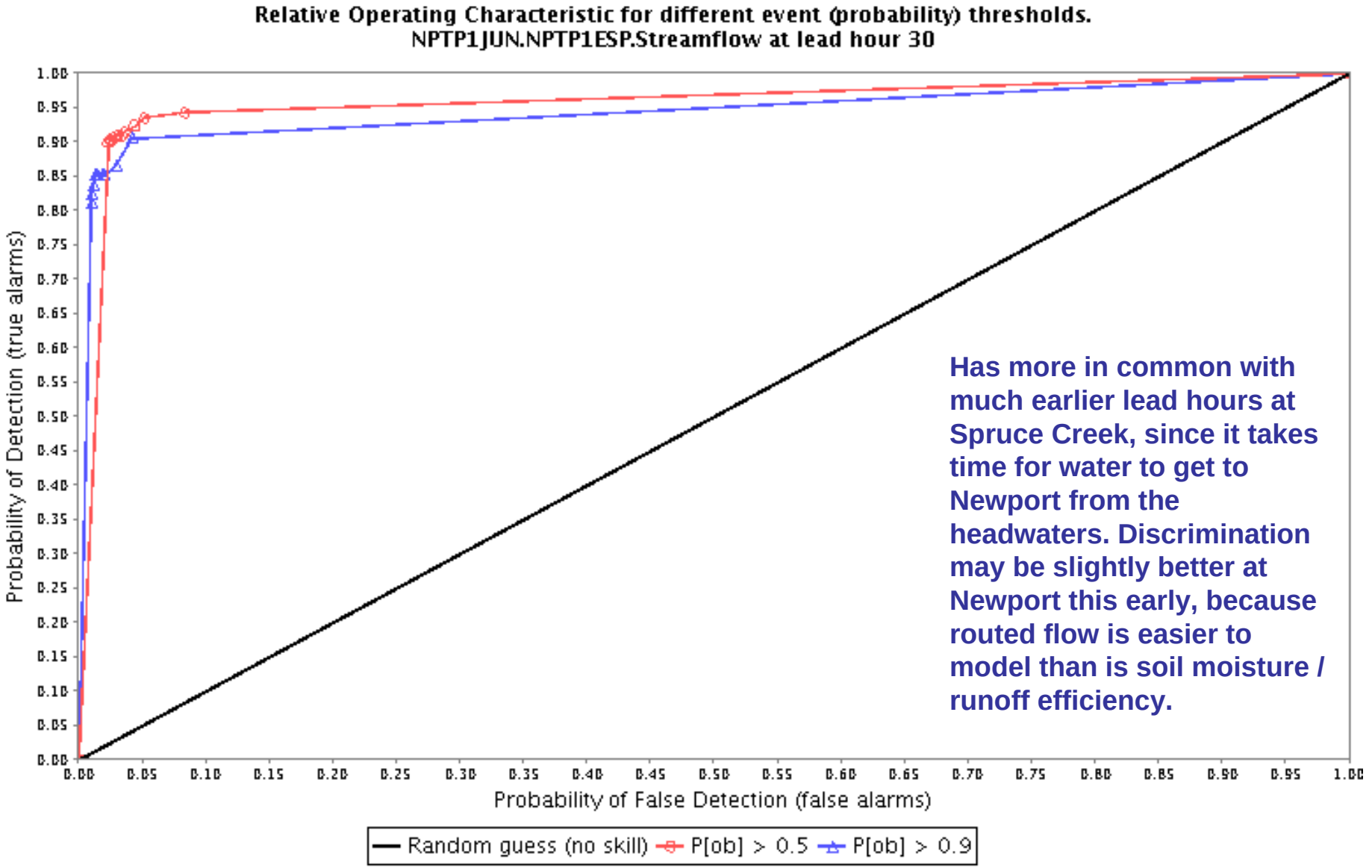
RMSE of the Ensemble Mean versus Observations





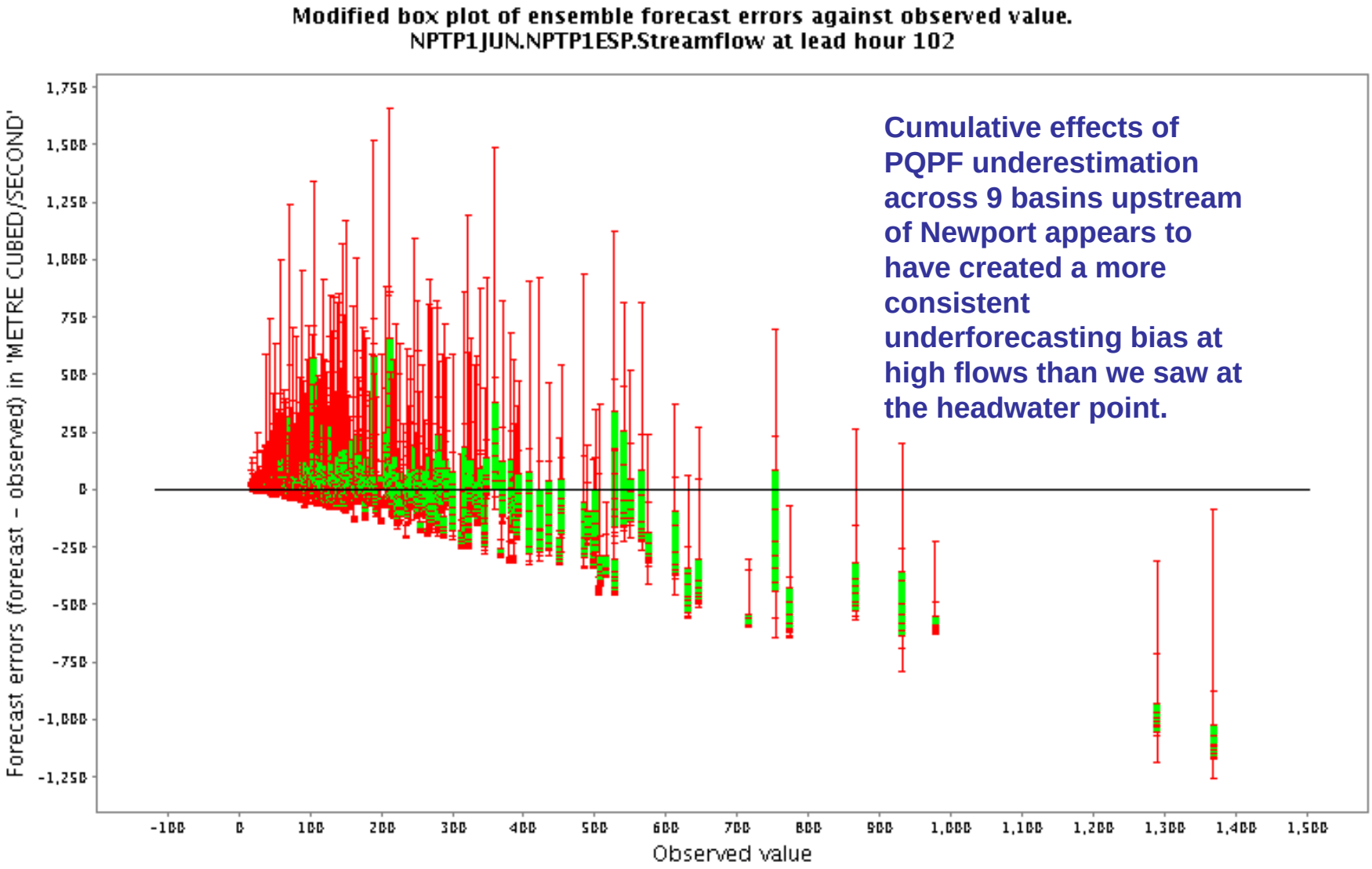
Sampling error due to missing forecast dates is the only source of these small variations in climatology-based PQPF error.

ROC Curve at Newport at Lead Hour 30



Has more in common with much earlier lead hours at Spruce Creek, since it takes time for water to get to Newport from the headwaters. Discrimination may be slightly better at Newport this early, because routed flow is easier to model than is soil moisture / runoff efficiency.

Newport Streamflow
Box Plots: Error versus Obs Value



Cumulative effects of PQPF underestimation across 9 basins upstream of Newport appears to have created a more consistent underforecasting bias at high flows than we saw at the headwater point.