# Probabilistic Forecast Verification

2004 DOH/RDM Science Workshop

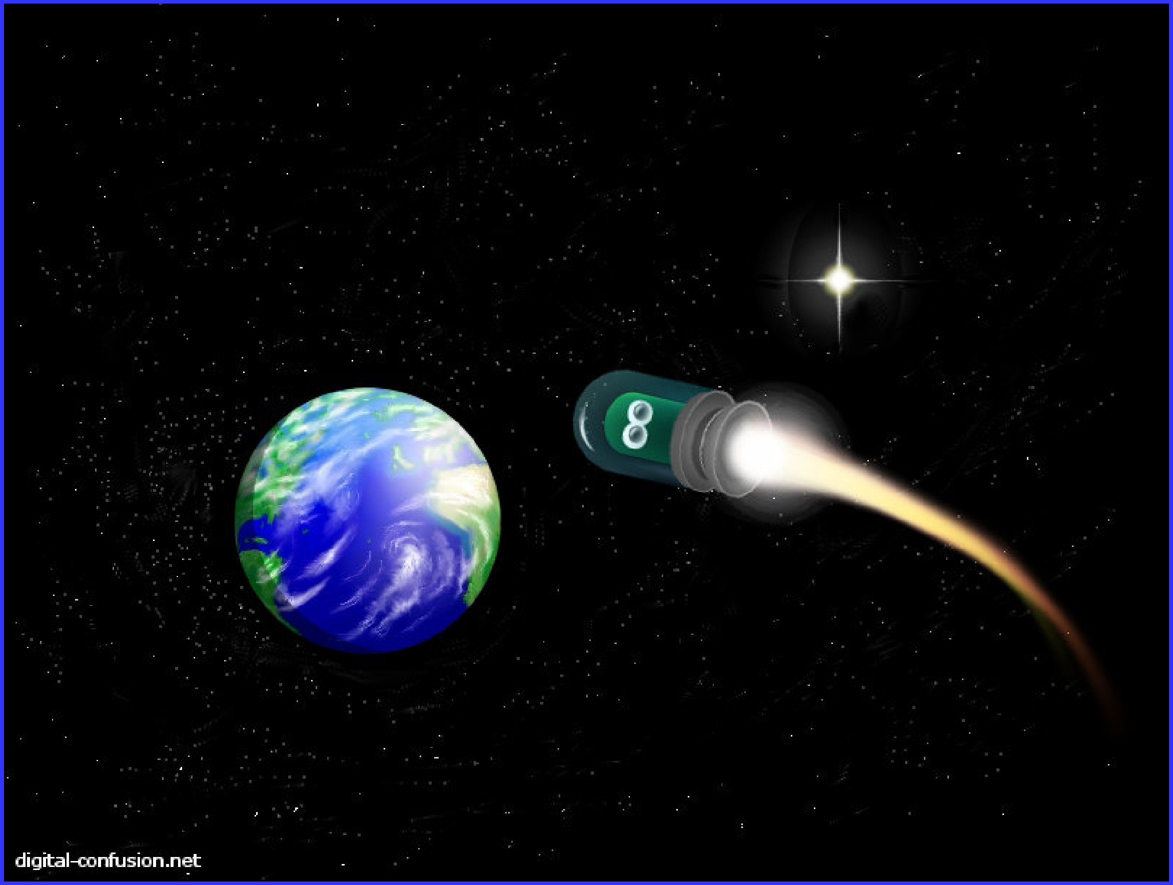
Silver Spring, MD
10 June 2004

Kevin Werner
WRH/SSD

# Outline

- Verification caveats
- Verification examples
    - Discrimination
    - Reliability
- Verification of MRF project at CBRFC
- Unanswered Questions

# OK, I made a probabilistic forecast… How can I tell if its any good?



“Probabilistic forecasting means never having to say I’m sorry” – Craig Peterson

# Probabilistic Forecast Verification 101

Caveats:

(1) A large (> ~20) number of *independent* observations are required.

(2) No “one size fits all” measure of success.

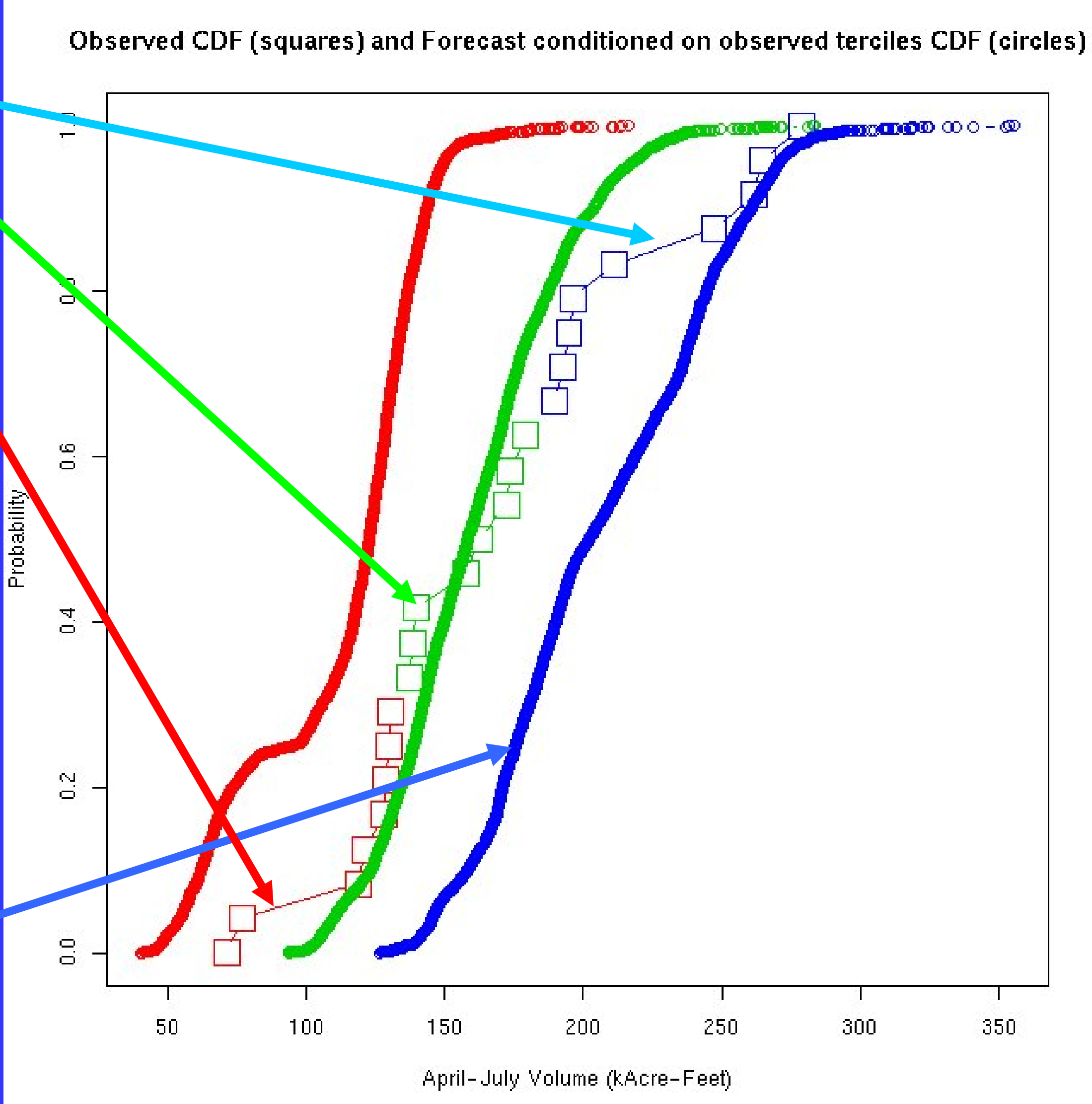
(3) Concepts are similar to deterministic forecast evaluation; However the application of the concepts is different.

# DISCRIMINATION Example

All observation CDF is plotted and color coded by tercile.

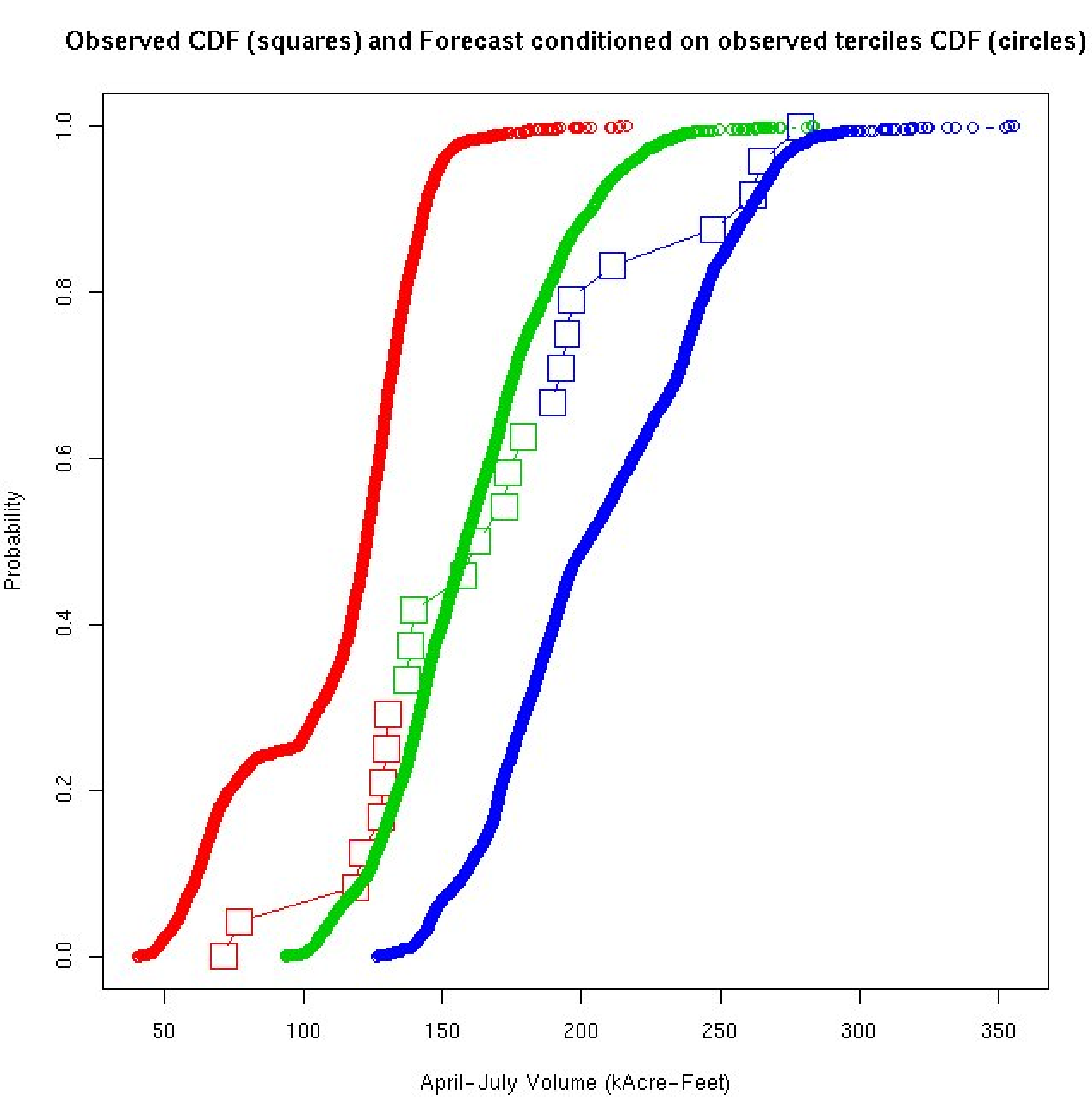
Forecast ensemble members are sorted into 3 groups according to which tercile its associated observation falls into.

The CDF for each group is plotted in the appropriate color. i.e. high is blue.



# DISCRIMINATION Example

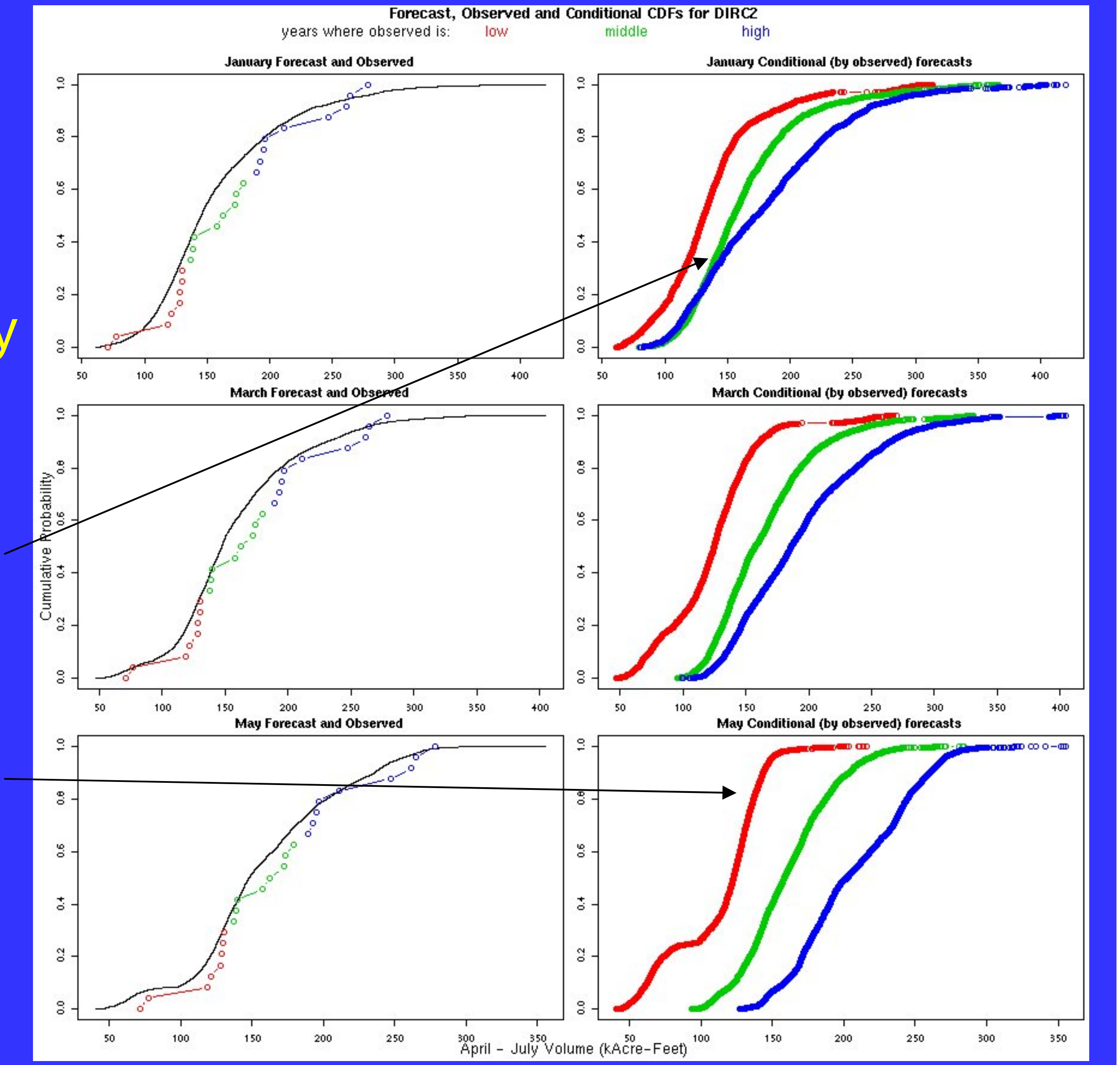
In this case, there is relatively good discrimination since the three conditional forecast CDFs separate themselves.



# DISCRIMINATION

How well do April – July volume forecasts discriminate when they are made in Jan, Mar, and May?

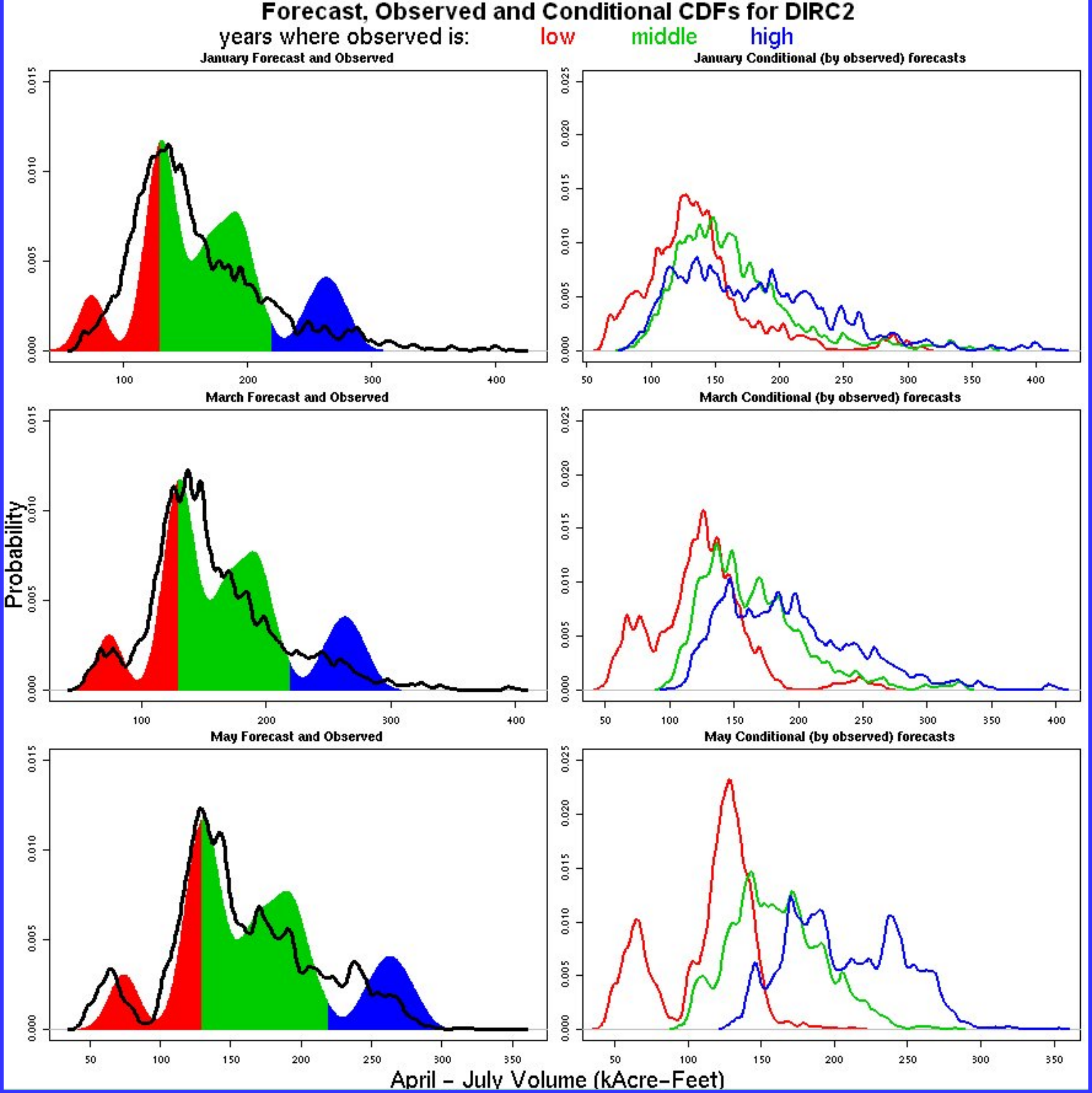
Poor discrimination in Jan between forecasting high and medium flows.  Best discrimination in May.



# Discrimination

Another way to look at discrimination using PDF’s in lieu of CDF’s.

The more separation between the PDF’s the better the discrimination.



# Reliability

“Reliability, or calibration, or conditional bias, pertains to the relationship of the forecast to the average observation for specific values of the forecast. Reliability measures sort the forecast/observation pairs into groups according to the value of the forecast variable, and characterize the conditional distributions of the observations given the forecasts.” Wilkes (1995)

Whereas discrimination examines the relationship between given observations and the subsequent forecasts, reliability examines the relationship between forecasts and the subsequent observations.

# Reliability Diagram

Reliability measures sort the forecast/observations pairs into groups according to the value of the forecast variable relative to an arbitrary value, and characterize the conditional distributions of the observations given the forecasts.

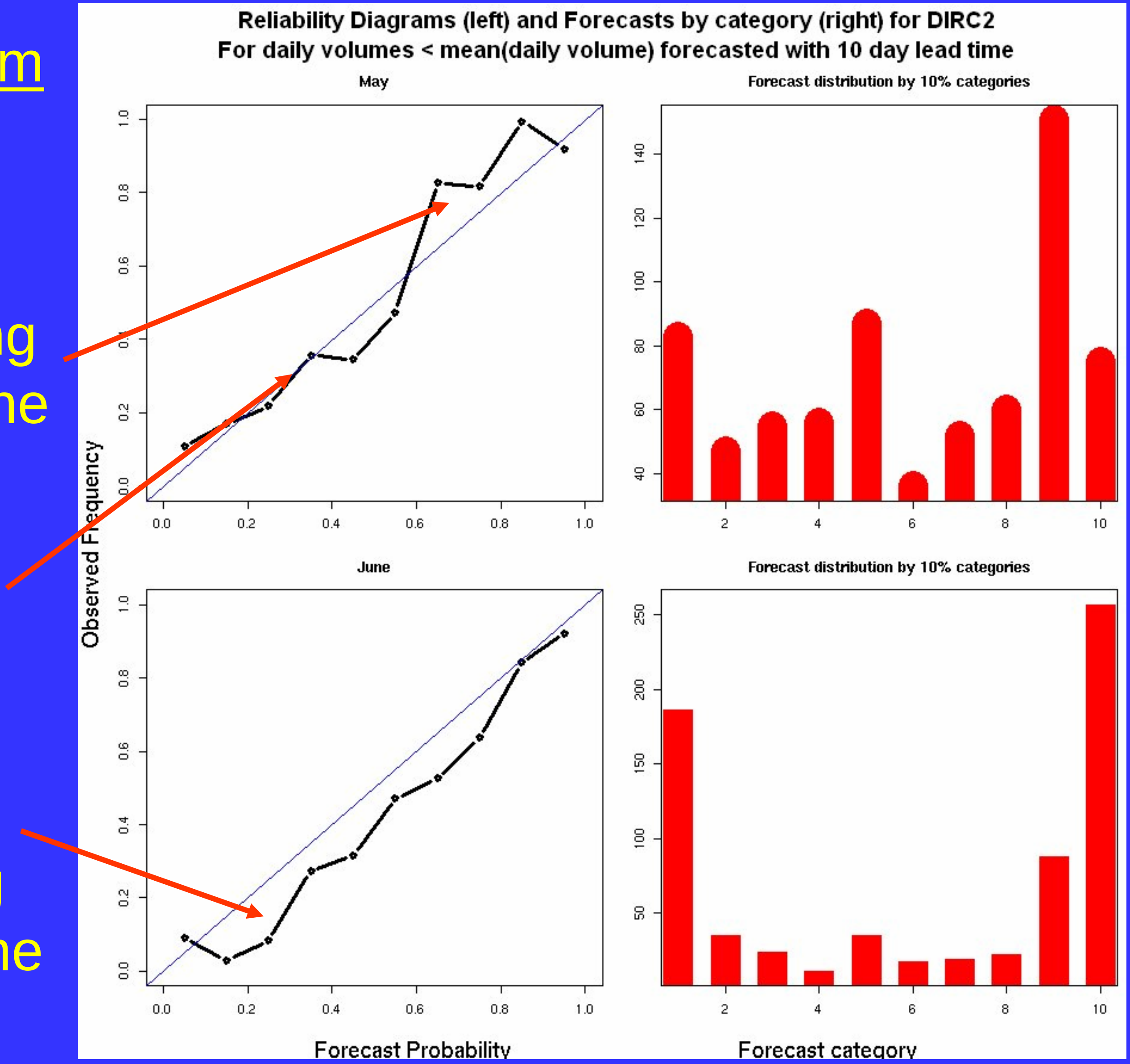
Traditional reliability diagrams transform a probabilistic forecast into a forecast of probability that an arbitrary value, such as flood stage or normal or …, will be exceeded. On one hand this limits the robustness of reliability as a verification measure. On the other, if the threshold value is of paramount importance, traditional reliability diagrams may be the most important verification measure.

# Reliability Diagram Example

Under Forecasting if area is above the diagonal

Perfect if on the diagonal

Over Forecasting if area is below the diagonal



# CBRFC MRF Project Verification

Daily probabilistic forecasts for each day in the snowmelt season.

Needed to verify…

# ESP Reforecast

Probabilistic forecast (or model) verification requires a large dataset. This was accomplished through reforecasting.

Reforecasts done for every basin for every day between 1979 – 1999.

Reforecasts made with both reforecasted MRF and historical MAT/MAPs.

# ESP Reforecast

For EACH reforecast day (i.e. 1/1/79, 1/2/79, … , 7/1/99)…

(1)NWSRFS carryover states created.

(2)MAT and MAP ensembles created from MRF reforecast.

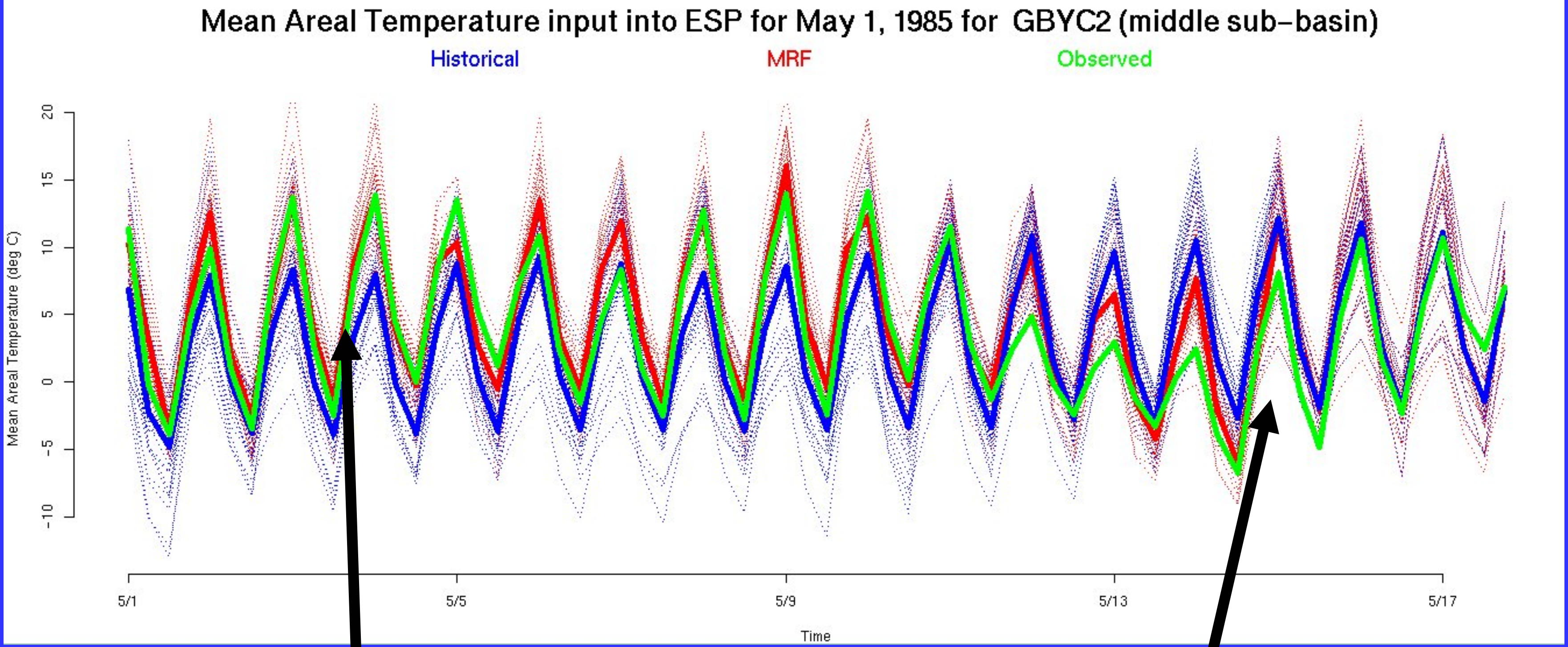
(3)Flow ensembles (i.e. *.CS time series) created from ESP.

# ESP Reforecast Example

Following example from Granby, CO (GBYC2) reforecast for May 1, 1985.

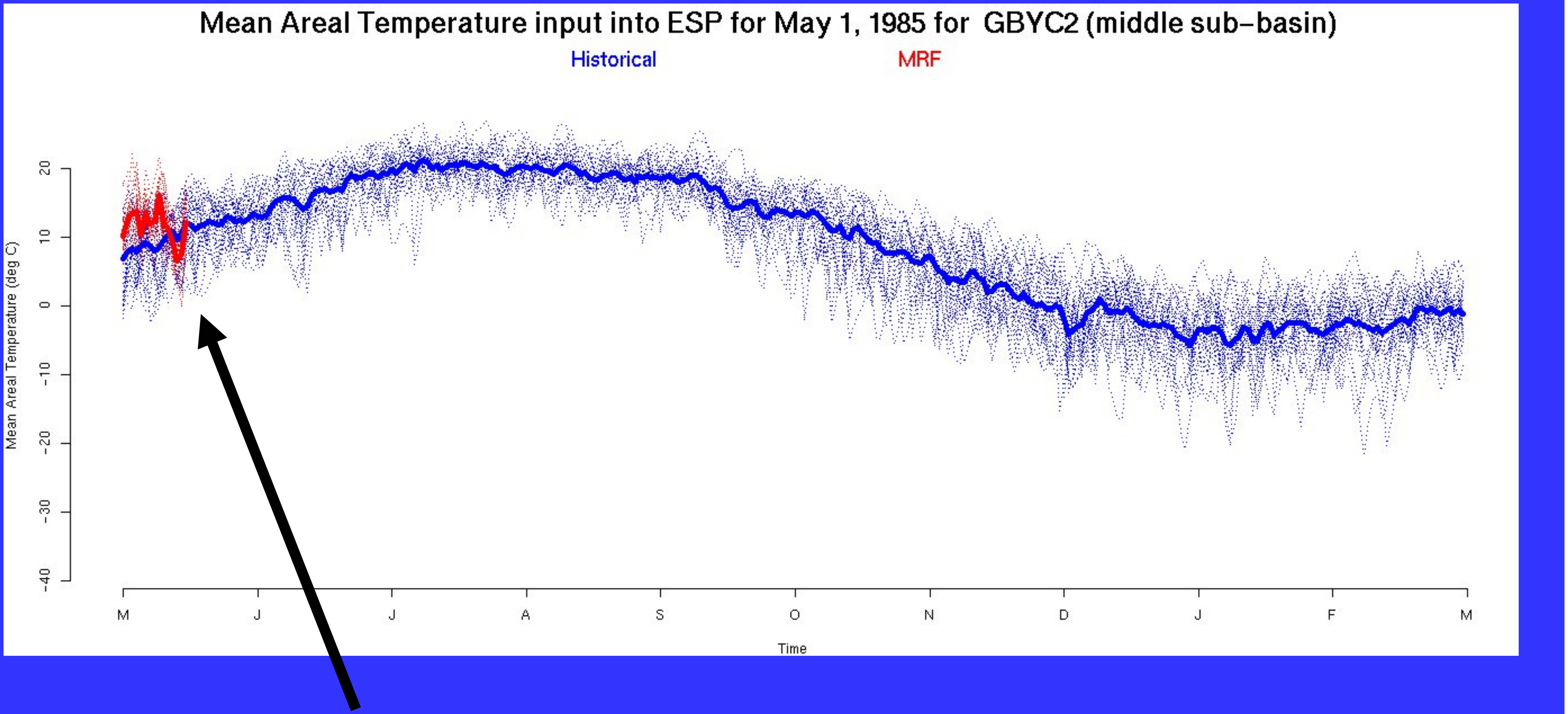
This is a snowmelt situation.

# Temperature Input into ESP



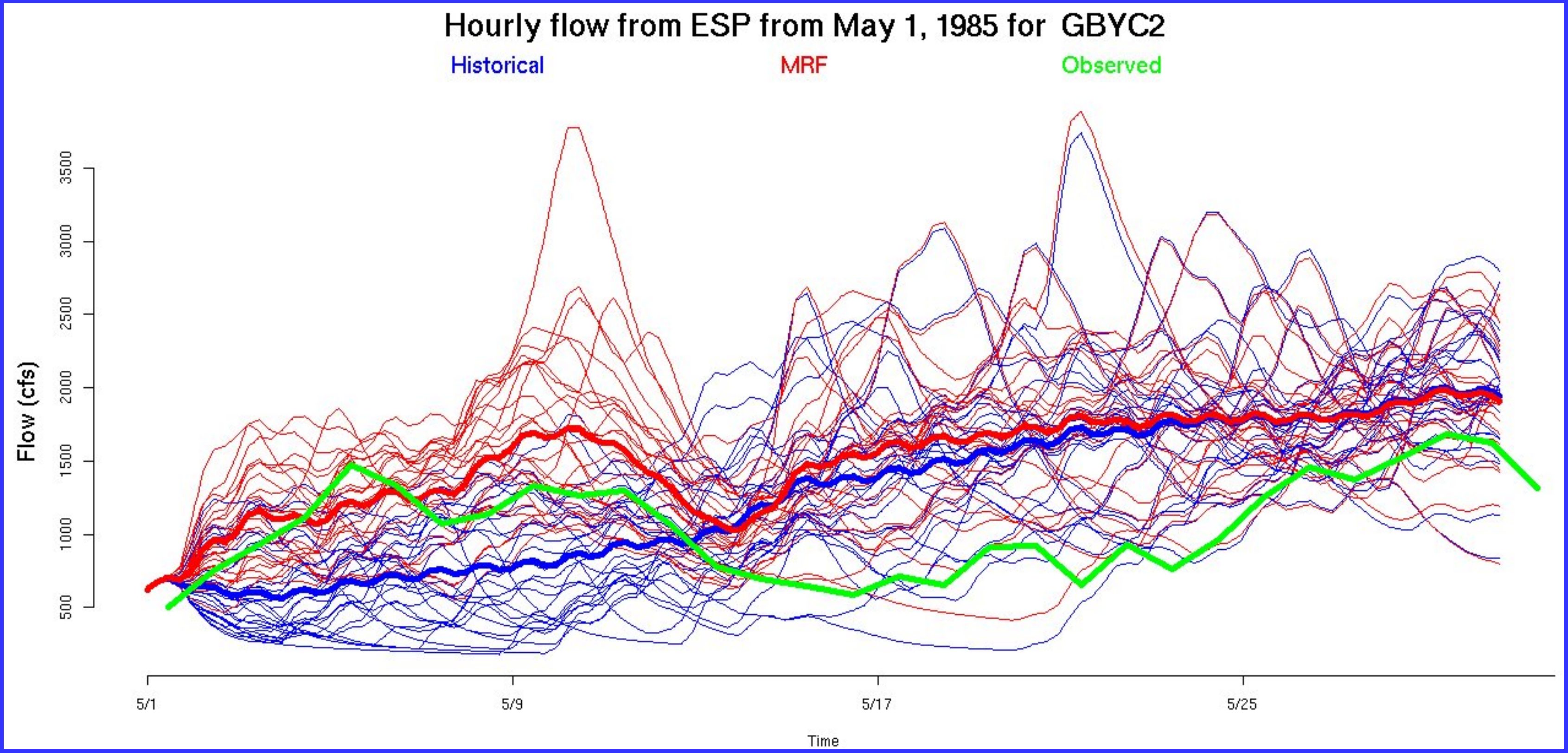
MRF derived MAT/MAPs are attached to historical years (“ensembles”) and ‘fed’ to ESP. Note MRF is warmer in first week

# Input into ESP



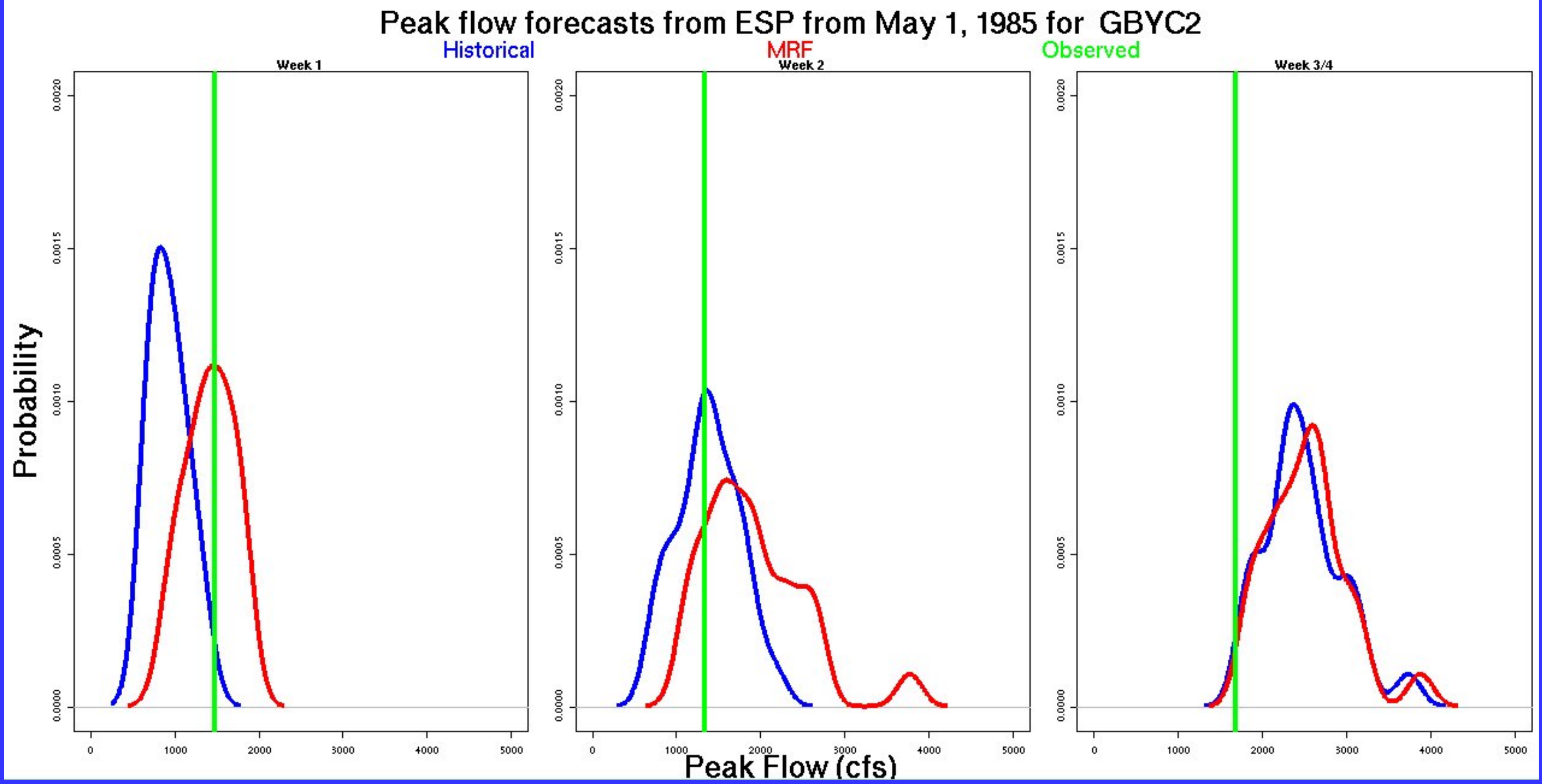
MRF derived MAT/MAPs related to the entire year of historical ensembles.

# ESP flow time series



Hourly instantaneous flow ensembles are created by ESP and saved. MRF shows higher flows than historical when it is warmer (during the first week). These may be converted into probabilistic forecasts…

# ESP peak flow



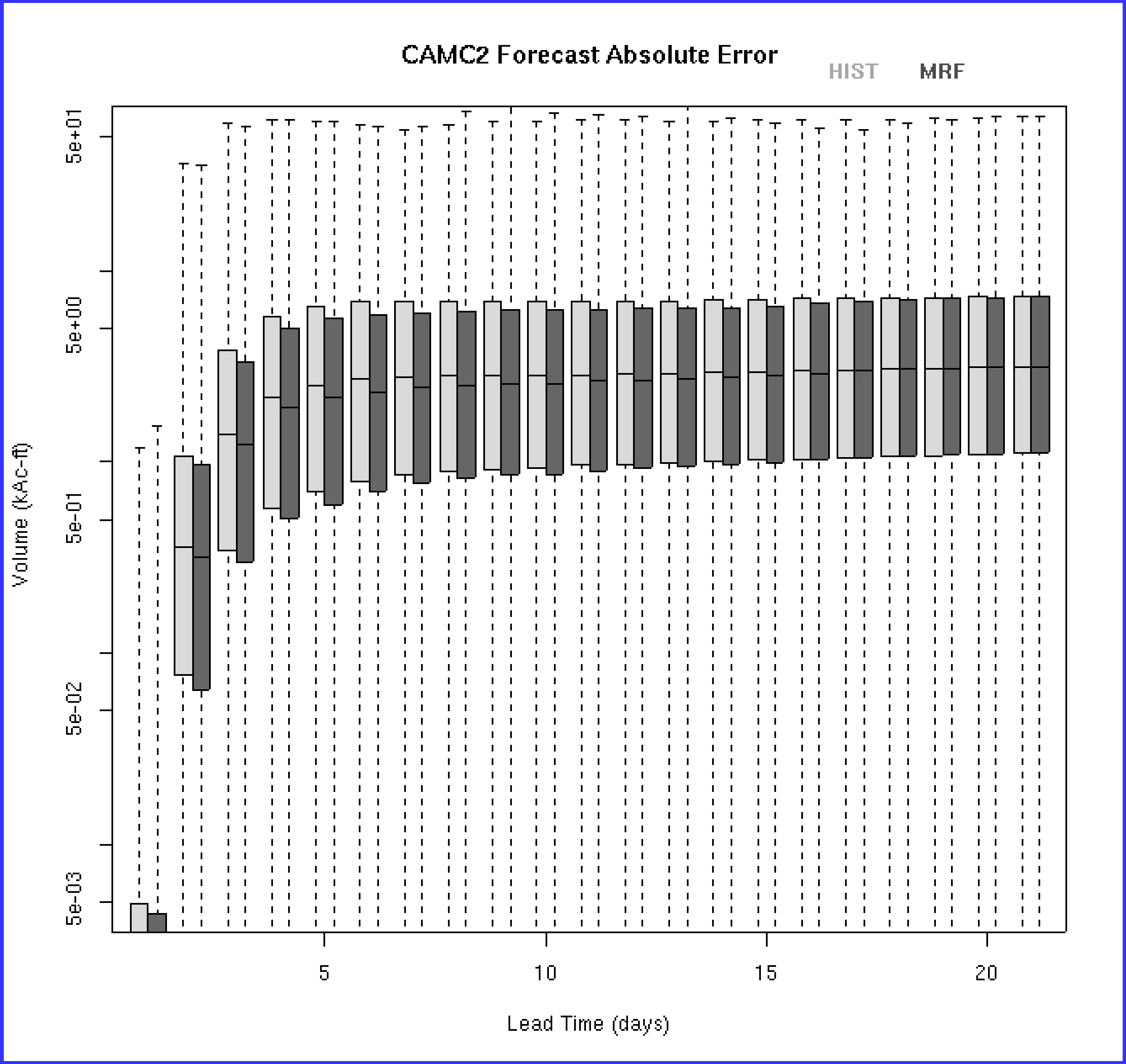
Peak flow forecasts shown as Probability Density Functions (PDFs). MRF shows higher probabilities in higher flows for two weeks.

## ESP Forecast Verification

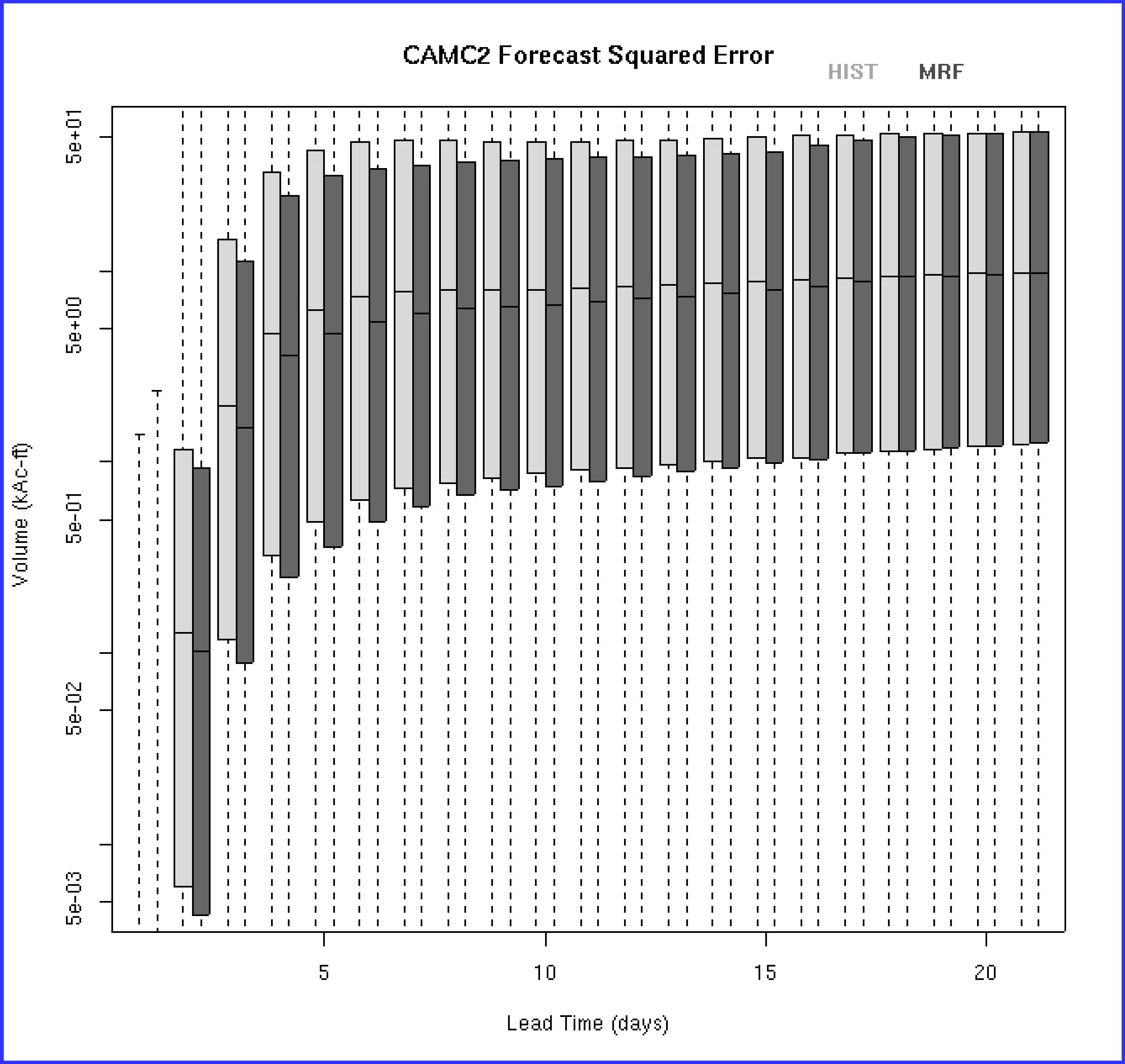
ESP forecasts may be verified as DETERMINISTIC forecasts. Traditional verification statistics such as Mean Absolute Error (MAE) and Mean Squared Error (RMSE) may be tallied from each forecast trace within an ensemble to show mean error statistics for the entire ensemble.

In this case, all forecasts made between April 1 and July 30 are aggregated by forecast lead time.

# ESP Forecast Verification



# ESP Forecast Verification



## ESP Forecast Verification

ESP forecasts are verified as a PROBABILISTIC forecast empirically derived from the ESP flow ensembles.

The Ranked Probability Score (RPS) and Ranked Probability Skill Score (RPSS) will be used quantify forecast skill improvement resulting from MRF.

RPS values will be calculated based on ESP reforecasts using MRF derived MAT/MAPs described here as well as purely historical MAT/MAPs.

# Ranked Probability Score (RPS)

The Ranked Probability Score (RPS) is used to assess the overall forecast performance of the probabilistic forecasts.

Similar to Brier Score but includes more than two categories.

A perfect forecast would result in a RPS of zero.

Gives credit for forecasts close to observation…
Penalizes forecasts further from the observation.

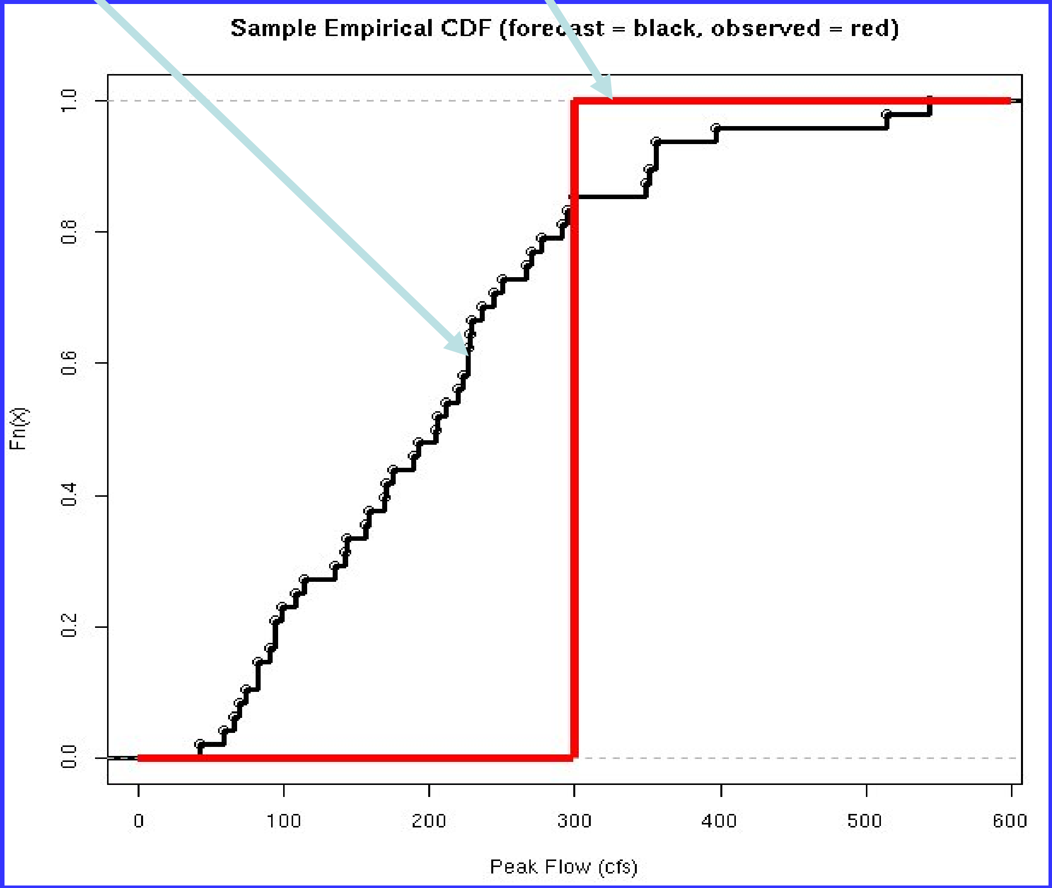
Looks at the entire distribution ( all traces ).

# RPS Formulation

Goal: Compare forecast CDF to observed CDF

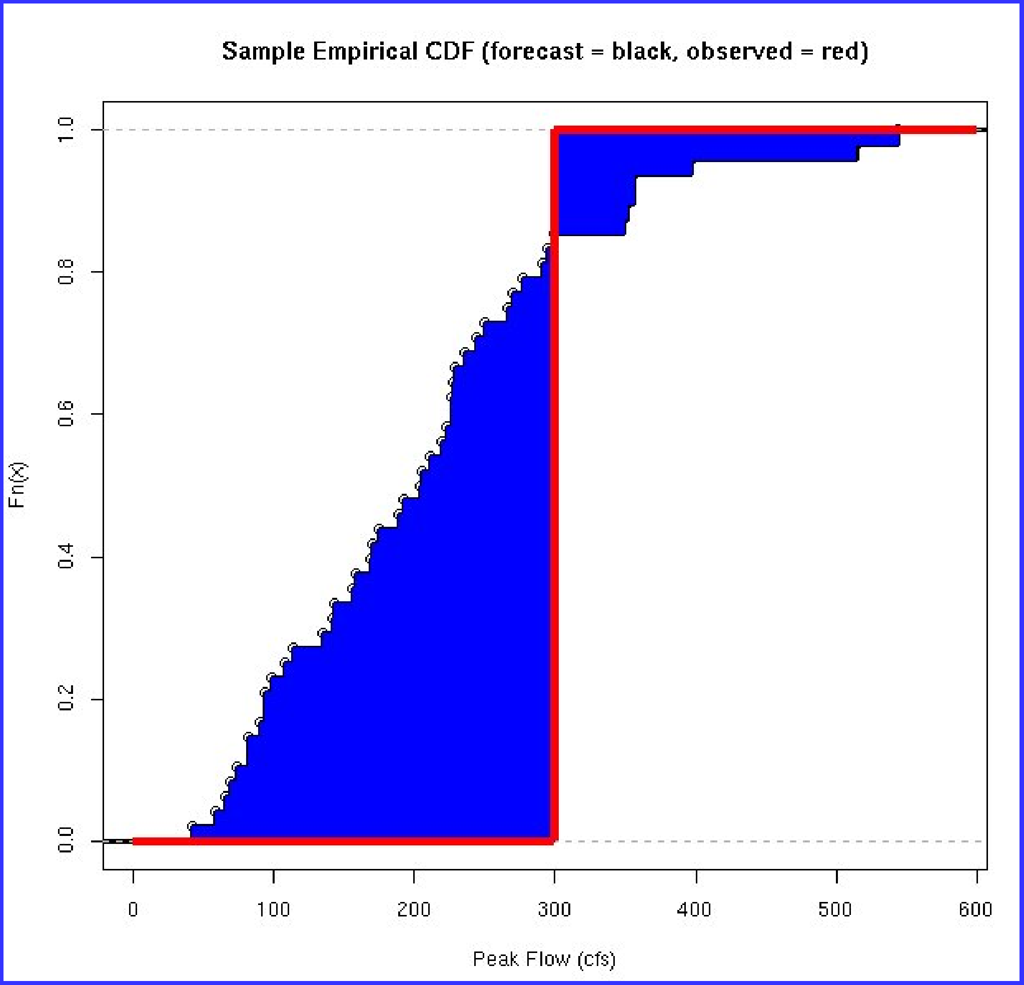
Notes:

1. Here an empirical distribution is assumed (not necessary).
2. Observation is one value, in this case 300 cfs.



# RPS Formulation

Graphically, the RPS is this area:



# RPS Formulation

Mathematically, RPS is given by:

$$RPS = \sum_{i=bin\#1}^{bin\#n} [P(forecast < i) - P(observed < i)]^2 / n$$

Where the summation indices are over n bins whose number and spacing are determined by the user. In order to best approximate the area between the forecast and observed CDFs, a large number of bins should be chosen. However, the larger the number of bins the more computationally intense the calculation becomes.

# Ranked Probability Skill Score RPSS

Useful to compare the forecast of interest to a reference forecast, e.g., climatology.

It is expressed as a percent improvement, e.g., over climatology ( or reference forecast ).

Perfect score is 100%.

Negative score indicates forecasts performed worse than reference forecast.

**Credit: Presentation-”Evaluation of NWS Ensemble Sreamflow Prediction” – Kristie Franz – U. of AZ**

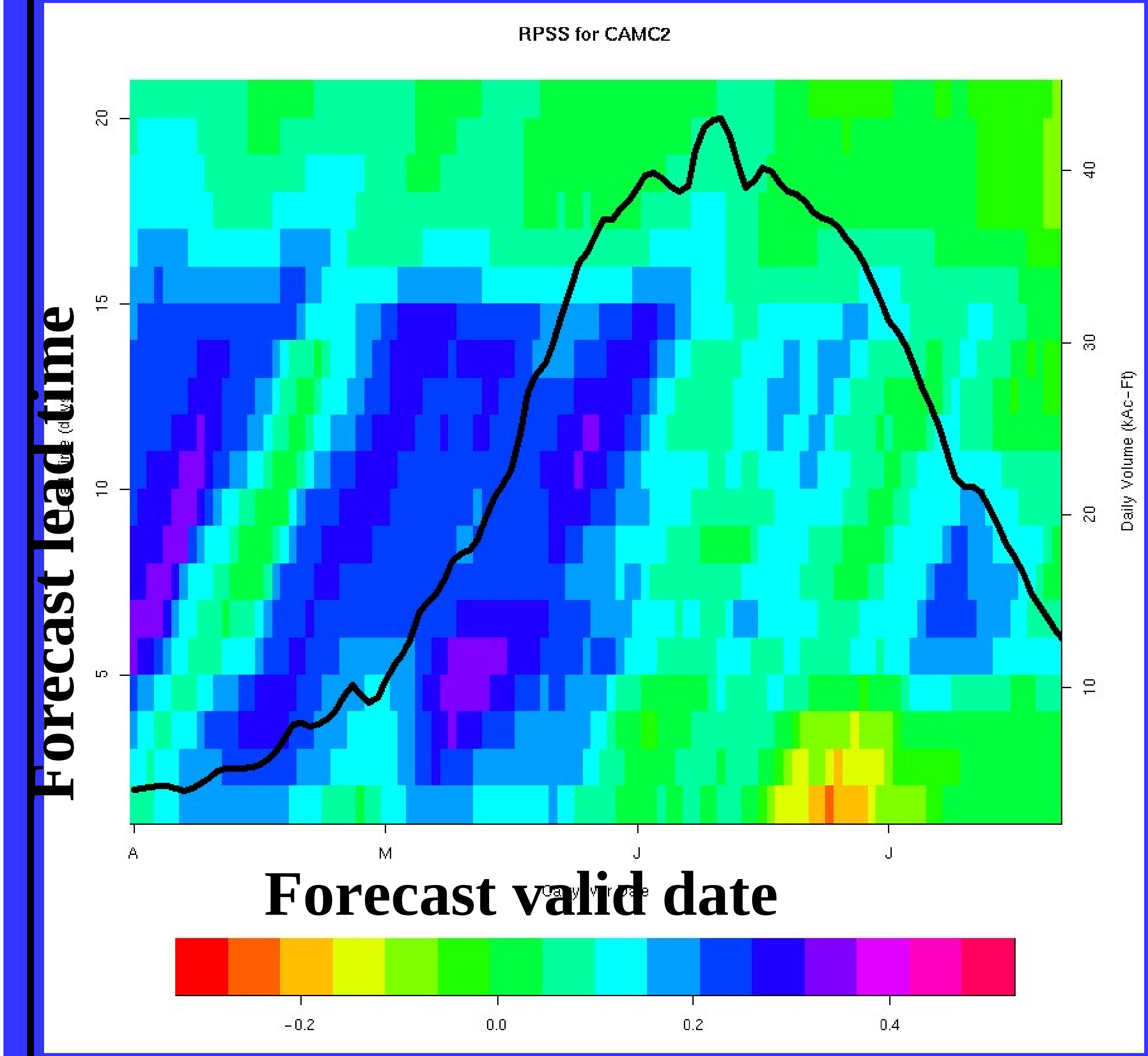
# Ranked Probability Skill Score RPSS

$$\mathrm{RPSS} = \frac{\mathrm{RPS}_{f} - \mathrm{RPS}_{cl}}{0 - \mathrm{RPS}_{cl}} \times 100\%$$

$\mathrm{RPS}_{f}$ = Rank Probability Score (forecasts)

$\mathrm{RPS}_{cl}$ = Rank Probability Score (climatology)

**$\mathrm{RPS}_{\mathrm{f}}$ and $\mathrm{RPS}_{\mathrm{cl}}$ must be calculated with the same bins!**

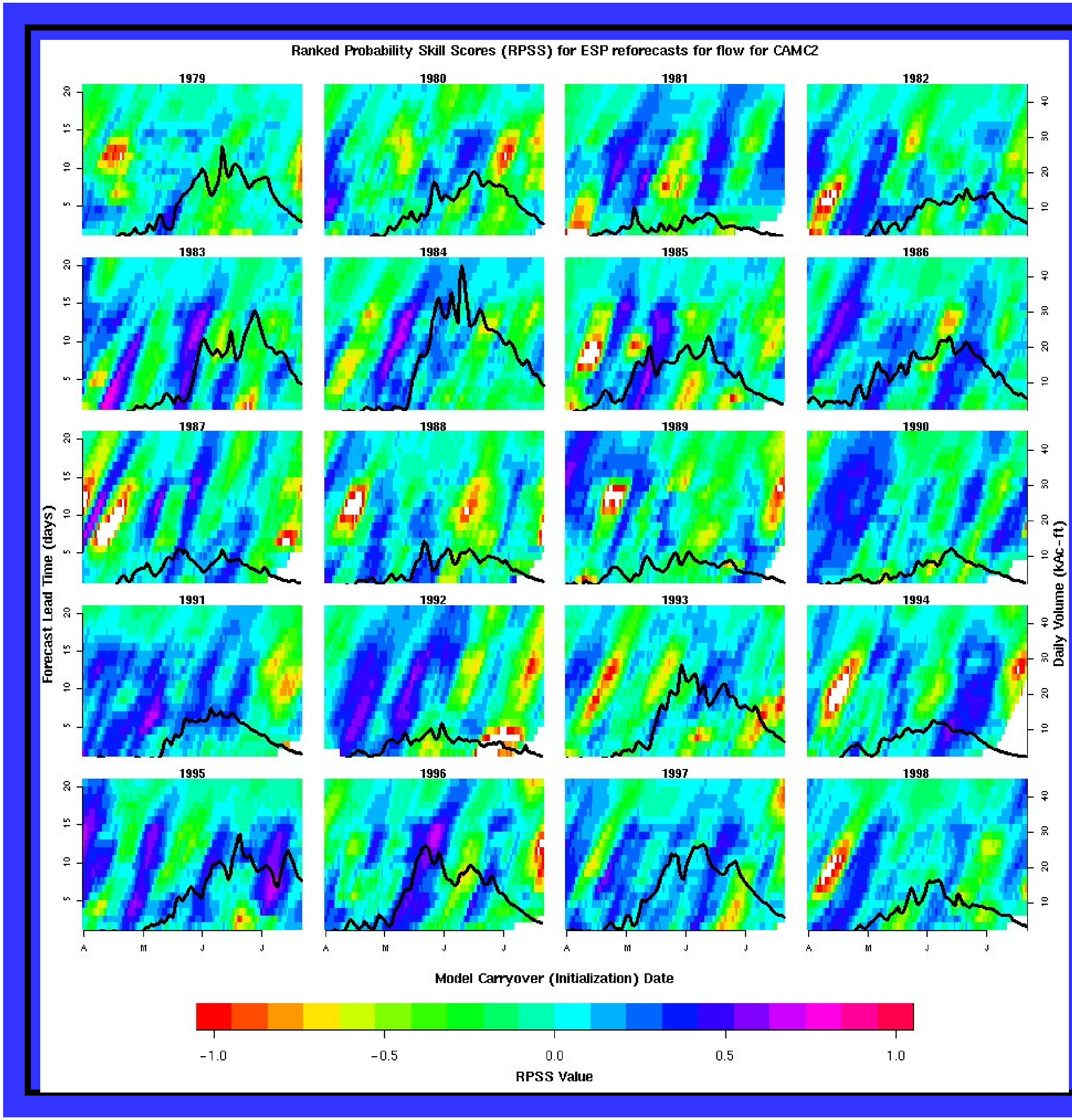


# ESP Forecast Verification

Mean hydrograph and RPSS values…

Good forecast skill improvements during rising limb of hydrograph.

# ESP Forecast Verification



Ranked Probability Skill Scores (RPSS) for ESP reforecasts for flow for CAMC2

## RPS Strengths

•Evaluates entire forecast distribution.

•No arbitrary threshold selection

•RPSS indicates skill over a reference forecast.

## RPS Weaknesses

•Sometimes difficult to interpret

•Multiple methods to compute statistic

# Unanswered Questions

- What are the “best” verification metric(s)?
    - Is there a “simple” metric that could be used to measure overall skill in a program (i.e. AHPS)?
    - Need to familiarize users (including RFCs).
- How to determine the signal to noise threshold?
- How can the verification metrics be related to the hydrology science?

# Credits:

Franz, Kristie and Sorooshian, Soroosh, 2002: Verification of NWS Probabilistic Hydrologic Forecasts, M.S. Thesis, Univ of Ariz.

Hamill, T.M., 1997: Reliability Diagrams for Multicategory Probabilistic Forecasts. Wea. Forecasting, 12, 736-741.

Hersbach, Hans, 2000: Decomposition of the Continuous RPS for Ensemble Prediction Systems, Wea. Forecasting, 15, 559-570.

Wilks, D.S., 1995: Statistical Methods in the Atmospheric Sciences: An Introduction. Academic Press, 467 pp.