Hydrologic forecast verification: The Upper Mississippi River Basin 2006-2008

Kristie Franz
Department of Geological and Atmospheric Sciences
Iowa State University

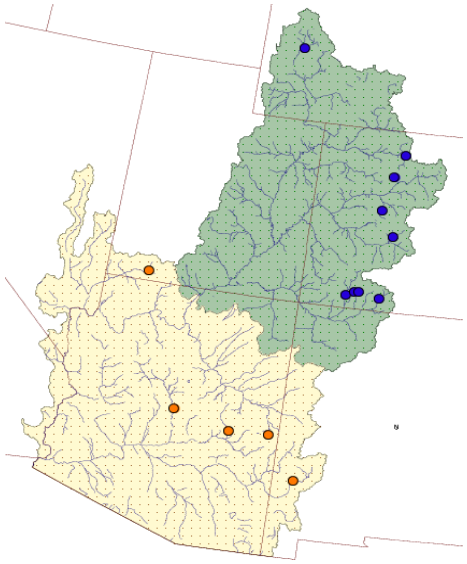
Mike DeWeese (NCRFC), Julie Demargne (OHD), Joe Bauman (ISU)

NWS-related work

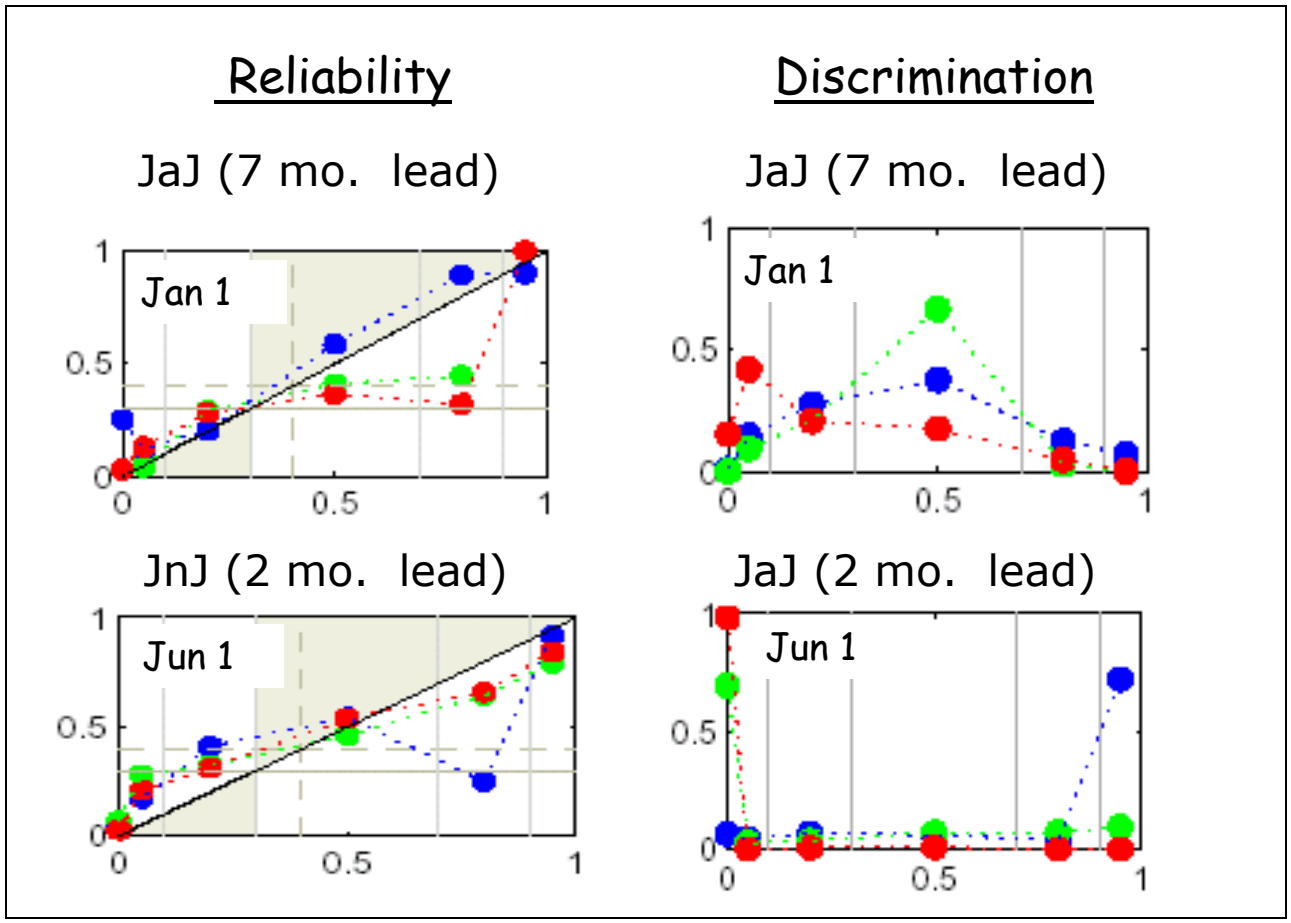
- ESP verification
  - Hindcast analysis (Franz et al., 2003)
  - OHRFC Archive investigation (Franz & Sorooshian, 2002)
  - ESPVS development (w/ RTI)

- Modeling
  - SNOW17 analysis (Franz et al., 2008 & in press)
  - SNOW17 energy balance modification (P. Butcher M.S. meteorology student, paper in prep.)
  - Flood prediction/testing HEC-HMS (collaboration is Des Moines WFO, S. Lincoln, M.S. Environmental Science student)
  - Data assimilation (w/Hogue and Margulis, UCLA)

ESP verification

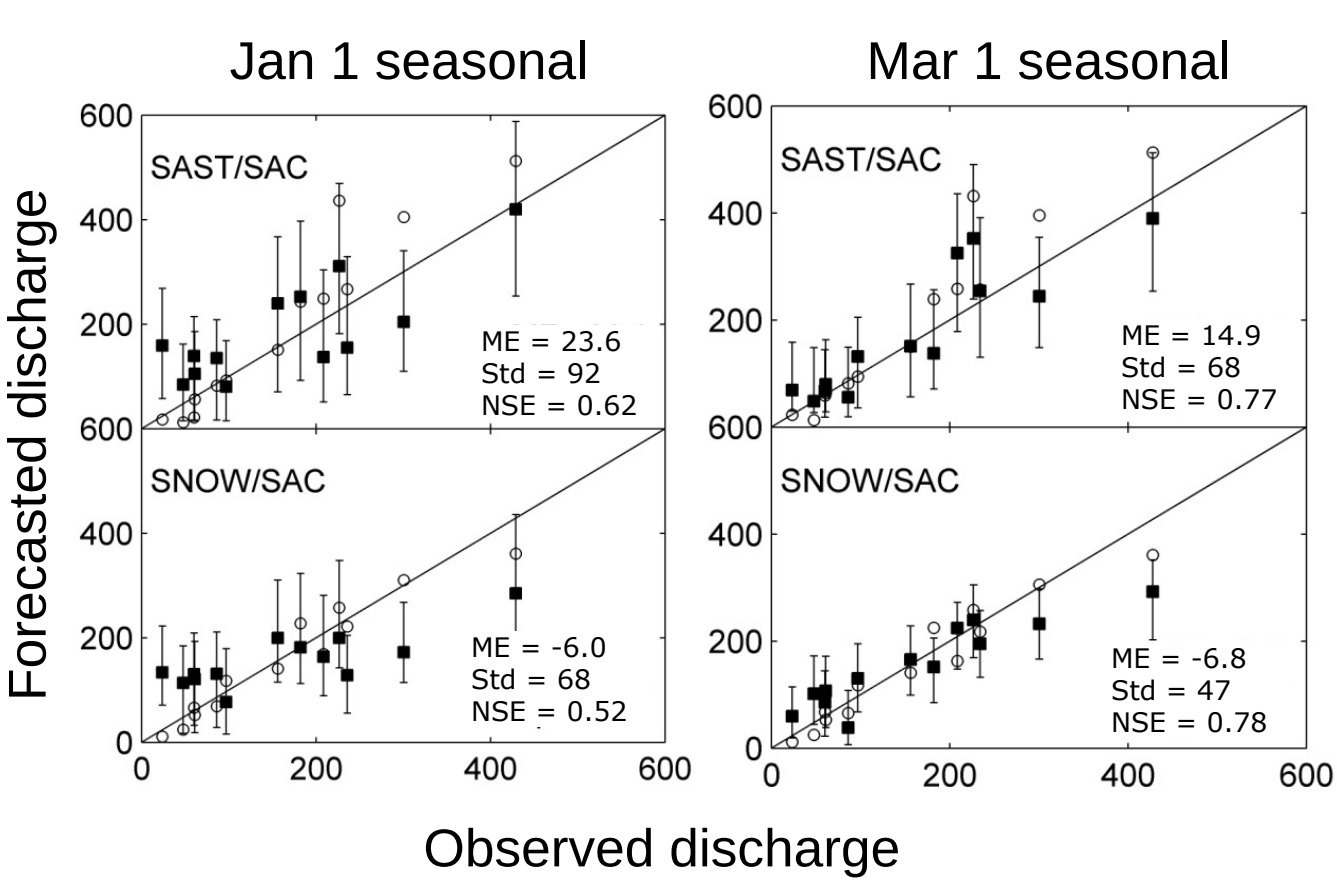


- CBRFC hindcasts



Franz et al., 2003, JHM

Model verification via hindcasts



Median forecast

Historical simulation

90% ensemble interval

Franz et al., 2008, JOH,
Franz et al., in press, JHM

ME = mean error
Std = standard dev.
NSE = Nash Sutcliffe

Current verification work

- Motivation
  - 2008 “500-year” floods in Midwest
  - Hearing:
    - “what can be done to improve the flood forecasts?”
    - “why were the forecasts so bad?”
  - Made me wonder…
    - Are these misconceptions?
    - Were the forecasts really that bad?
    - Maybe we were talking about 2 different types of forecasts (i.e. 100-year flood versus peak discharge)?

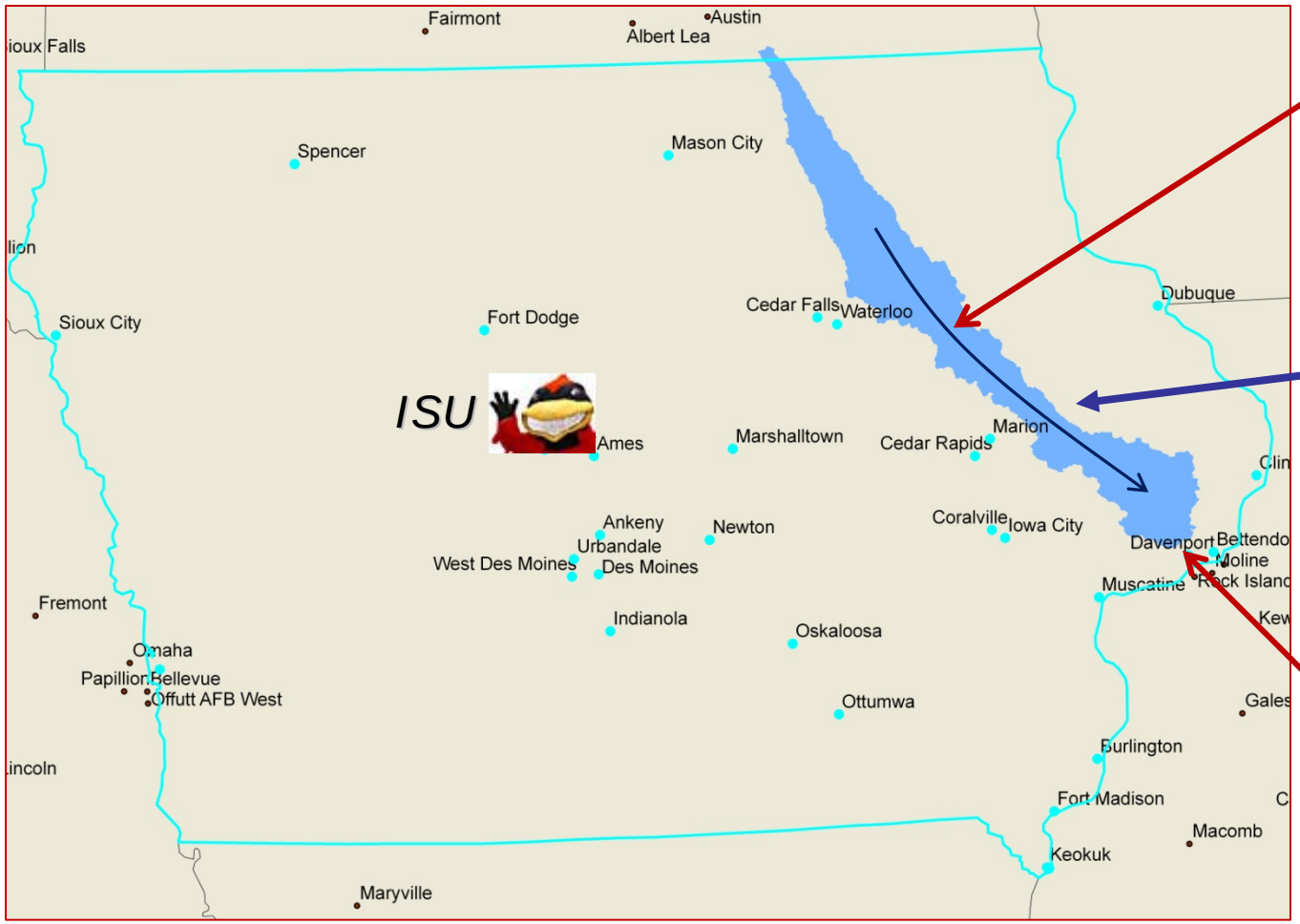
Photo: http://www.doobybrain.com/2008/06/18/houses-floating-down-the-cedar-river/

Verification Collaboration

- Julie Demargne & Mike DeWeese

- Goals:
  - Apply verification metrics to NCRFC forecasts
    - Compare and contrast metrics
      - Identify redundancies and inconsistencies
      - Identify key metrics and what they indicate
    - Generate verification data for NCRFC archive (esp. 2008 floods)

- Preliminary analysis from: Wapsipinicon Rv. on deterministic forecasts



Independence (as needed)

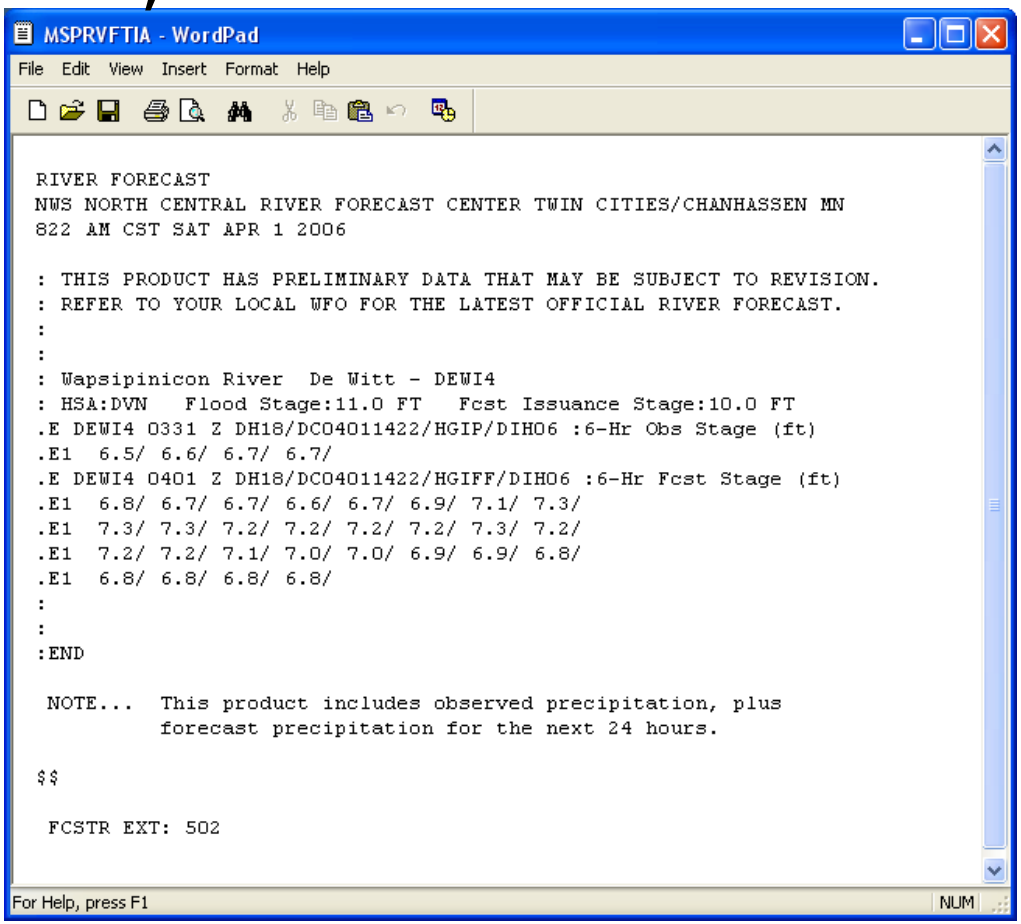
Animosa (as needed)

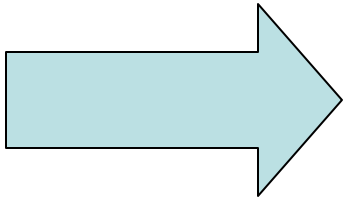
DeWitt (daily, some missing forecasts)

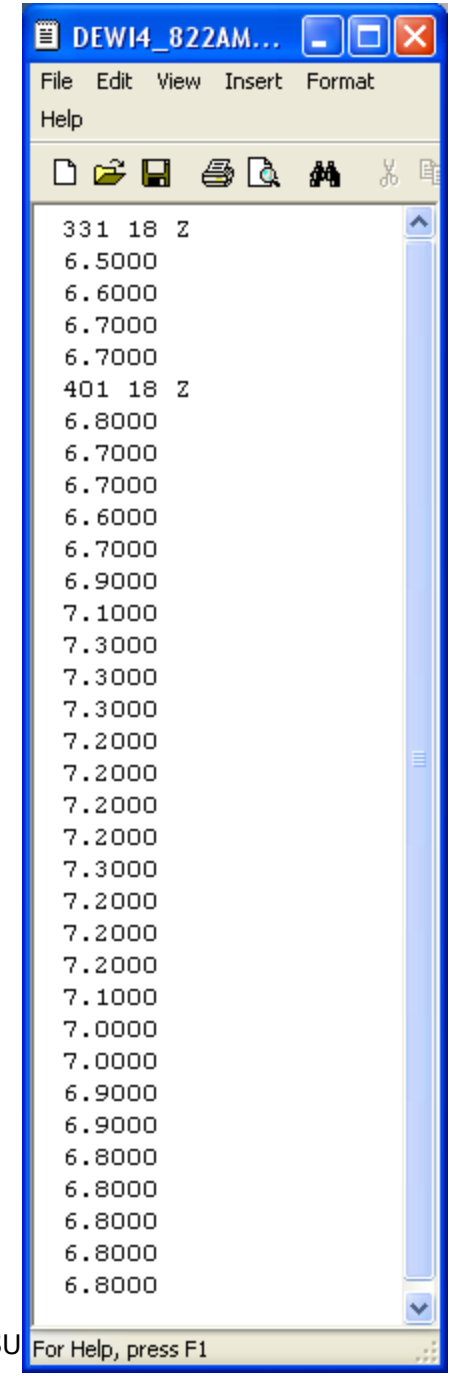
Forecast points evaluated

NCRFC archives (2006-2008

1,073 SHEF data files







Archive problems

- Forecaster notes are ignored in automated data extraction
- Format changes (i.e. headers) caused failed extractions
  - Some manual correction and processing required

Archive problems

- No archived discharge data
  - Mismatch in stage & discharge data timesteps
    - USGS: instantaneous daily max and min, & daily mean
    - NWS: 6-hour instantaneous forecast
  - Preliminary analysis on daily max stage
  - Forecasts that were issued after 6pm local were ignored

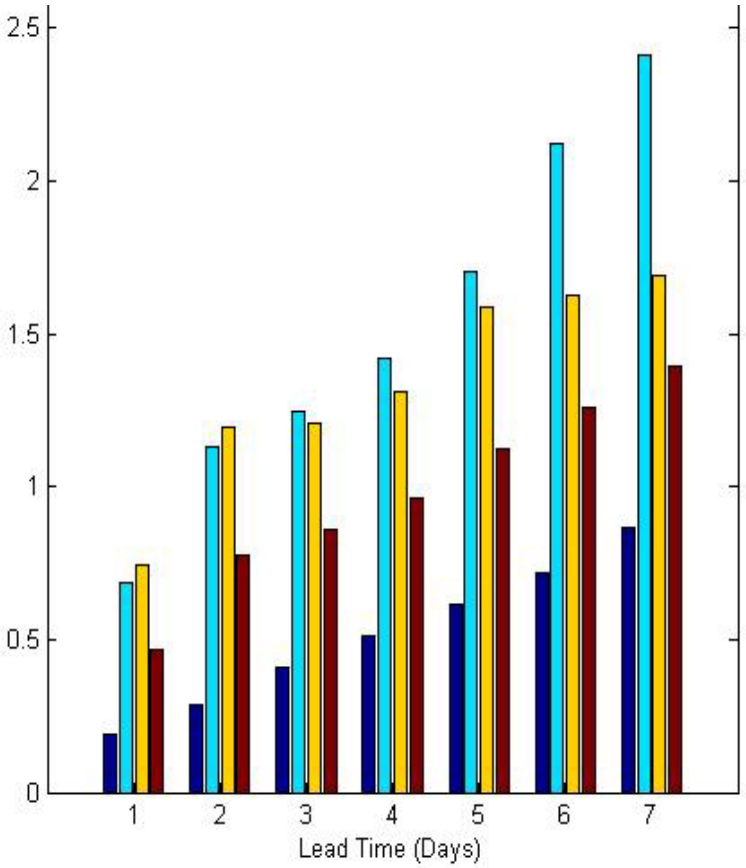
Verification Methods

- Accuracy measures
- Categorical statistics
- Skill scores using persistence
- Statistics summarized for individual sites and for watershed

Accuracy Statistics

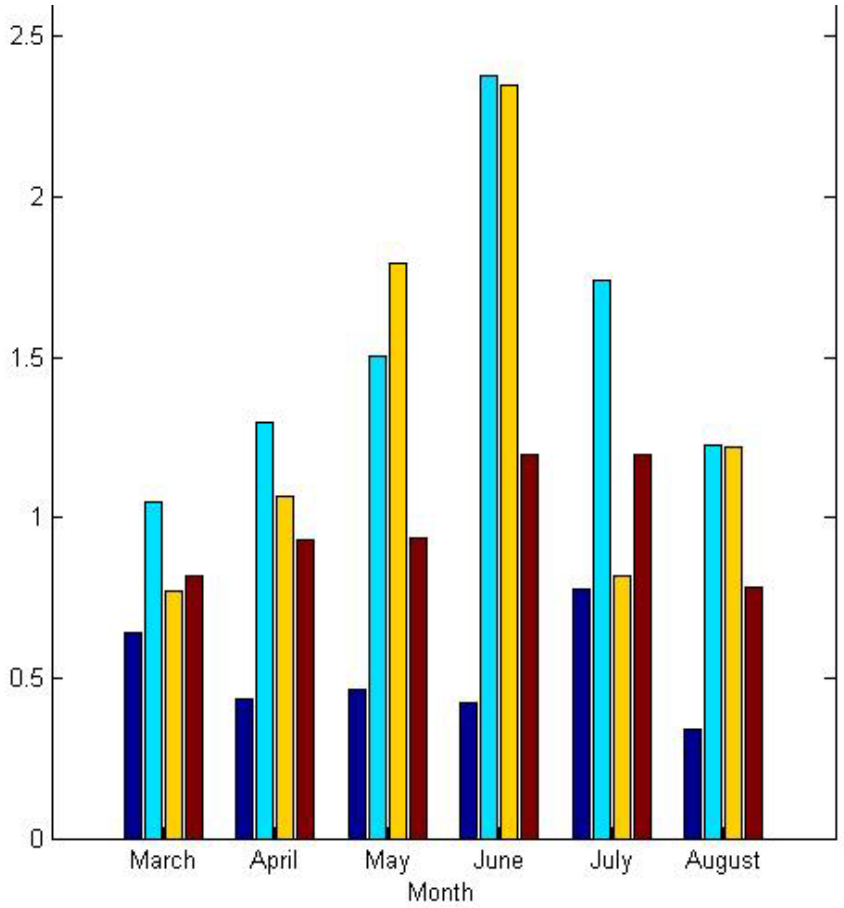
- Mean Absolute Error
- Root Mean Square Error
- Mean Error
- Min Absolute Error
- Max Absolute Error
- Nash-Sutcliffe
- Percent Bias
- Correlation

Mean Absolute Error (lead time)



- Error increases with lead time
- Forecasts improve downstream
  - DeWitt most accurate

Mean Absolute Error (monthly)



- Highest average error in June and July
  - Sample issue, basin conditions, tiles?

Question

- What magnitude error is “good”?

Independence

Flood Stage: 12.0’
Major Stage: 15.0’

Animosa

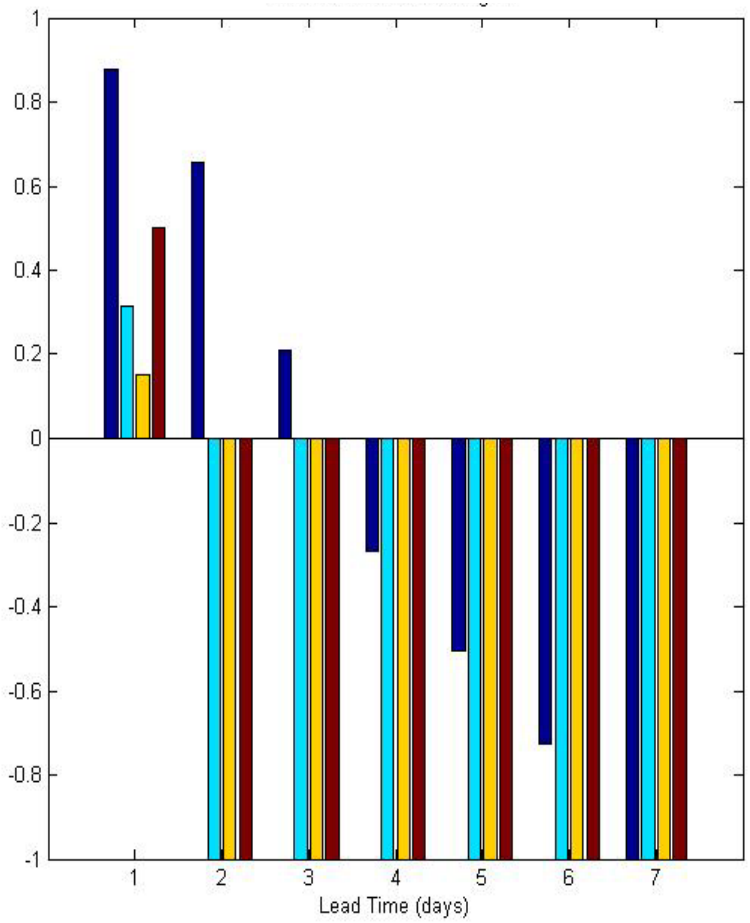
Flood Stage: 14.0’
Major Stage: 19.0’

DeWitt

Flood Stage: 11.0’
Major Stage: 12.5’

Nash-Sutcliffe (lead time)

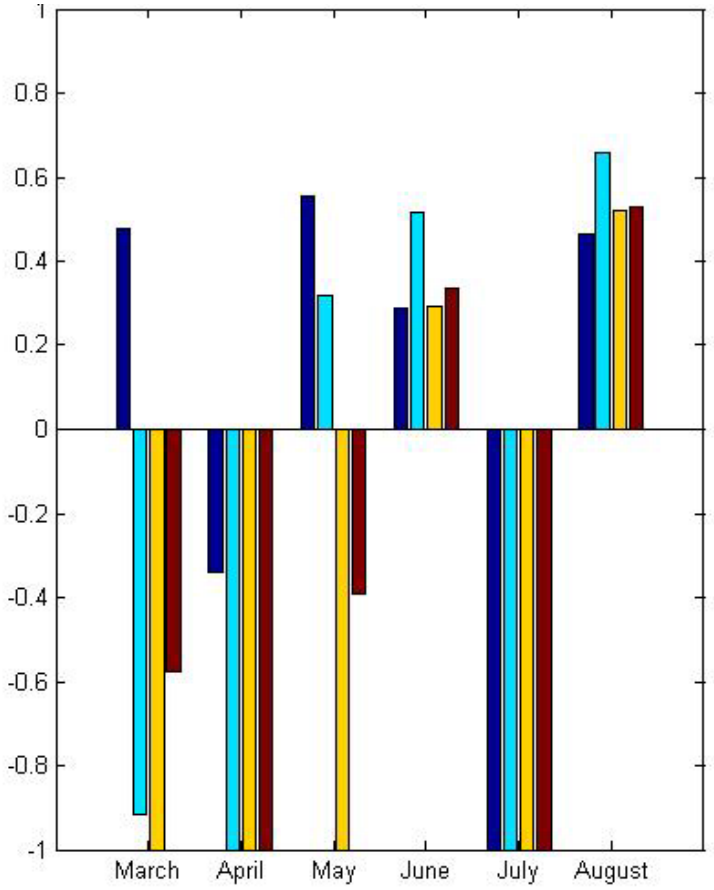
$$NSE = 1 - \left(\frac{\sum_{t=1}^{N} (x_t - y_t)^2}{\sum_{t=1}^{N} (y_t - \bar{y}_t)^2} \right)$$



DeWitt
Anamosa
Independence
All Sites

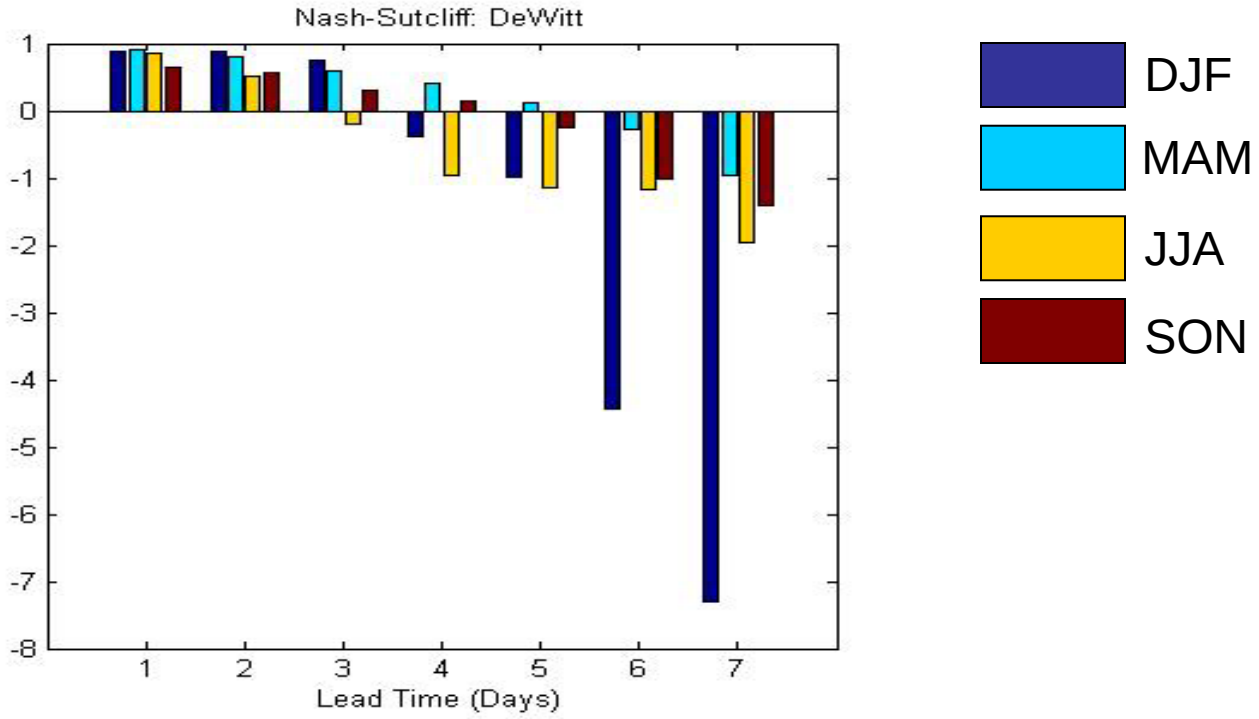
- Ratio of the forecast error to the observation variance.
- Perfect score = 1;
  0 = mean of obs better predictor than forecast
- Predictability drops off after day 1
- Improvement downstream

Nash-Sutcliffe (monthly)



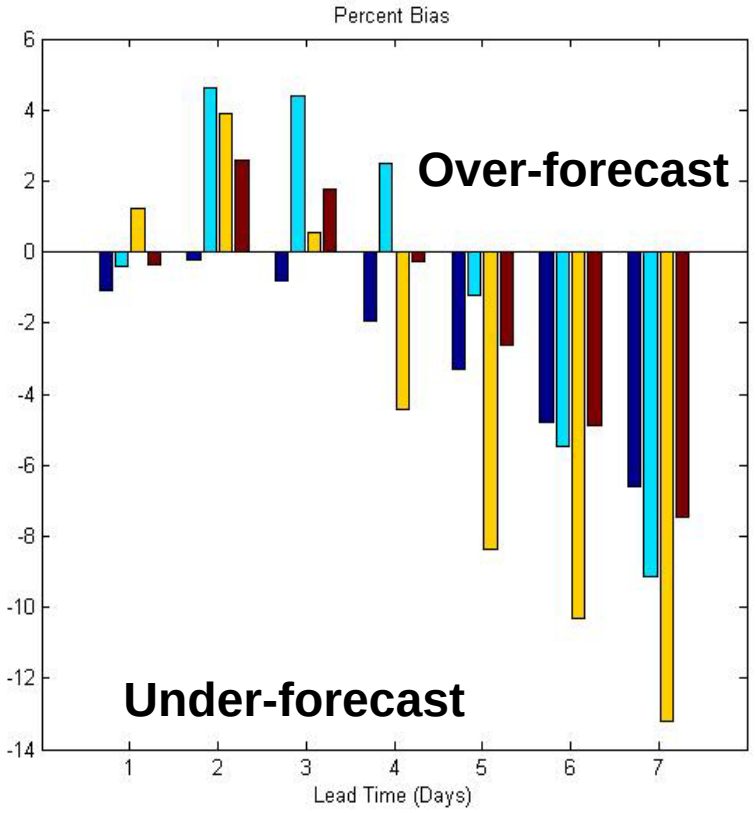
- Scores affected by very poor scores in July 2006, 2007, and 2008.
- No predictability in April or July on average

Nash-Sutcliffe (DeWitt)



- Looking at scores by month and lead time

Percent Bias



- Similar trends
  - poorer scores with increased lead time
  - June & July largest biases
- More negative w/ increased lead time (underforecasting due to QPF?)

$$\%BIAS = \left[\sum_{t=1}^{N} \left(Q_{sim,t} - Q_{obs,t} \right) / \sum_{t=1}^{N} \left(Q_{obs,t} \right) \right] \cdot 100$$

Categorical Statistics

- 2x2 contingency table
- Observations conditioned on:
  - Below Flood Stage, Flood Stage, and Major Stage.

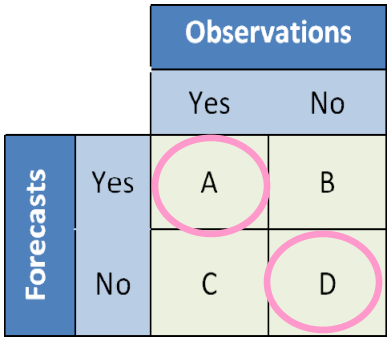
		Observations	
		Yes	No
Forecasts	Yes	A	B
	No	C	D

Probability of detection
Probability of false detection
Categorical Bias
Gilbert skill score
Critical success index

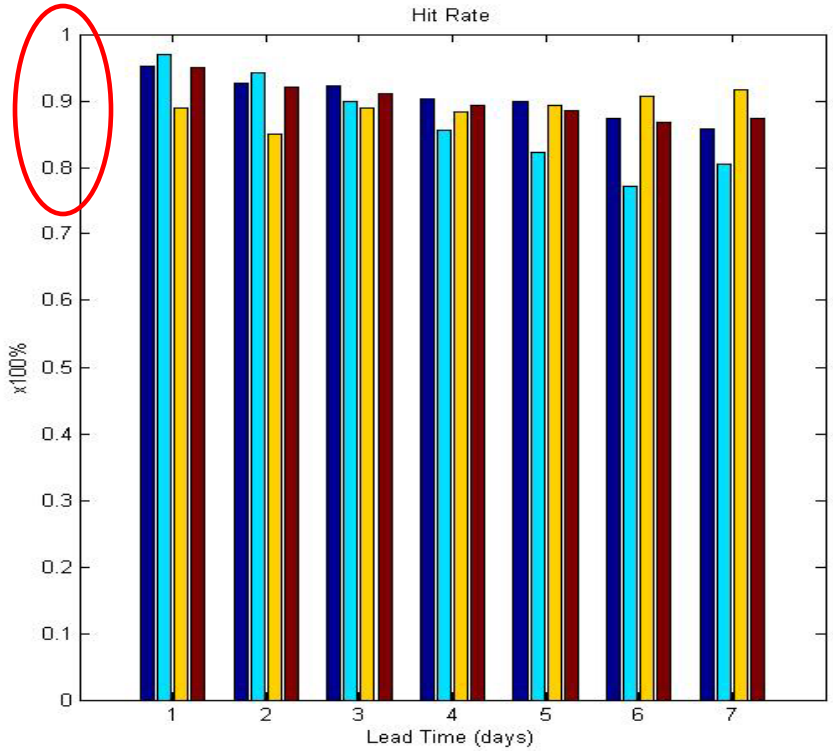
Percent correct

- PC ((A+D)/n)

  - not overly beneficial statistic since flooding events are rare compared to non-flooding events.



Percent correct



DeWitt

Anamosa

Independence

All Sites

- Increases with lead time in Independence!
- Score is high with no specific trend.
- Looks good, but is it informative?

Probability of detection

		Observations	
		Yes	No
Forecasts	Yes	A	B
	No	C	D

- (A/(A+C))
  - Proportion of occurrences that were correctly forecasted
  - Probability that event was forecasted given that event was observed
  - AKA probability of detection

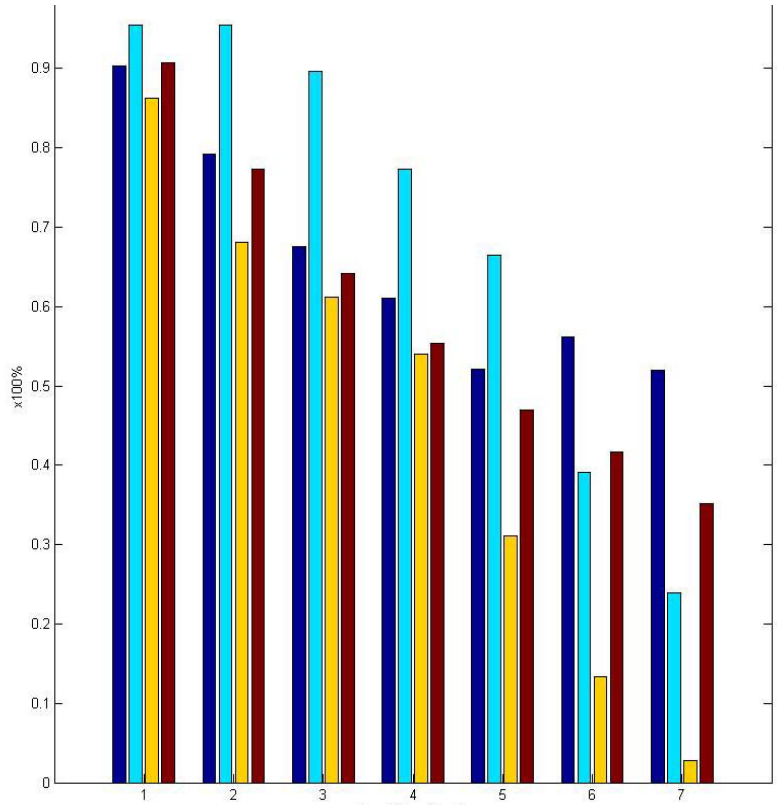
Critical Success Index

		Observations	
		Yes	No
Forecasts	Yes	A	B
	No	C	D

- CSI (A/(A+B+C))
  - Often used for rare events
  - Conditioned probability of a hit given that the event was either forecasted or observed, or both
  - takes into account the positive and negative occurrences
  - does not consider forecasts of non-occurrence

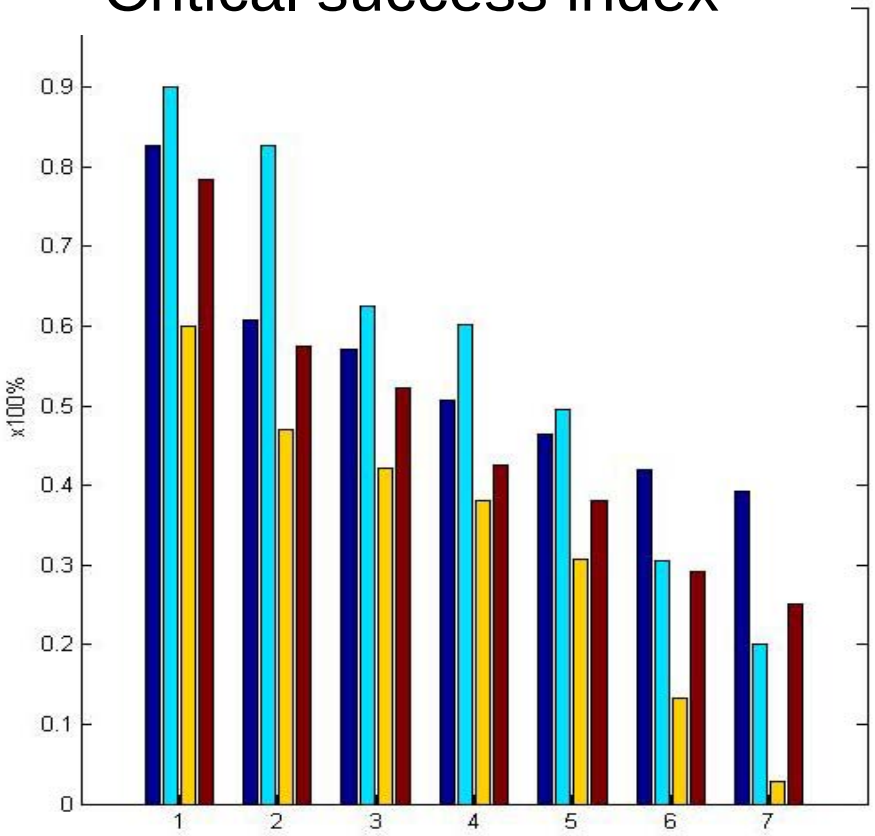
DeWitt
Anamosa
Independence
All Sites

Probability of detection



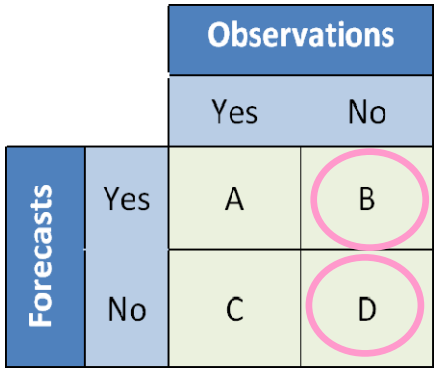
Lead time (days)

Critical success index



Lead time (days)

POFD & ROC Curves

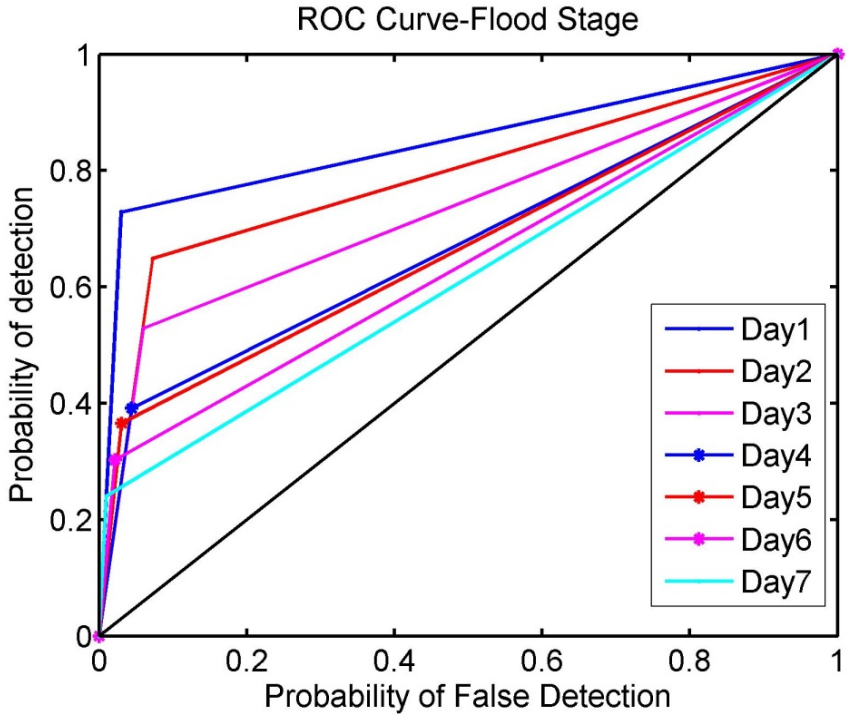


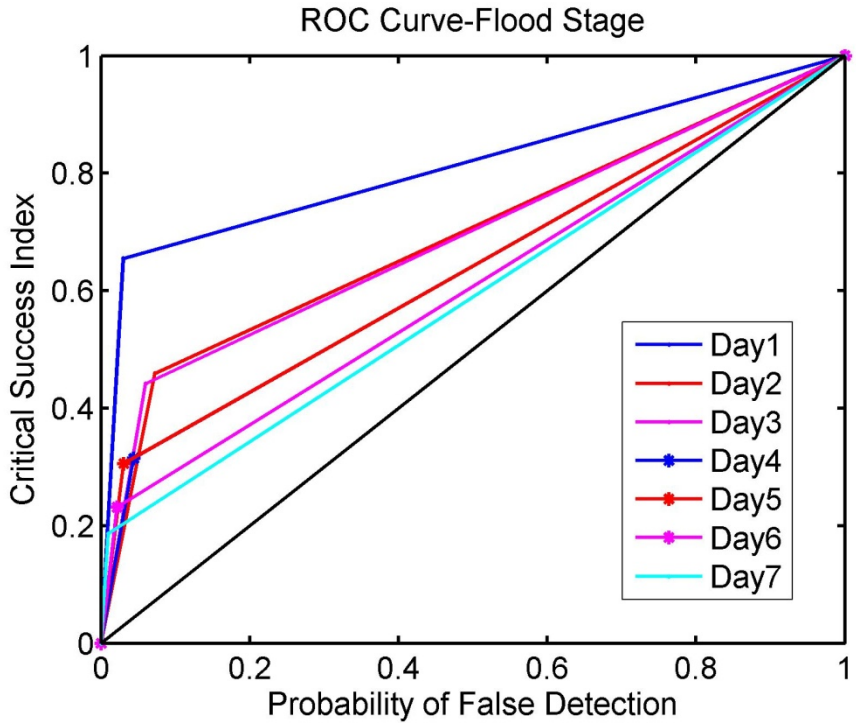
- Probability of false detection (B/(B+D))
- ROC Curves- 'Receiver Operating Characteristic'
  - Signifies the ability to accurately predict an event.
  - Often used with multi-valued forecasts
  - Commonly a plot of Probability of detection vs. Probability of false detection (false alarm rate)

    OR Critical Success Index vs. Probability of false detection
- Same information as bar graphs, but are easier to view

ROC Curves

Low probability of false detection for flood stages
Skill decreases with lead time
CSI indicates larger decrease in skill from day 1 to days 2 and 3;
Nash Sutcliffe also showed a large decline in skill after day 1





Other measures

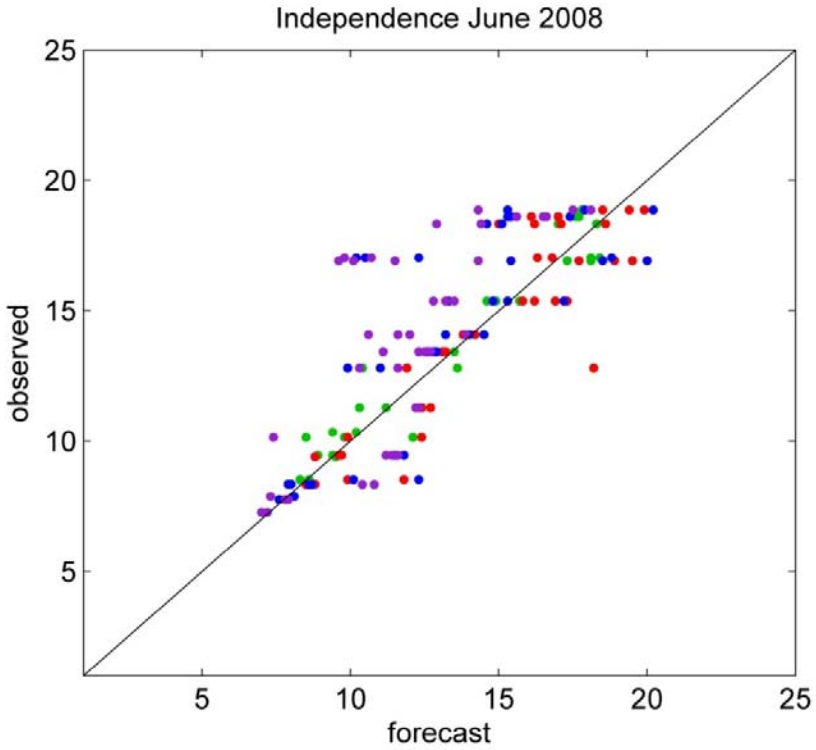
- Categorical Bias ((A+B)/(A+C))
- Gilbert skill score
- Skill scores against persistence
  - As in the accuracy statistics, we are finding that the Mean Absolute Error and Nash-Sutcliffe provide the best/concise assessment.
- Scatter plot/Joint distribution (next slide)

Joint distribution of forecasts & observations:

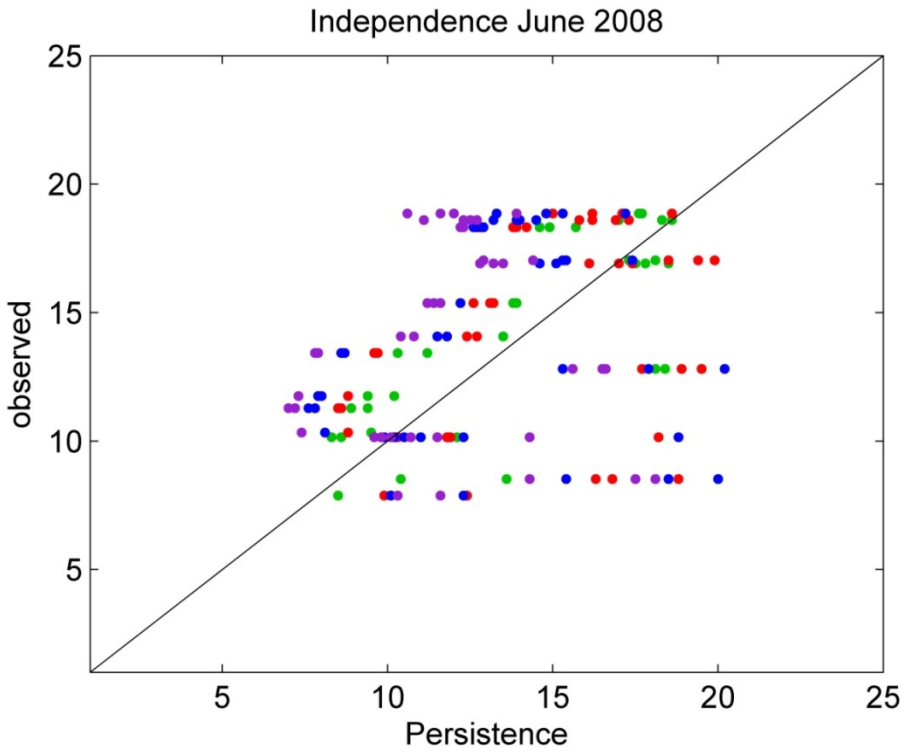
Independence June 2008

- Day 1
- Day 2
- Day 3
- Day 4

NCRFC forecast



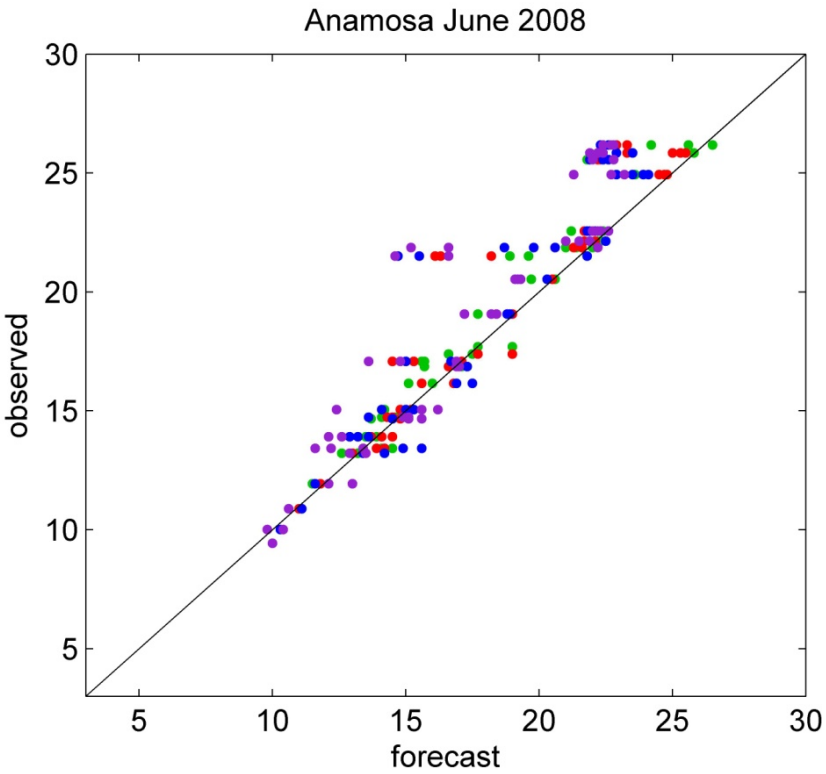
Persistence



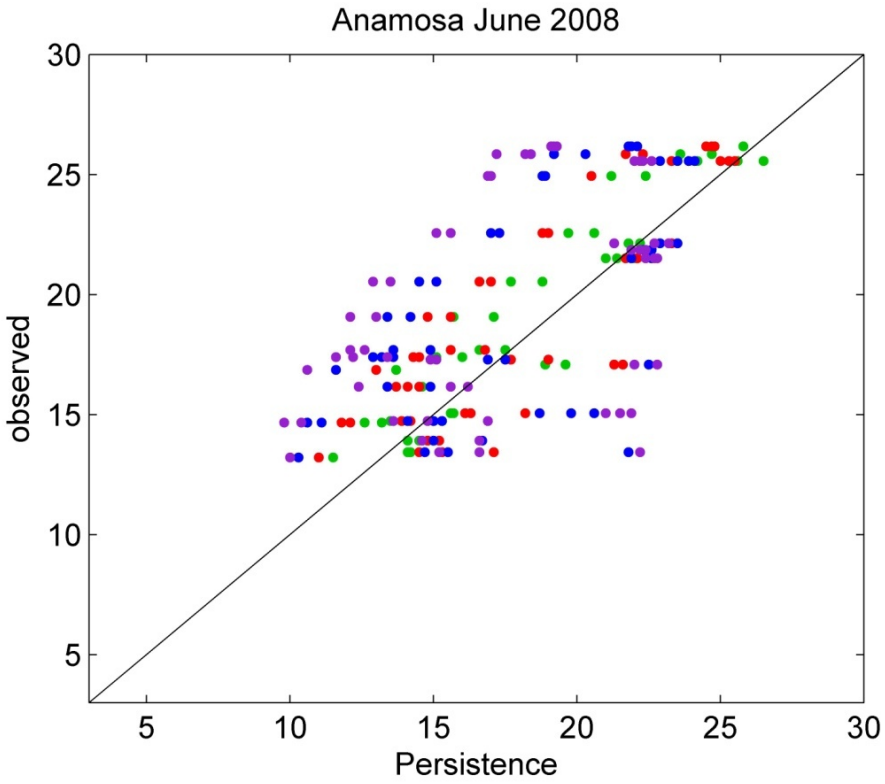
Joint distribution of forecasts & observations: Animosa June 2008

- Day 1
- Day 2
- Day 3
- Day 4

NCRFC forecast



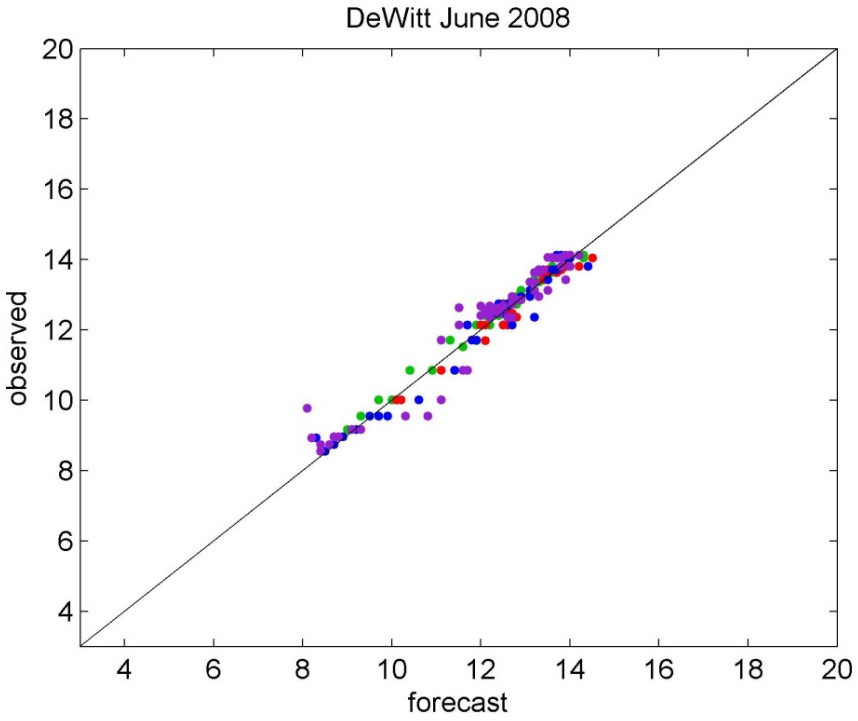
Persistence



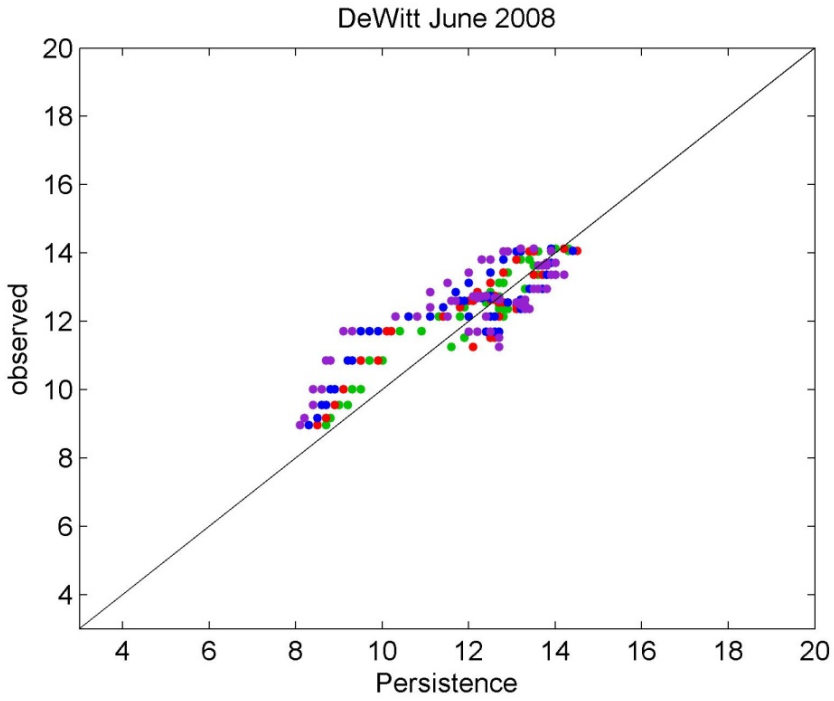
Joint distribution of forecasts & observations: DeWitt June 2008

- Day 1
- Day 2
- Day 3
- Day 4

NCRFC forecast



Persistence



Preliminary Assessment

- Nash-Sutcliffe and Mean Absolute Error found to be useful thus far in providing basic view of skill
  - Others include bias, correlation
- The Critical Success Index may be more consistent with other measures compared to probability of detection.
- As might be expected, forecasts are better than persistence, but difference decreases with increased lead time.
- Forecast performance improves downstream in Wapsipinicon
  - More samples?
  - More information?
  - Better modeling?
  - Scale limitations?

Concluding remarks

- Forecasts are not always issued on a regular basis
  - Some are as needed
  - During flooding may be several per day
  - How to we combine these samples to get a proper and fair regional assessment and comparison?
- Incomplete archives limited the type of analysis that could be done
  - Mismatch in timestep of observation and forecast
  - No analysis of time to peak
  - Calendar day-based analysis rather arbitrary, but USGS historical data available as daily values
  - **Need to archive data at same resolution as forecast when possible.**

Concluding remarks

- Statistics –What is a good value for error?
  - Comparison across sites
  - How do we normalize data, particularly within categories with different ranges?
  - Need to expand current verification data set to compare
- How to best display the statistics?
- How many stage or discharge categories do we need?
- Separating model from forecast/input initial condition error

Future Work

- Des Moines River (1,421 files)
  - Reservoir inflow?
- Cedar River (2,338 files)
- ISU Mesonet is now archiving real-time USGS streamflow data for Iowa
- Assess improvements in forecasting skill from year to year.
- Additional metric evaluation and development