NWS Hydrology Forecast Verification Team: 5th Meeting

04/22/2008 – 1 pm EST

Outline

- Review IVP Exercise
- WGRFC Verification Case Study
- Next Meetings

IVP Exercise: Review

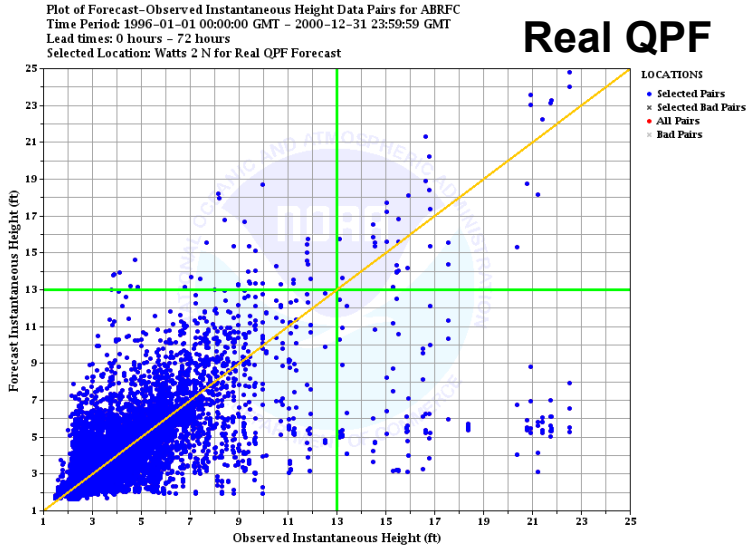
- 6-hr stage hindcasts for 3 lead days available for 4 years for Watts Basin at ABRFC

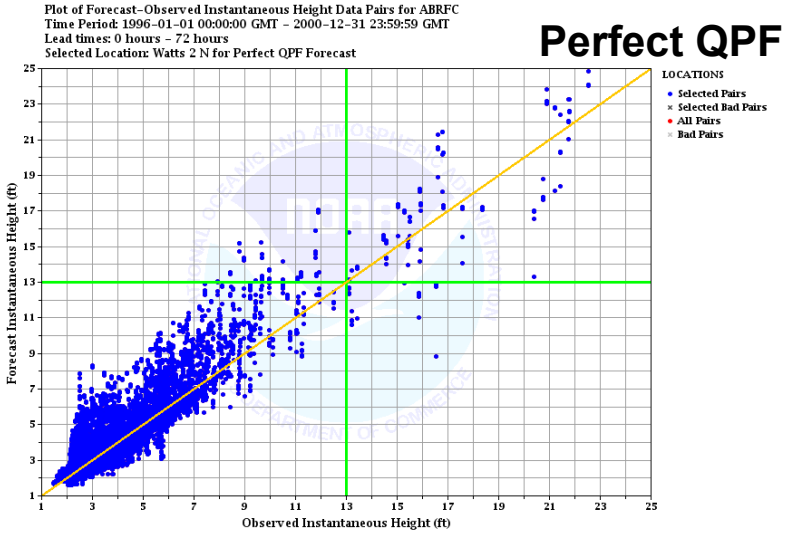
- 4 types of stage forecasts:
  - Zero QPF (XJ),
  - Real QPF (zero after 24 hours) (XK),
  - Perfect QPF (XL),
  - Persistence (reference forecast) (FR)

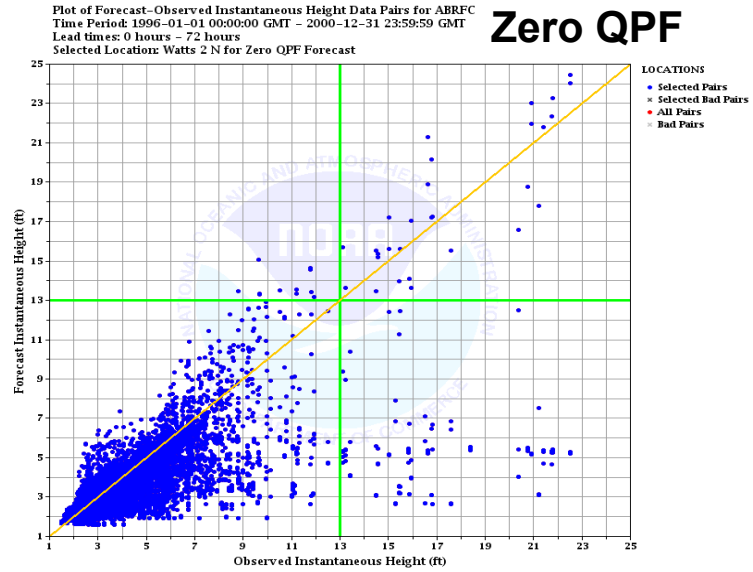
- Exercise goals:
  - Inter-compare the quality of the 4 sets of forecasts
  - Evaluate impact of QPF error and hydrologic error on stage
  - Gain expertise with IVP

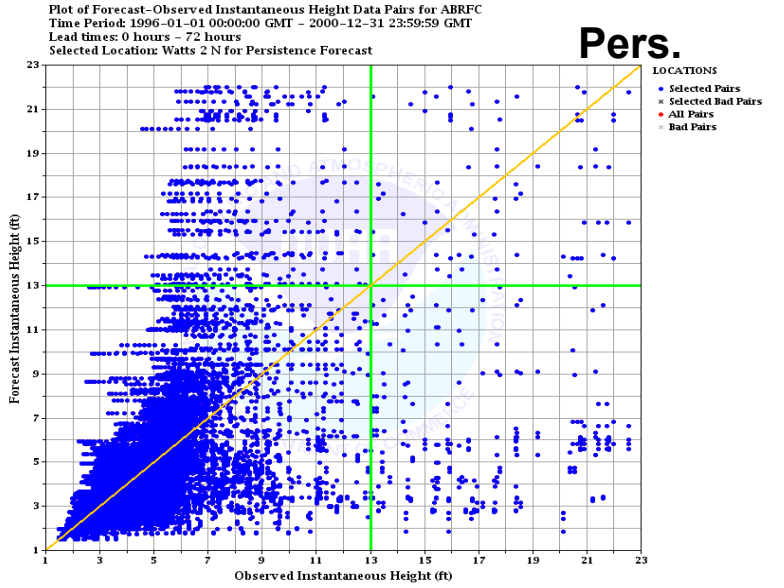
Scatter Plots #1-4: for each forecast type (Zero, Real, Perfect, Persistence) and 3 lead days

Analyze the spread for each of the 4 forecast types. Why is the spread so large for most forecast types? Why is the correspondence between observations and forecasts higher for perfect QPF and lower for persistence forecast? Is there a tendency for over- or under-forecasting?





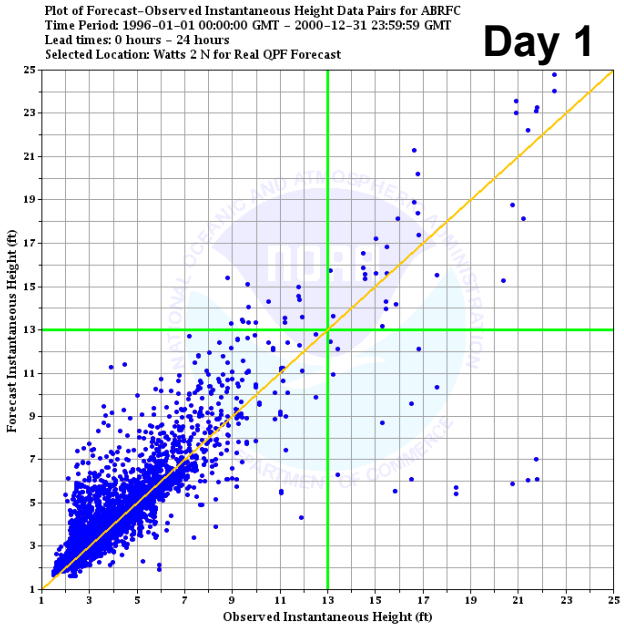


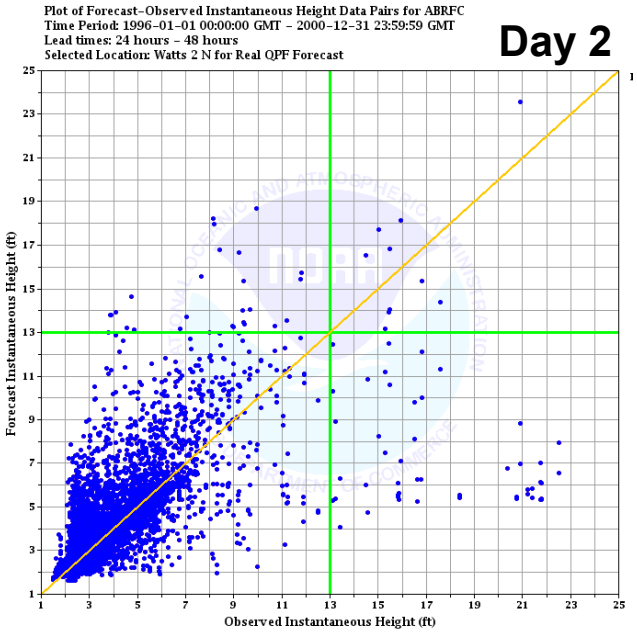


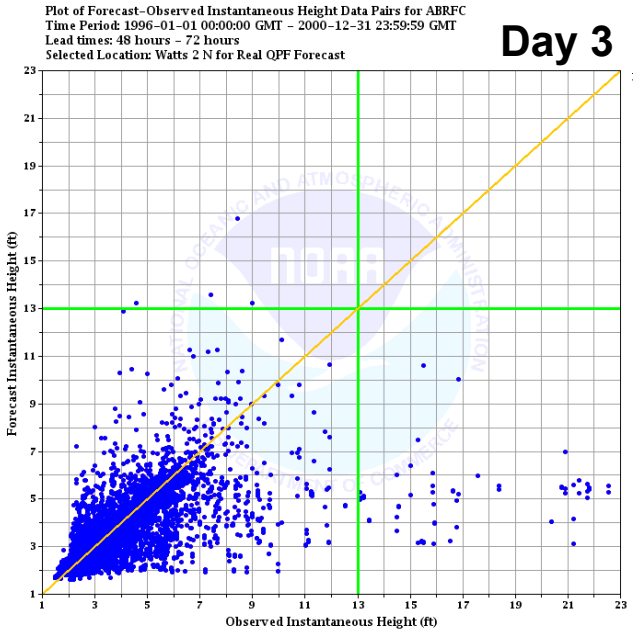
Scatter Plots #5-7 for Real QPF Forecasts for each lead day

Compare the spread for the 3 lead days and give a few characteristics for these Real QPF Forecasts for each lead day. Is there a tendency for low events to over- or under-forecast? What about high events?

Real QPF (Zero after 24 hrs)





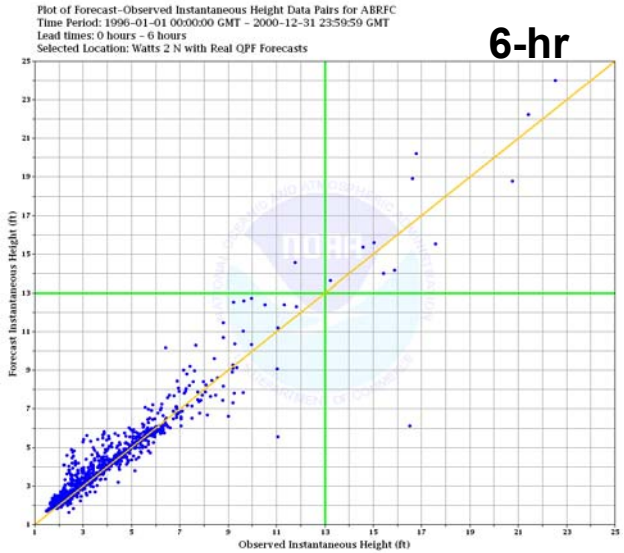


Note: the 6-hr forecasts for the four 6-hr lead times relative to a given lead day are all pooled together to display the forecast-observed pairs on this plot and compute verification statistics for each lead day, as shown in the other plots.

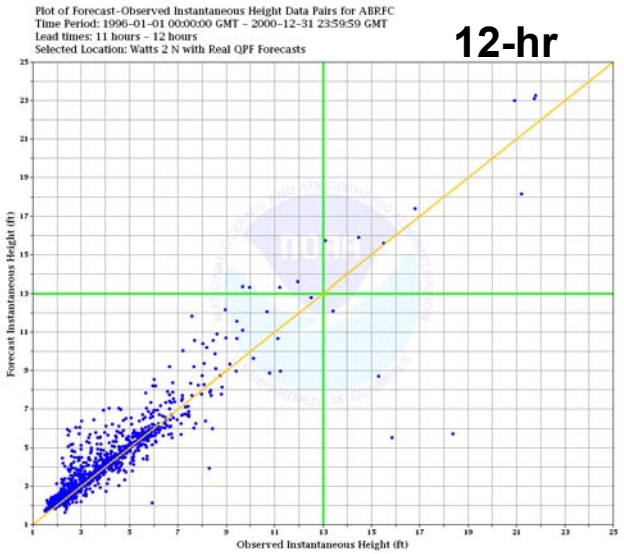
Additional scatter plots for Real QPF Forecasts for each individual 6-hr lead time

Note: the spread varies a lot between individual 6-hr lead times, for low events and more especially for high events. It would be better not to pool together 6-hr forecasts from different lead times when computing verification statistics for these Real QPF Forecasts.

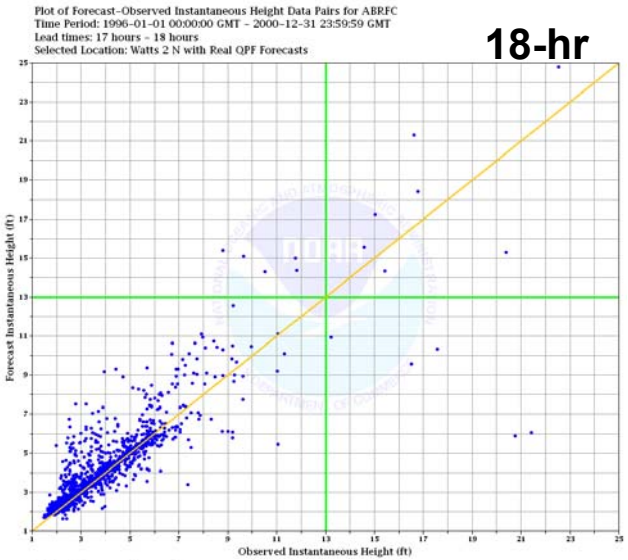
Real QPF (Zero after 24 hrs)



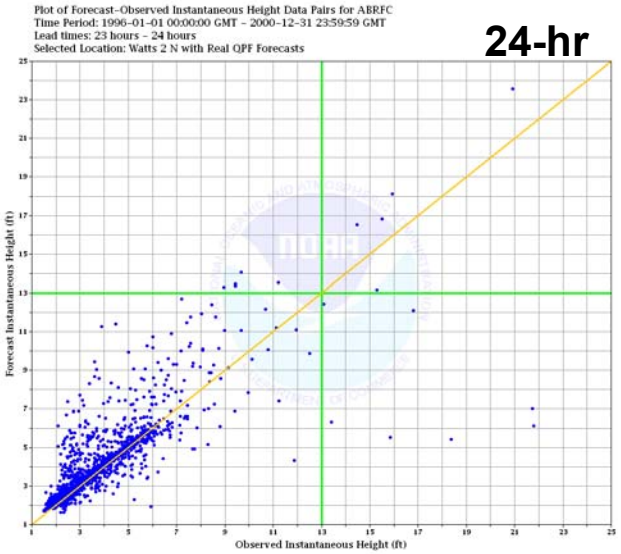
6-hr



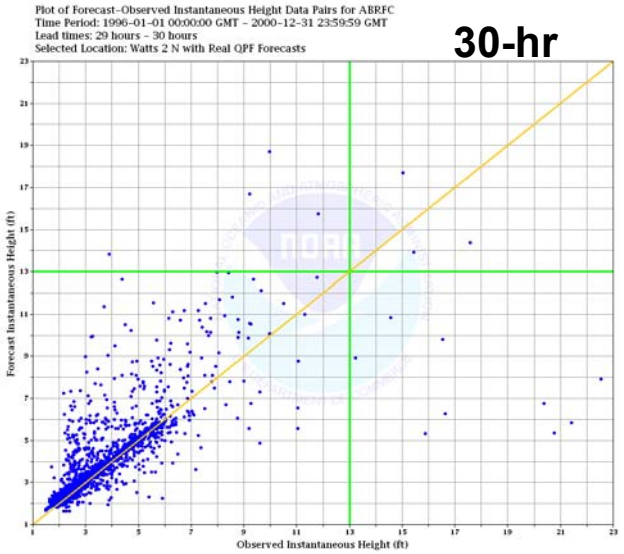
12-hr



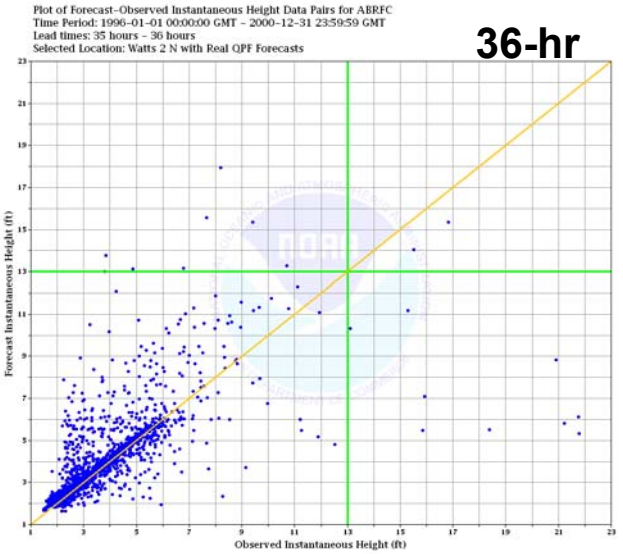
18-hr



24-hr



30-hr



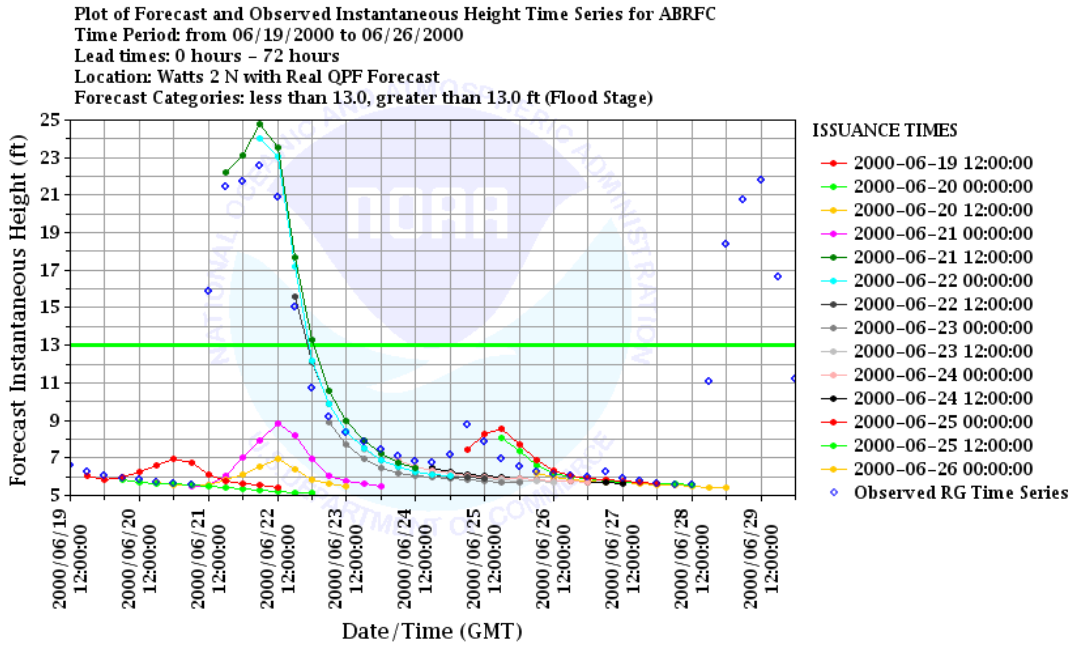
36-hr

Time Series Plots #8-9 for Real QPF Forecasts for a given event

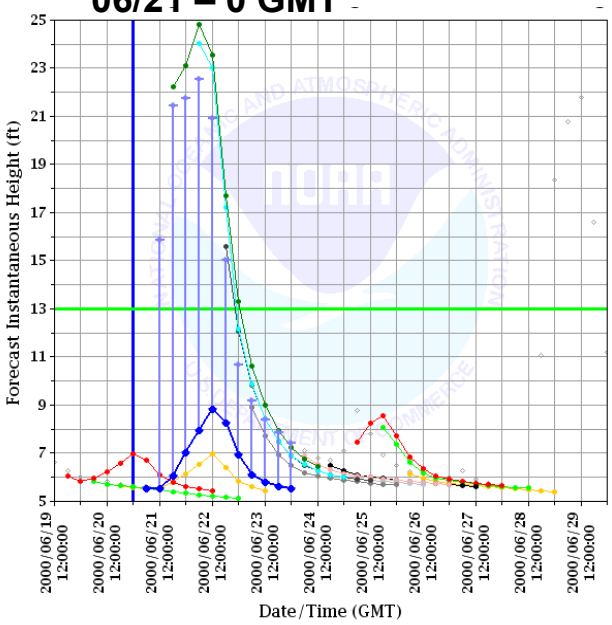
Analyze how the forecast errors (including timing error) vary with the issuance time for this specific flood event. Can you guess what run-time MODS were made during the event?

Real QPF

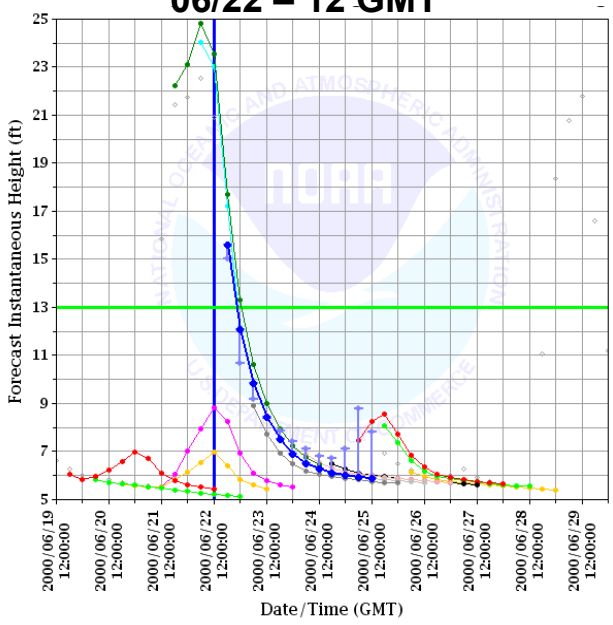
Event of 06/22/2000



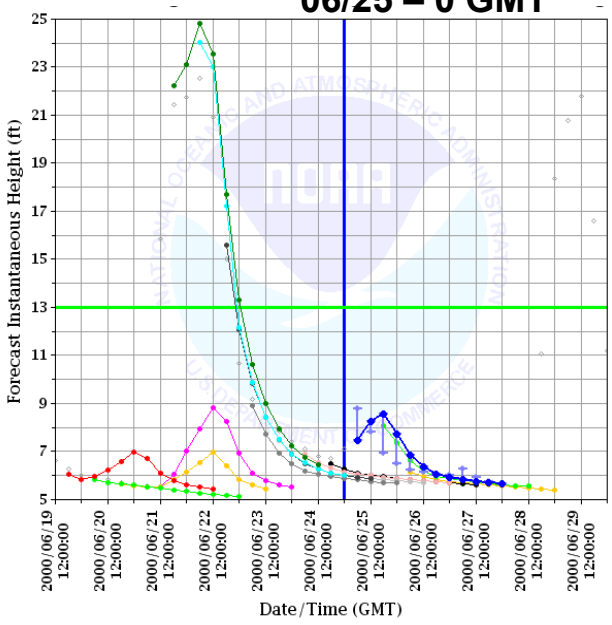
06/21 – 0 GMT



06/22 – 12 GMT

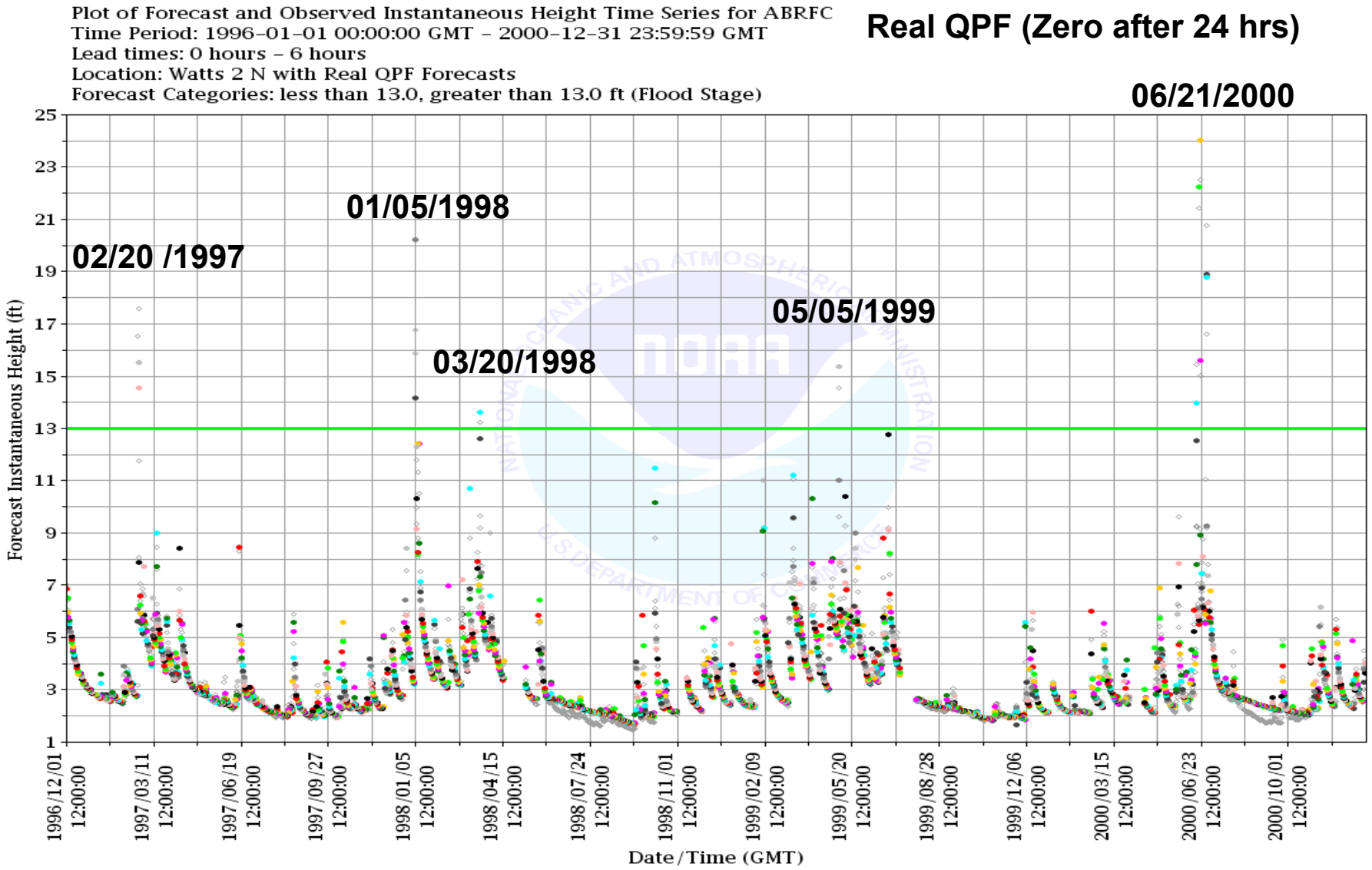


06/25 – 0 GMT



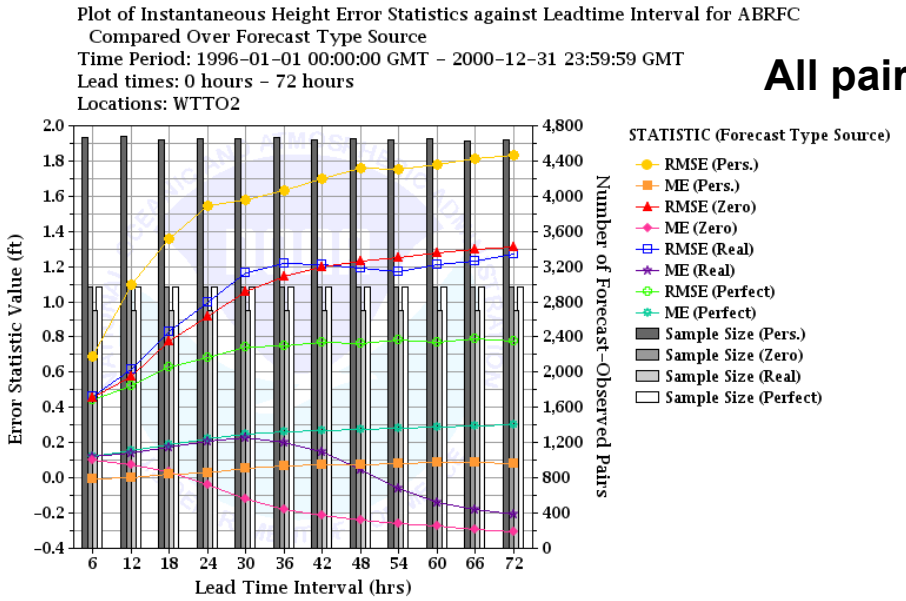
Additional time series plot for Real QPF Forecasts for the whole verification period

Note: since there are only 5 flood events in the whole period, the statistics relative to above Flood Stage are not statistically significant. It would be better to compute statistics for the Action Stage or a lower stage threshold to increase the number of high events.

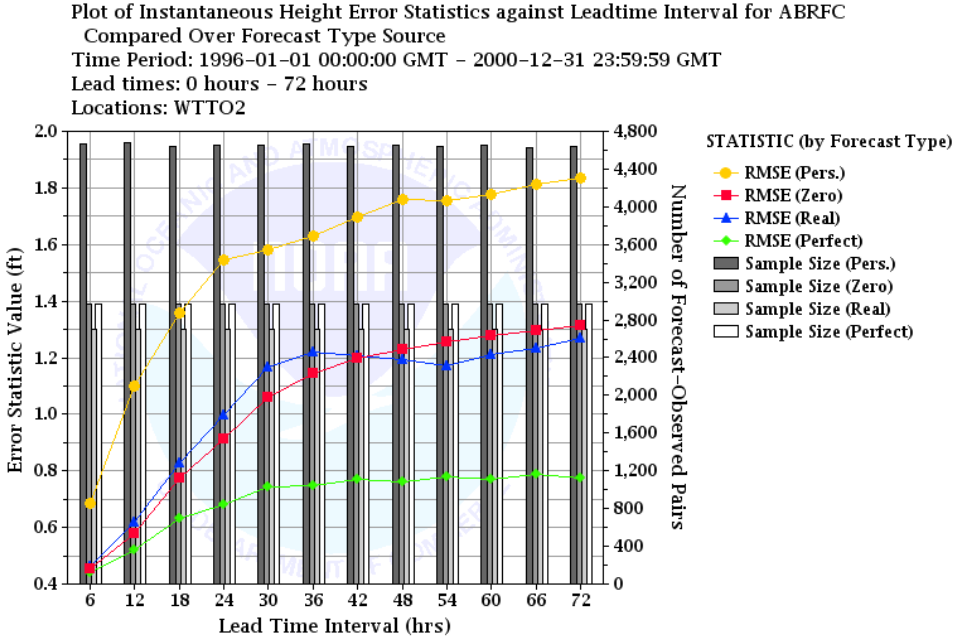


Error Statistics Plot #10 relative to lead time: RMSE, ME and Sample Size for 4 forecast sets

Analyze the variations of RMSE and ME with lead times for the four types of forecasts. Are Real QPF Forecasts more accurate than Zero QPF Forecasts? How different are their additive biases? What can you say about Perfect QPF Forecasts?



Error Statistics Plot #11 relative to lead time: RMSE and Sample Size for 4 forecast sets

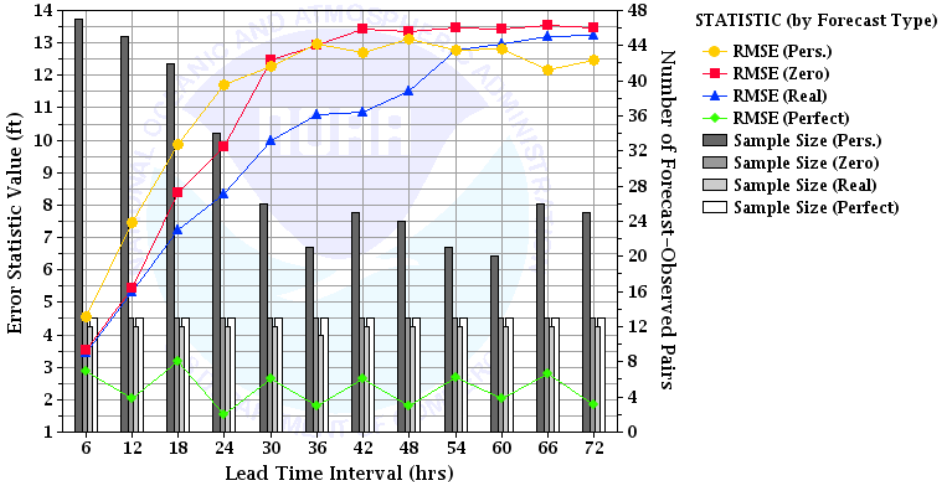


Analyze the variations of RMSE with lead times for all pairs and the subsets of pairs defined with the 4 different conditions. When is the sample size too small for robust verification results?
What do the variations of RMSE for perfect QPF for the Observation ≥ FS and Forecast ≥ FS conditions tell you?

Error Statistics Plots #12-15 relative to lead time and with conditioning : RMSE and Sample Size for 4 forecast sets

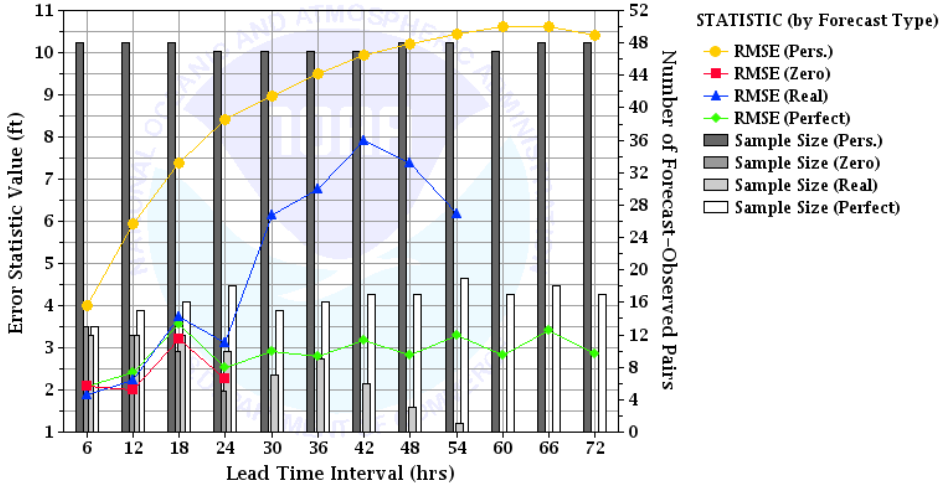
Plot of Instantaneous Height Error Statistics against Leadtime Interval for ABRFC
Compared Over Forecast Type Source
Time Period: 1996-01-01 00:00:00 GMT – 2000-12-31 23:59:59 GMT
Lead times: 0 hours – 72 hours
Observed Category: Obs. above Flood Stage
Locations: WTTO2

Obs ≥ FS



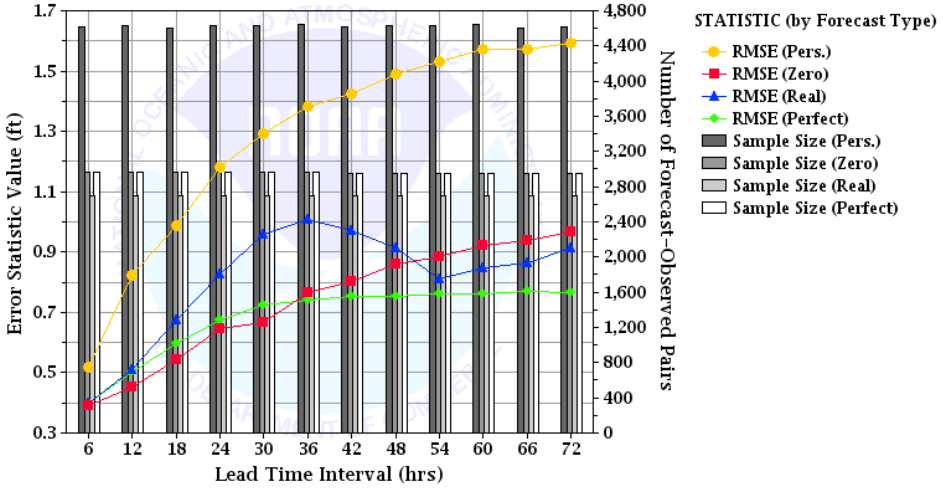
Plot of Instantaneous Height Error Statistics against Leadtime Interval for ABRFC
Compared Over Forecast Type Source
Time Period: 1996-01-01 00:00:00 GMT – 2000-12-31 23:59:59 GMT
Lead times: 0 hours – 72 hours
Forecast Category: Fcst. above Flood Stage
Locations: WTTO2

Fcst ≥ FS



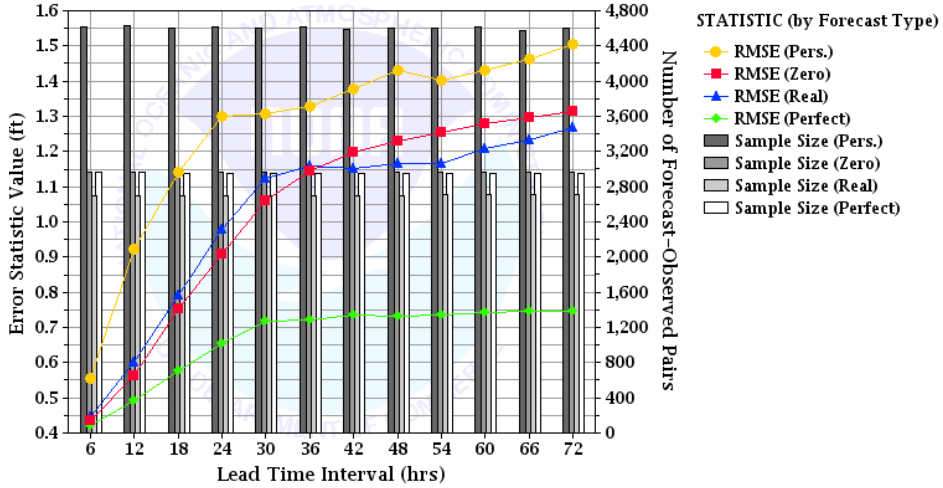
Plot of Instantaneous Height Error Statistics against Leadtime Interval for ABRFC
Compared Over Forecast Type Source
Time Period: 1996-01-01 00:00:00 GMT – 2000-12-31 23:59:59 GMT
Lead times: 0 hours – 72 hours
Observed Category: Obs. below Flood Stage
Locations: WTTO2

Obs < FS

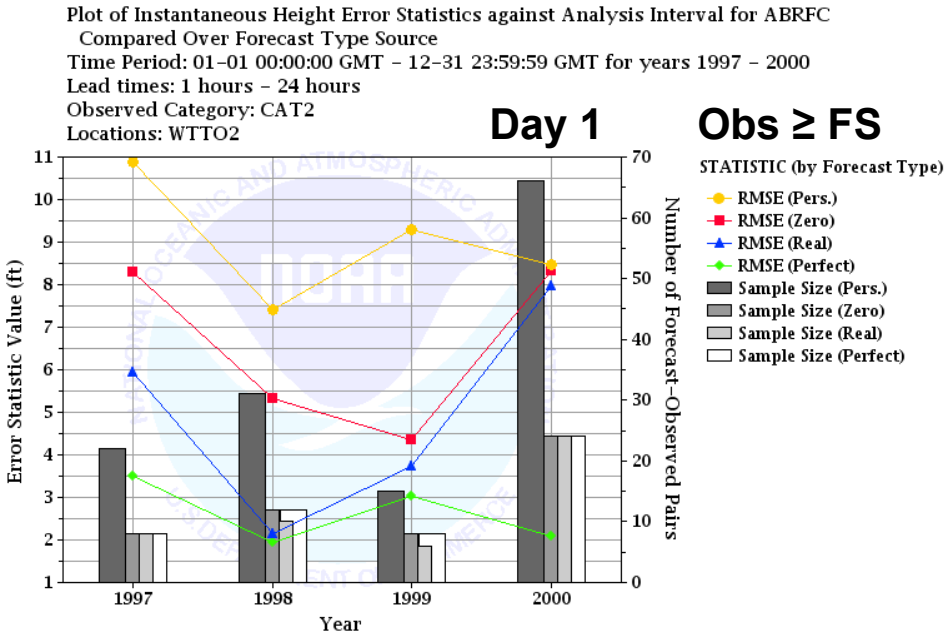


Plot of Instantaneous Height Error Statistics against Leadtime Interval for ABRFC
Compared Over Forecast Type Source
Time Period: 1996-01-01 00:00:00 GMT – 2000-12-31 23:59:59 GMT
Lead times: 0 hours – 72 hours
Forecast Category: Fcst. below Flood Stage
Locations: WTTO2

Fcst < FS



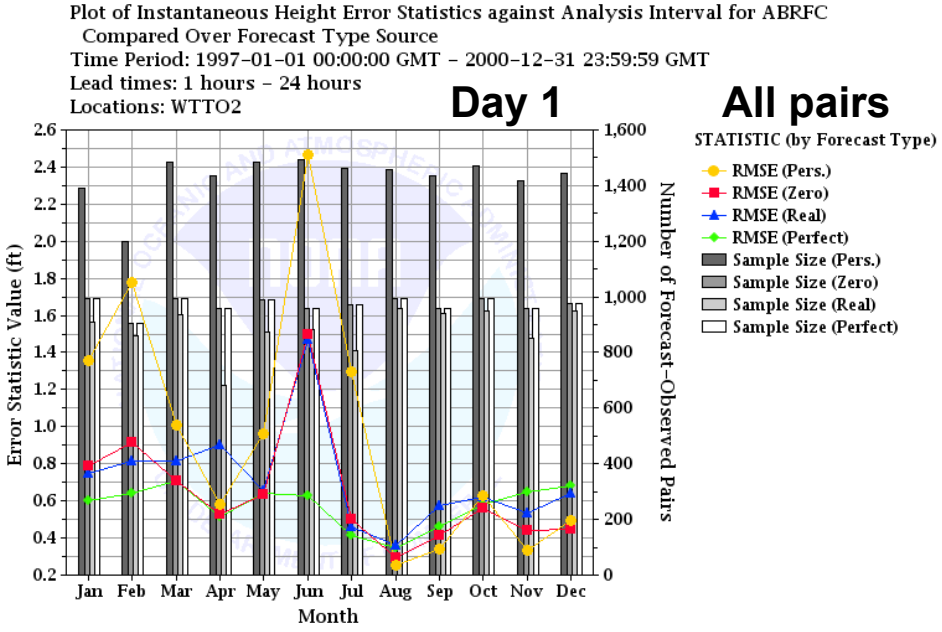
Error Statistics Plot #16 relative to years: RMSE and Sample Size for 4 forecast types



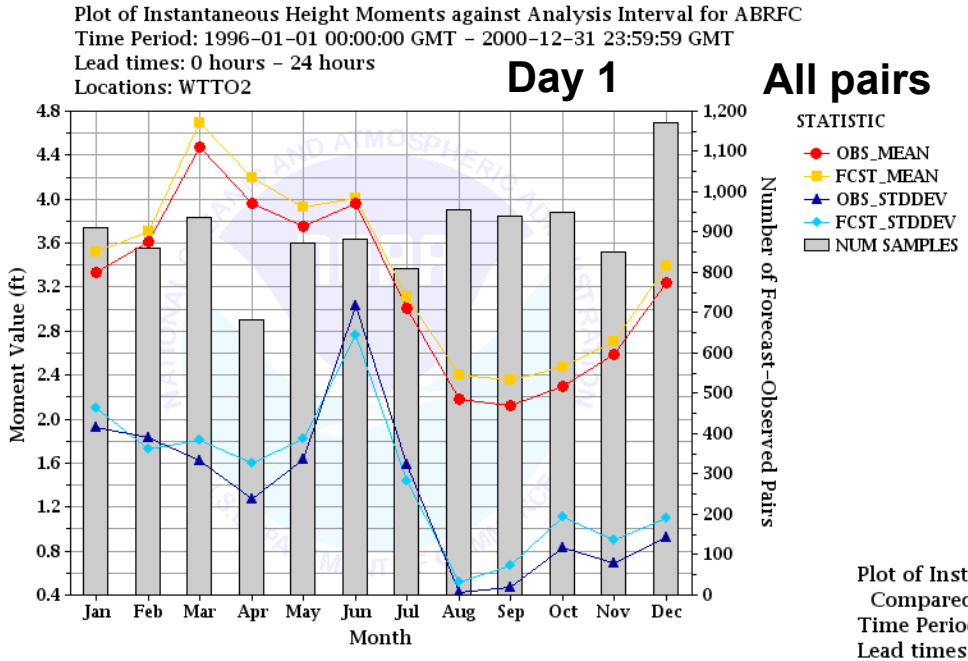
Analyze the variations of RMSE with years for the 4 different forecasts.
Given the sample size, can you say that the QPF have improved the forecasts for specific years for flood events? Can you say that the models have improved over the years for flood events?

Error Statistics Plot #17 relative to months: RMSE and Sample Size for 4 forecast types

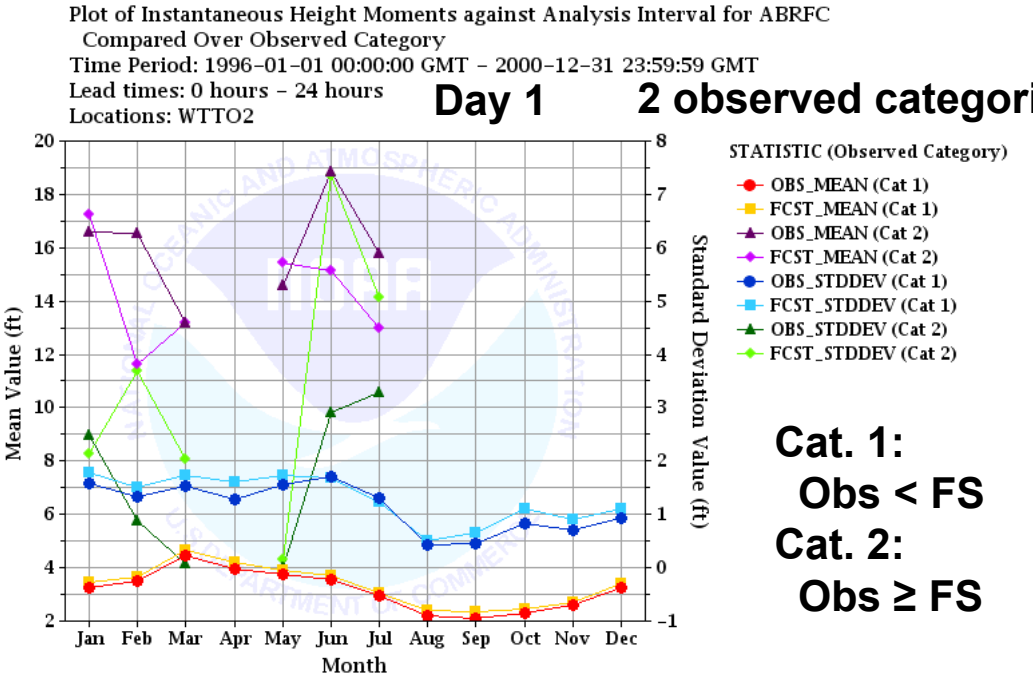
Analyze the variations of RMSE with months for the 4 different forecasts for lead day 1.
By comparing the Zero QPF forecasts and the Real QPF forecasts, when did the Real QPF improve the forecast accuracy?
By analyzing the Perfect QPF forecasts, when did the models perform better and worse?



Moments Plots #18-19 relative to months: Mean and Standard Deviation for observations and forecasts for lead day 1



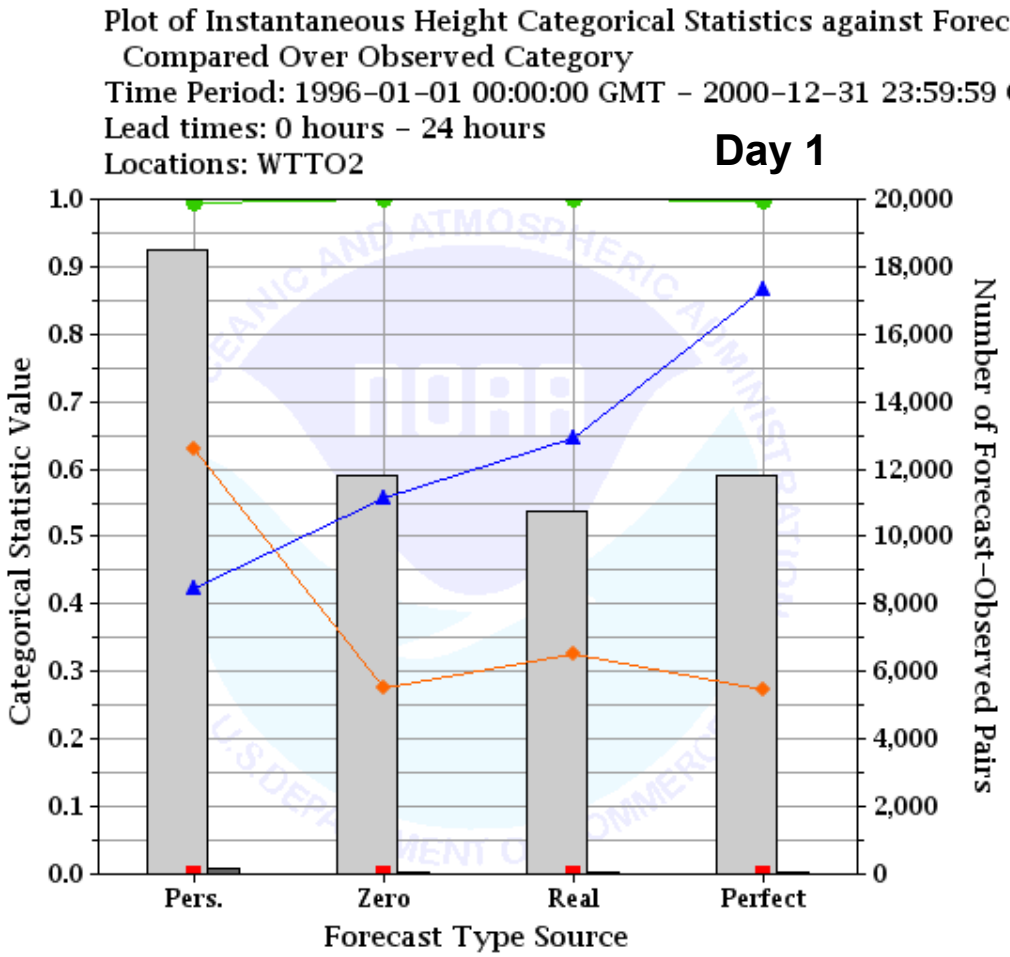
Analyze how well the forecast distributions correspond to the observed distribution for the different months for all pairs and then the subsets of pairs defined with the observed categories.
Are the results similar for the two observed categories? What can you say about the results for flood events given the sample size?



Cat. 1:
Obs < FS
Cat. 2:
Obs ≥ FS

Categorical Statistics Plot #20: POD, HFAR, and Sample Size for Flood Stage Threshold for 4 forecasts types and for lead day 1.

Analyze how the POD and HFAR vary for the four forecast types for lead day 1. Since there are so many pairs in Category 1, the sample sizes for Category 2 are not visible on this plot. What plot could you use to display the sample sizes for Category 2 (observations ≥ Flood Stage) for the four forecast types? What can you say about sample sizes?



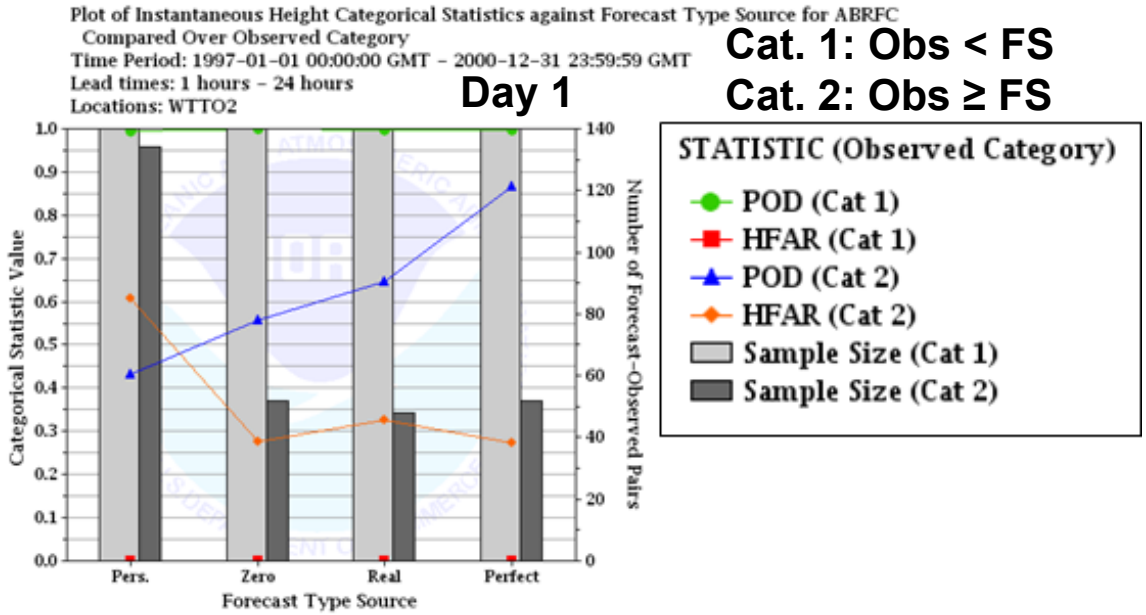
Day 1

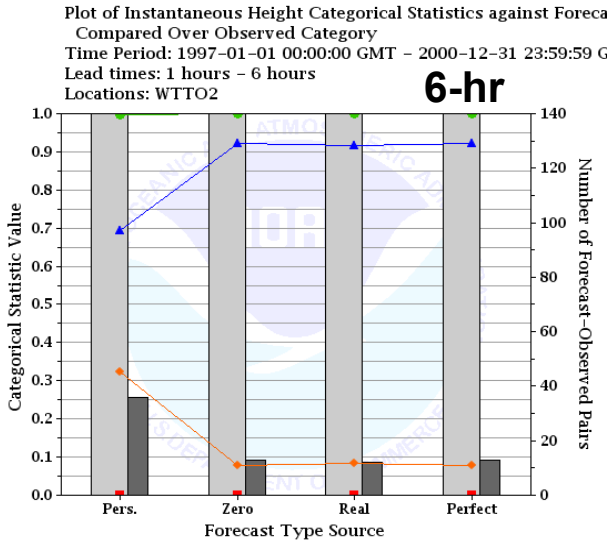
2 observed categories

STATISTIC (Observed Category)
- POD (Cat 1)
- HFAR (Cat 1)
- POD (Cat 2)
- HFAR (Cat 2)
- Sample Size (Cat 1)
- Sample Size (Cat 2)

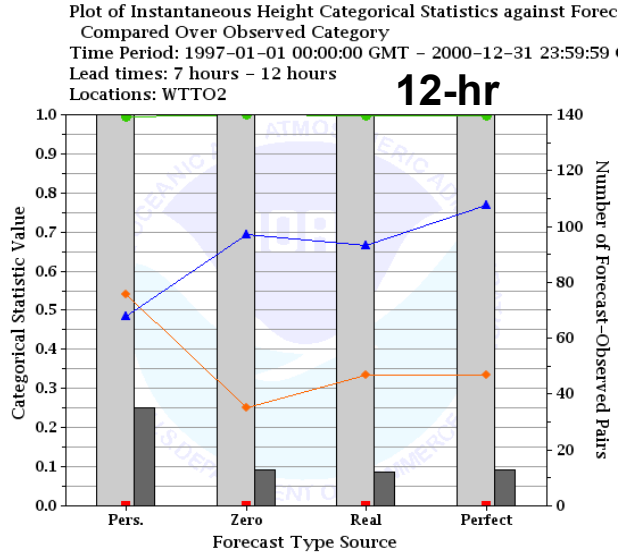
Cat. 1: Obs < FS
Cat. 2: Obs ≥ FS

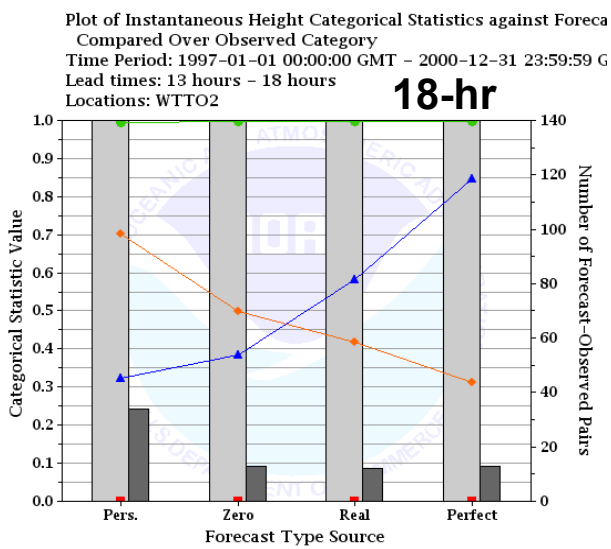
Additional categorical statistics plots: POD, HFAR, and Sample Size for Flood Stage Threshold for 4 forecasts types and for each individual 6-hr lead time

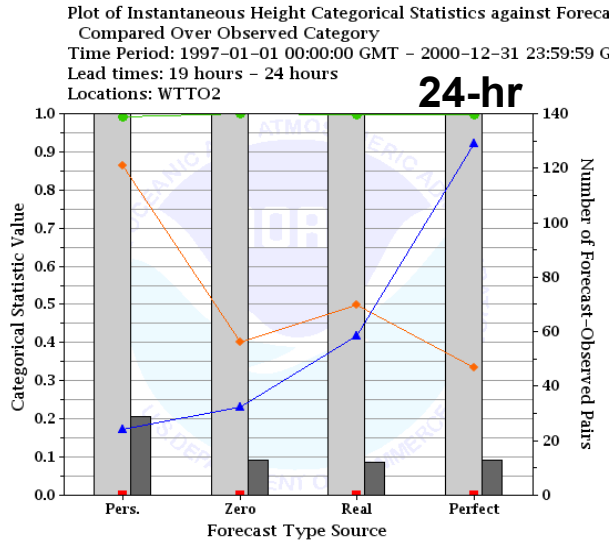




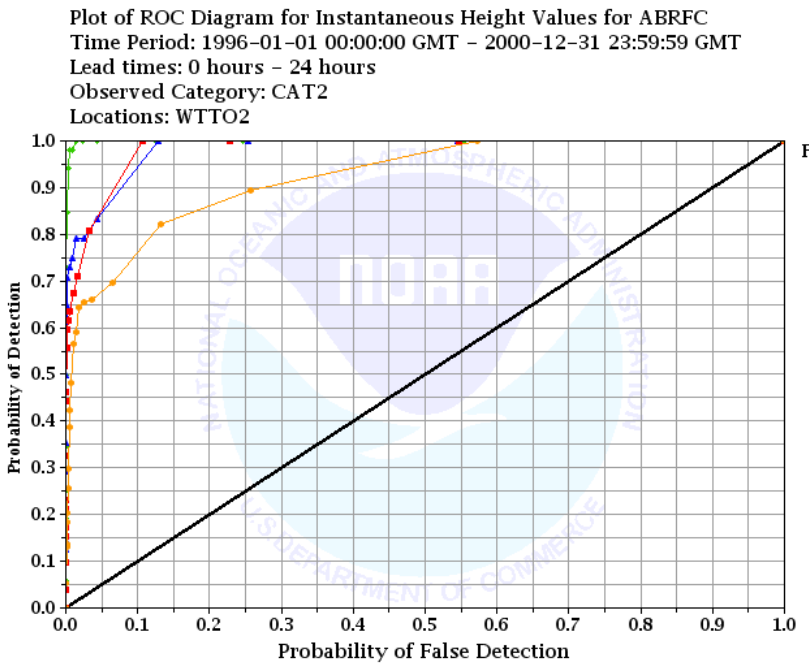
Note: POD and HFAR vary significantly between individual 6-hr lead times. Also the sample sizes are too small to get robust results.





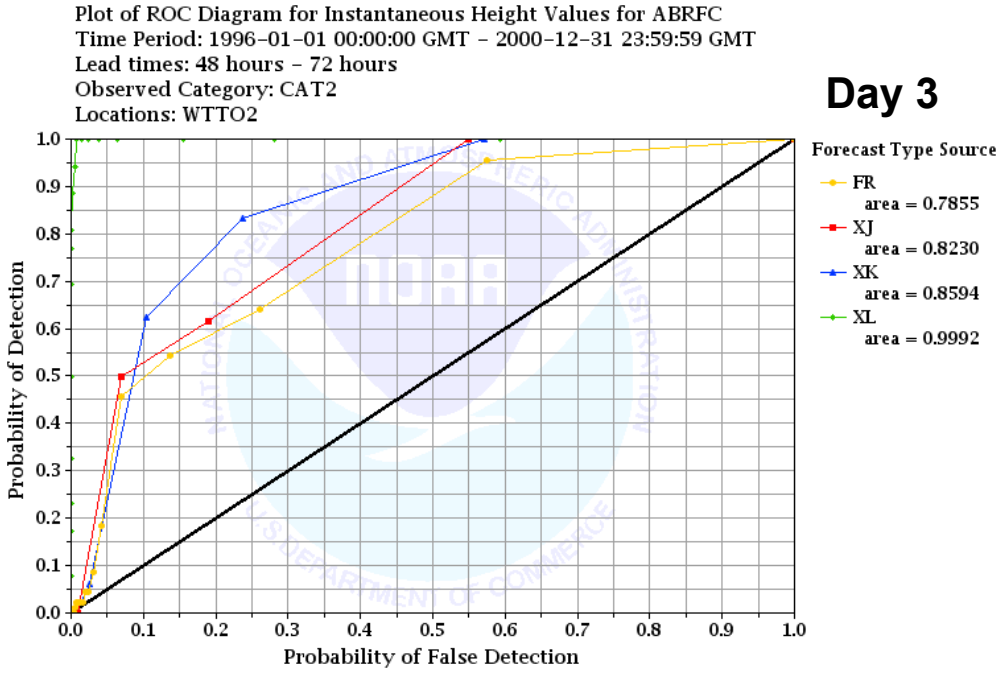


ROC Plots #21-22: ROC curve and ROC Area for 4 forecast types for Obs. ≥ FS



Day 1

Analyze how ROC curves and ROC areas vary with lead days and with forecast types.
Do they vary with lead days for perfect QPF? How about the other forecast types?



Day 3

IVP Exercise: Review

- 6-hr stage hindcasts available for 4 years for 1 single basin: only 5 flood events
  - All verification statistics relative to flood events are based on too few flood events to be statistically significant
  - Tip: to get more high events, define a lower threshold (e.g. action stage), extend period of record, and/or pool forecast-observed pairs from similar forecast points

- Verification statistics relative to 1 lead day are computed by IVP by pooling forecasts from different lead times (e.g., day-1 stats from all forecasts for 6-hr, 12-hr, 18-hr and 24-hr lead times)
  - ***But*** forecast quality varies a lot between individual 6-hr lead times
  - Tip: better to compute verification statistics for individual lead times

WGRFC Verification Case Study

Next meetings

- 6th meeting on 05/05/08 at 1:30 pm EST:
  - Overview of new EVS prototype
  - Present EVS exercise

- 7th meeting to review EVS exercise:
  - potential dates: June 2-5, 9-12

- Next Verification Case studies:
  - June 08: OHRFC CNRFC MARFC
  - July 08: NERFC? NCRFC? NWRFC?
  - August 08: ABRFC MBRFC? SERFC?
  - September 08: LMRFC APRFC

Thank you!