Verification Analytics

Nathan Barber

Tennessee Valley Authority

03/13/2019

Outline

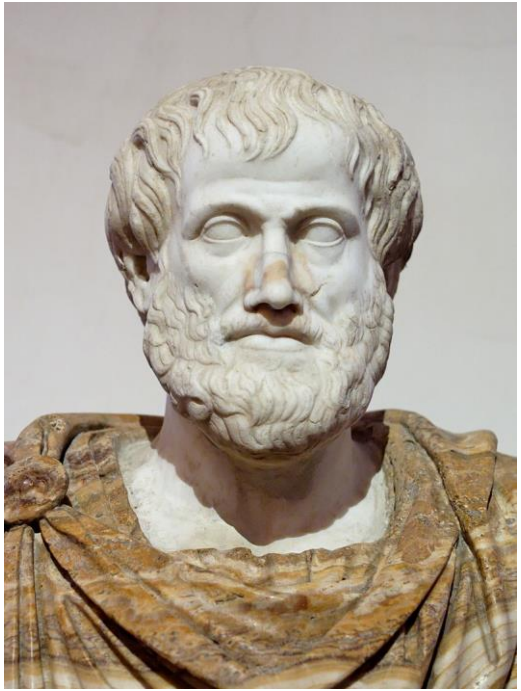
- The value of forecasts at TVA
- Historical practice of forecast verification
- Designing a verification analytics system
- Deliverables
- Next Steps
- Lessons Learned

Proposition: Using forecasts provides value by allowing TVA to make better water management decisions

Alternative Proposition 1: Using forecasts neither improves nor worsens TVA's water management decisions

Alternative Proposition 2: Using forecasts makes TVA's water management decisions worse

Alternative Proposition 3: Using forecasts can either make TVA's water management decisions better or worse.



How do we test?

What will we do differently if the rain shifts south?

What would happen at Great Falls if we didn't account for forecast inflow?

How will we handle an additional inch over Tims Ford reservoir?

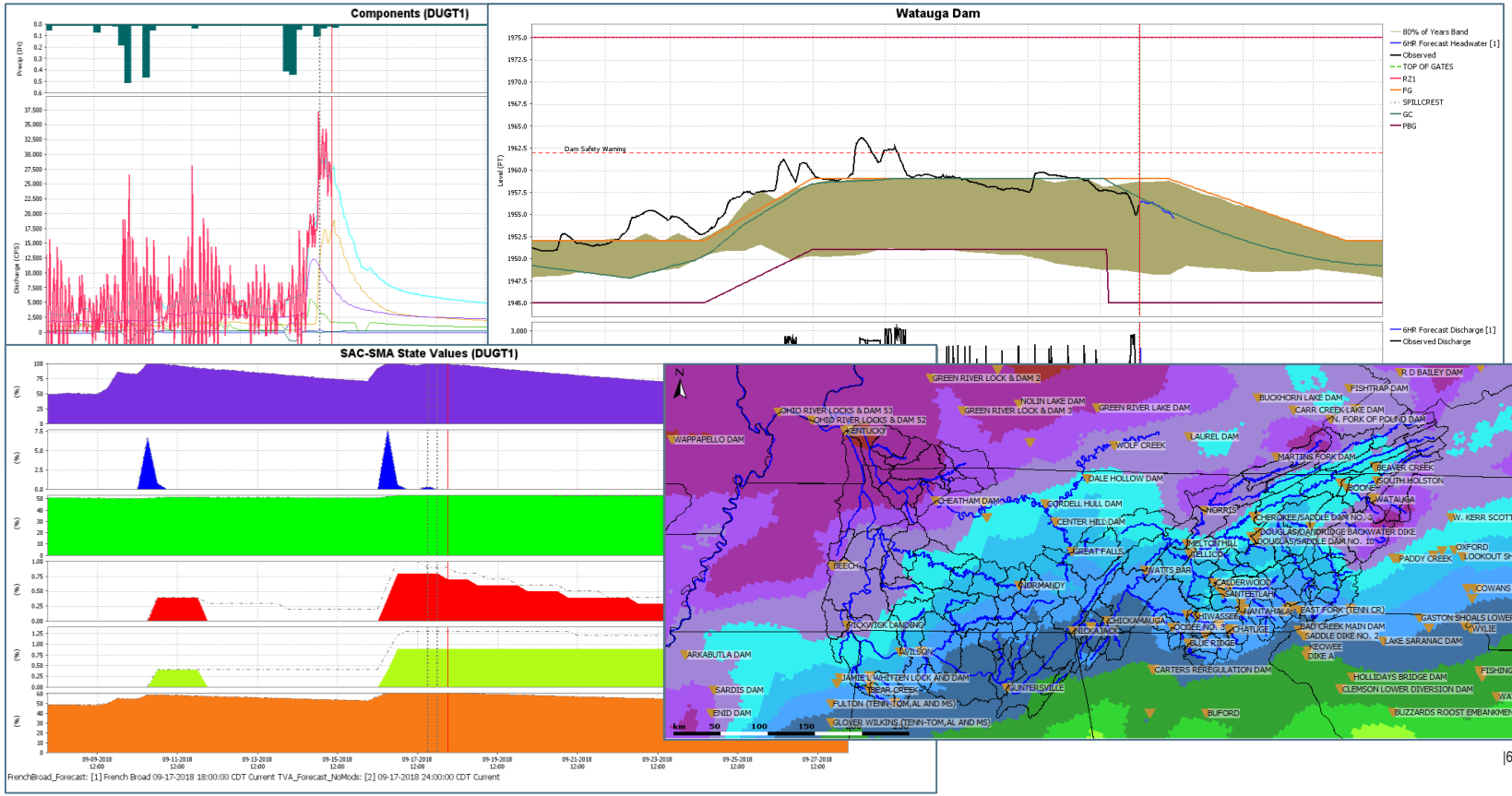
What would happen if we omitted out incremental cost or load forecasting in our optimization models?

What if we only reacted to rises in water temperature and assumed no future knowledge?

Thought Experiments

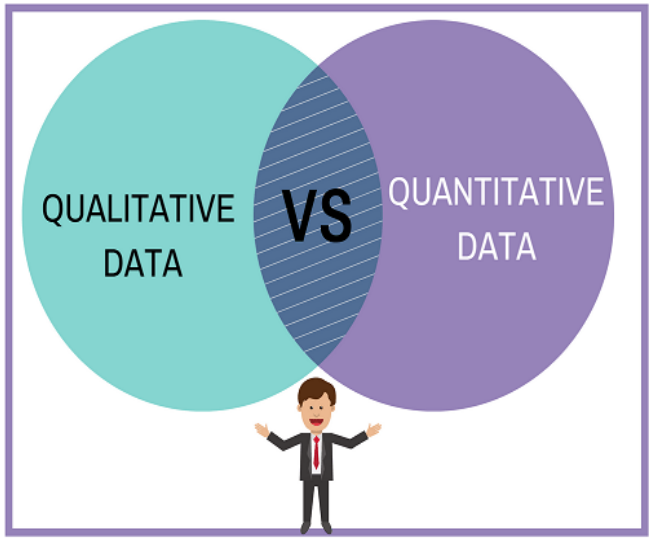
- LMRFC's QPF is generally "better" than WPC's QPF
- Forecasts generally improve as they are issued closer to an event
- Streamflow forecasts have greater skill once the precipitation is on the ground
- Streamflow forecasts tend to underforecast on the rising limb, followed by overforecasting at the peak
- QPF is generally reliable out to 3 days with respect to reservoir operations at TVA

The Value of Forecasts to TVA



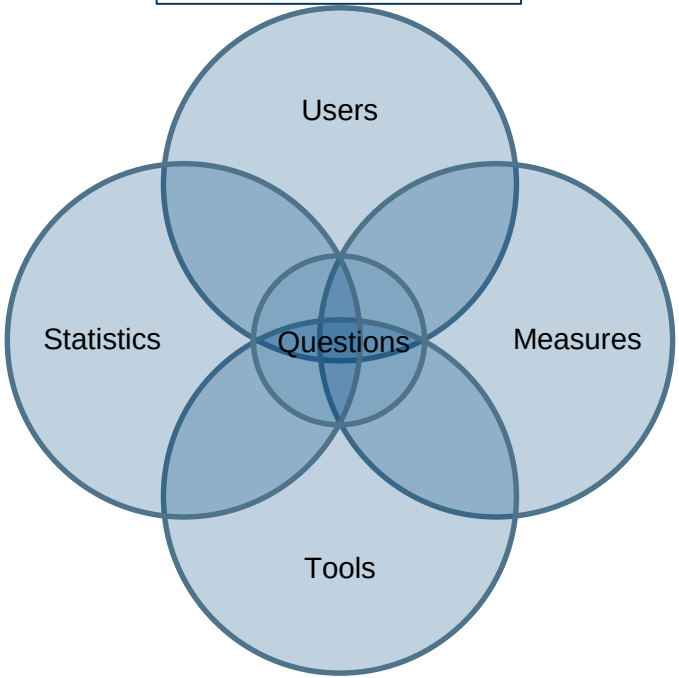
Quantitative Verification

- Validate/Verify expert knowledge
- Magnitude of error/skill
- Confidence Intervals
- Statistical Significance
- Statistically/Scientifically robust
- Objective rather than subjective
- Transform thought experiments into actual experiments
- **Use information to improve forecasts**



Designing a Verification System

Who will be using the system?

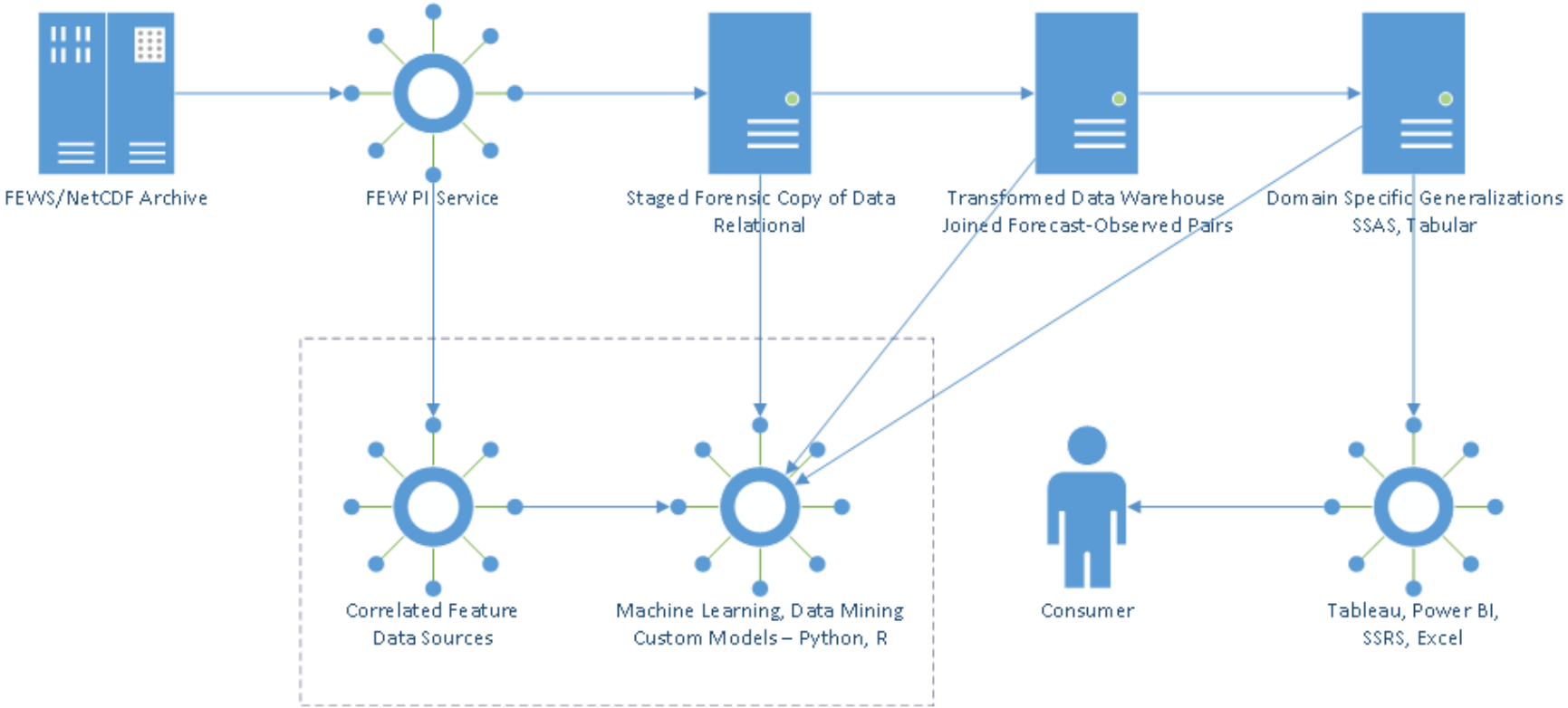


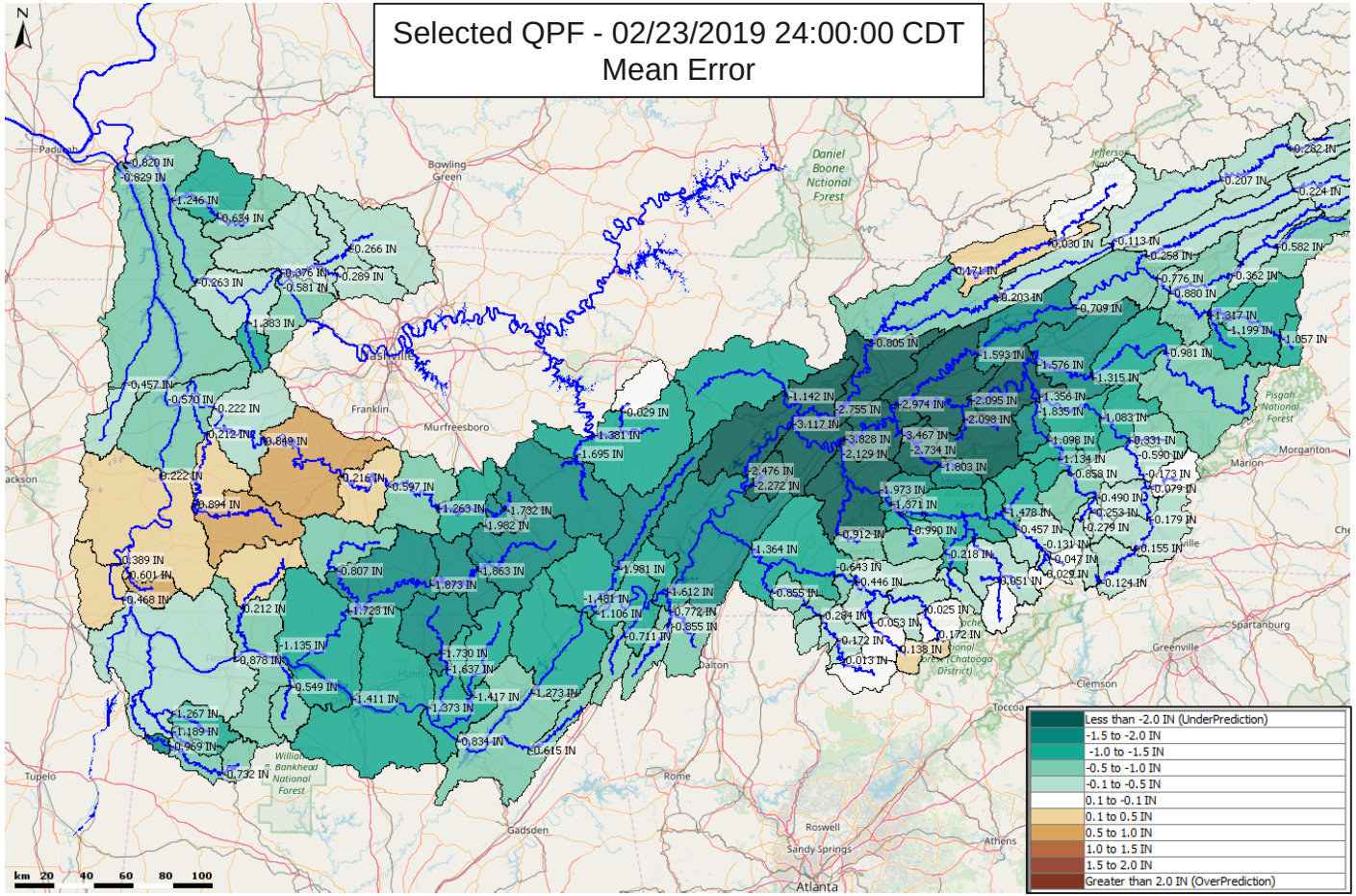
What statistics and scores are commonly employed in operational, scientific agencies?

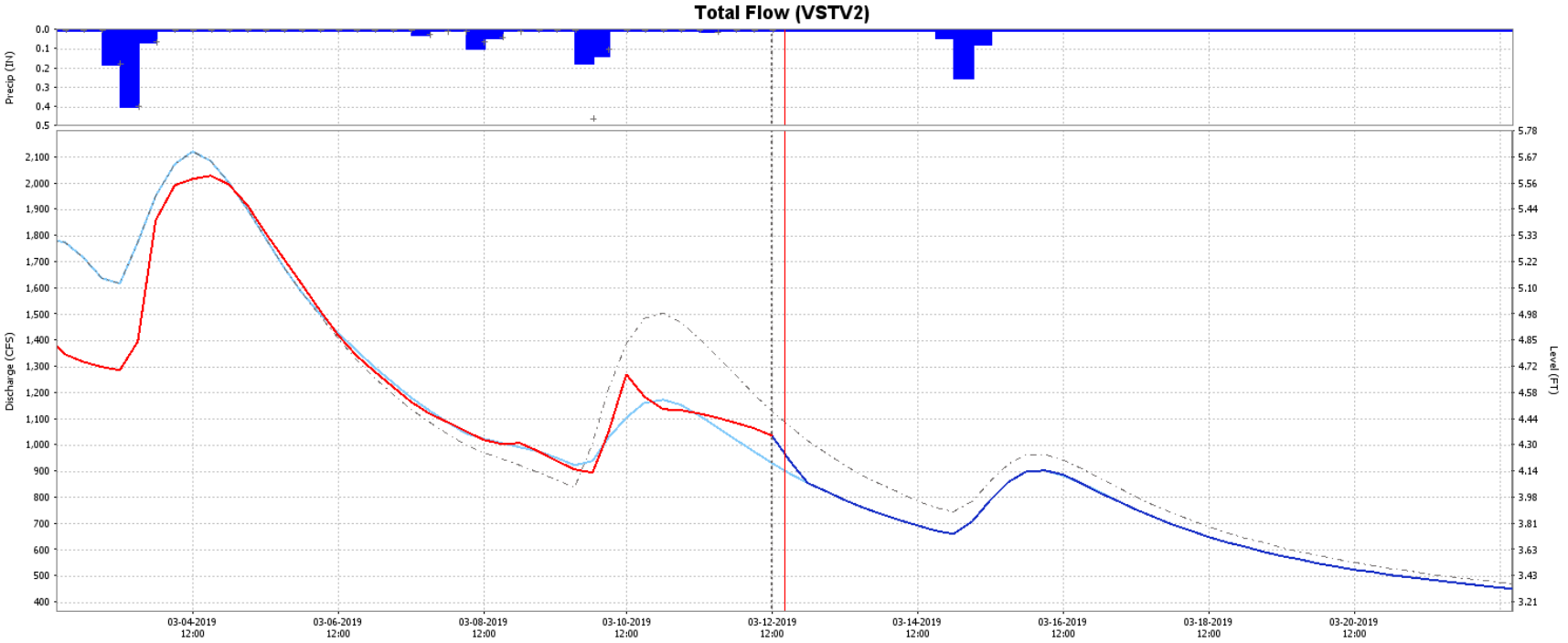
What data do we need to answer the question?

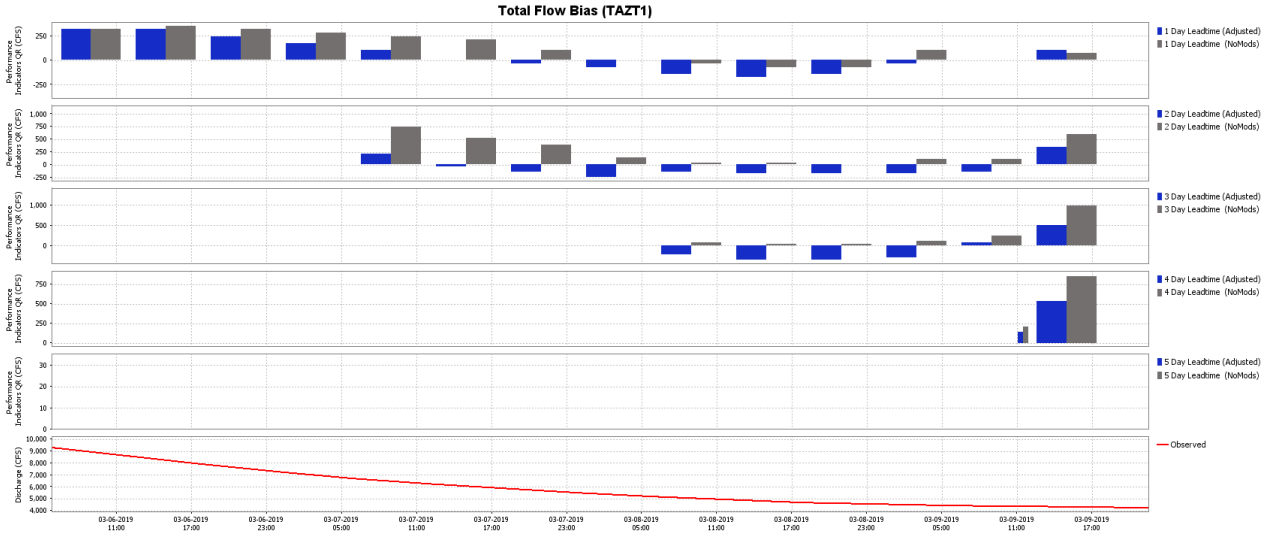
Which tools can provide the greatest ease in accessing the data and solving problems?

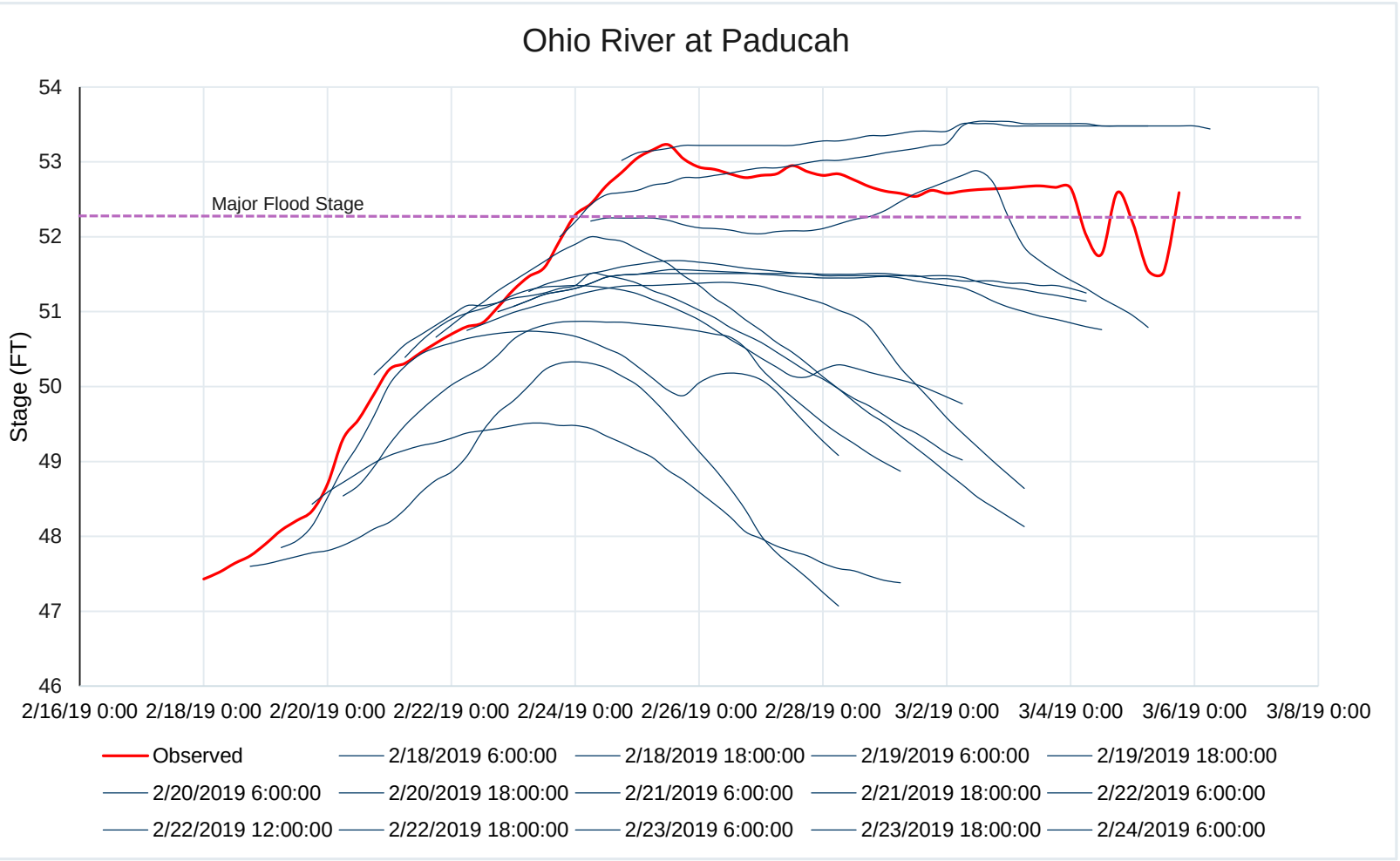
Software Architecture




Selected QPF - 02/23/2019 24:00:00 CDT
Mean Error
N
Less than -2.0 IN (UnderPrediction)
-1.5 to -2.0 IN
-1.0 to -1.5 IN
-0.5 to -1.0 IN
-0.1 to -0.5 IN
0.1 to -0.1 IN
0.1 to 0.5 IN
0.5 to 1.0 IN
1.0 to 1.5 IN
1.5 to 2.0 IN
Greater than 2.0 IN (OverPrediction)
km 20 40 60 80 100
Bowling Green
Nashville
Franklin
Murfreesboro
Daniel Boone National Forest
Pisgah National Forest
Spartanburg
Greenville
Clemson
Rome
Gadsden
Atlanta
Roswell
Sandy Springs
Athens
Tupelo
Marion
Morganton

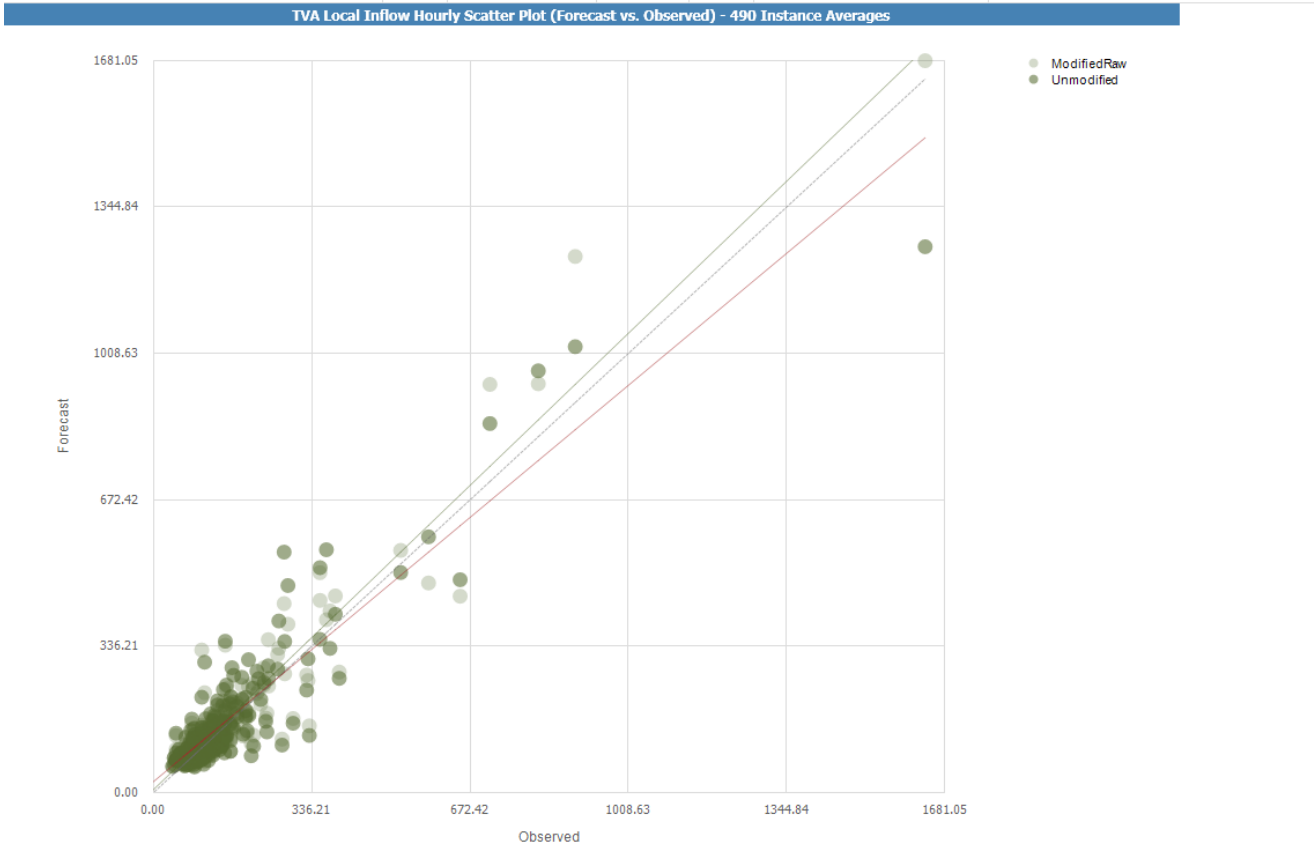

Total Flow (VSTV2)
Precip (IN)
0.0
0.1
0.2
0.3
0.4
0.5
Discharge (CFS)
2,100
2,000
1,900
1,800
1,700
1,600
1,500
1,400
1,300
1,200
1,100
1,000
900
800
700
600
500
400
Level (FT)
5.78
5.67
5.56
5.44
5.33
5.22
5.10
4.98
4.85
4.72
4.58
4.44
4.30
4.14
3.98
3.81
3.63
3.43
3.21
03-04-2019 12:00
03-06-2019 12:00
03-08-2019 12:00
03-10-2019 12:00
03-12-2019 12:00
03-14-2019 12:00
03-16-2019 12:00
03-18-2019 12:00
03-20-2019 12:00


Total Flow Bias (TAZT1)
Performance Indicators QR (CFS)
1 Day Leadtime (Adjusted)
1 Day Leadtime (NoMods)
2 Day Leadtime (Adjusted)
2 Day Leadtime (NoMods)
3 Day Leadtime (Adjusted)
3 Day Leadtime (NoMods)
4 Day Leadtime (Adjusted)
4 Day Leadtime (NoMods)
5 Day Leadtime (Adjusted)
5 Day Leadtime (NoMods)
Discharge (CFS)
Observed
03-06-2019 11:00
03-06-2019 17:00
03-06-2019 23:00
03-07-2019 05:00
03-07-2019 11:00
03-07-2019 17:00
03-07-2019 23:00
03-08-2019 05:00
03-08-2019 11:00
03-08-2019 17:00
03-08-2019 23:00
03-09-2019 05:00
03-09-2019 11:00
03-09-2019 17:00


Ohio River at Paducah
Major Flood Stage
Stage (FT)
54
53
52
51
50
49
48
47
46
2/16/19 0:00
2/18/19 0:00
2/20/19 0:00
2/22/19 0:00
2/24/19 0:00
2/26/19 0:00
2/28/19 0:00
3/2/19 0:00
3/4/19 0:00
3/6/19 0:00
3/8/19 0:00
Observed
2/18/2019 6:00:00
2/18/2019 18:00:00
2/19/2019 6:00:00
2/19/2019 18:00:00
2/20/2019 6:00:00
2/20/2019 18:00:00
2/21/2019 6:00:00
2/21/2019 18:00:00
2/22/2019 6:00:00
2/22/2019 12:00:00
2/22/2019 18:00:00
2/23/2019 6:00:00
2/23/2019 18:00:00
2/24/2019 6:00:00

ForecastProvider	Unmodified,ModifiedRaw	LeadTimeDays	006 (0.25 days)	PoolSeason	Winter,Spring,Summer,Fall
EventStart	10/29/2017	LocationGroup	Individual Locations	AverageBy	ForecastProvider,Date
EventEnd	10/29/2018	Location	Bear Creek Dam (BCRA1),Bear Creek nr	CompareBy	ForecastProvider

1 of 1 | 100% | Find | Next



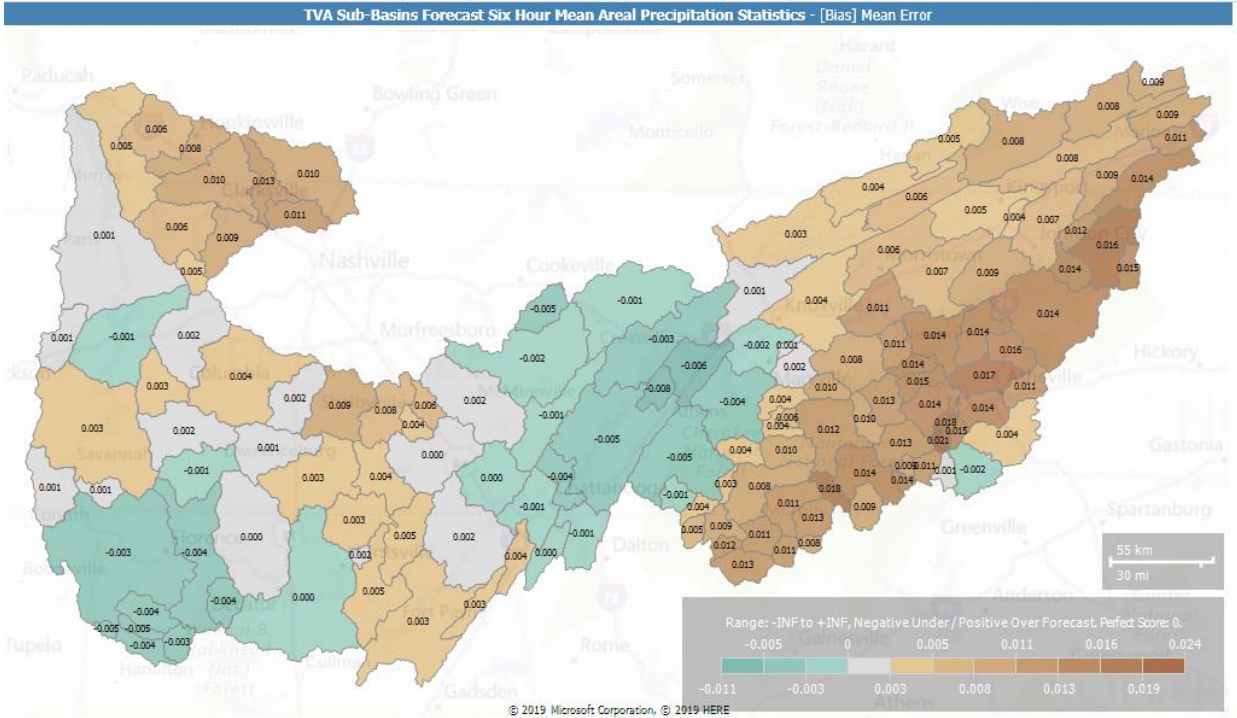

Correlation
1.2
1
0.8
0.6
0.4
0.2
0
-0.2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Extended,HRRR
HRRR
MeanAbsoluteError
0.014
0.012
0.01
0.008
0.006
0.004
0.002
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Extended,HRRR
HRRR
MeanError
0.002
0
-0.002
-0.004
-0.006
-0.008
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Extended,HRRR
HRRR
MultiplicativeBias
160
140
120
100
80
60
40
20
0
0
0.2
0.4
0.6
0.8
1
1.2
1.4
Extended,HRRR
HRRR

ForecastProvider: ML | LeadTimeDays: 012 (0.5 days) | RainfallClass: ZERO (0.0 : 0.0),LOW (0.0 : 0.0001),LOW

EventStart: 3/6/2018 | PoolSeason: Winter,Spring,Summer,Fall | Statistic: [Bias] Mean Error

EventEnd: 3/6/2019 | Layer: Sub-Basins | ShowData: True / False

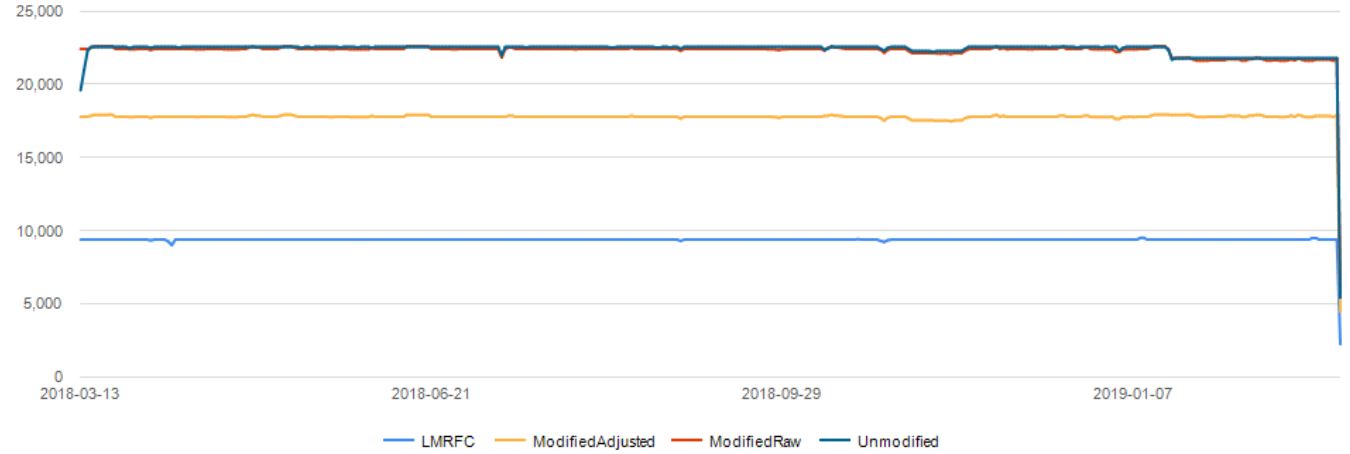
1 of 1 | 100% | Find | Next



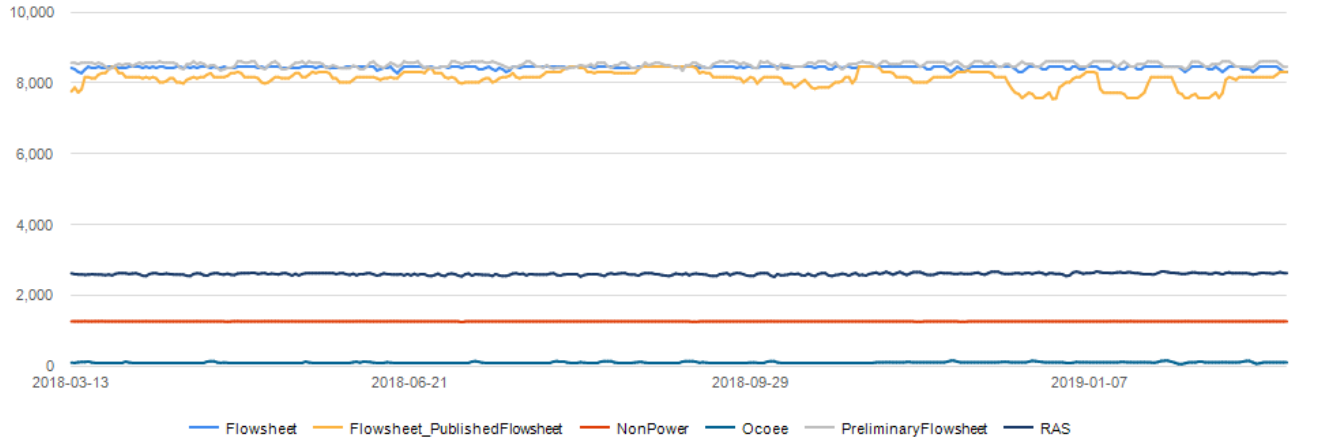
What is the average forecast error? Also called the (additive) bias. Does not measure the magnitude of the errors. Does not measure the correspondence between forecasts and observations, i.e., it is possible to get a perfect score for a bad forecast if there are compensating errors.

For Further Reading / Formula Details:
http://www.cawcr.gov.au/projects/verification/
https://en.wikipedia.org/wiki/Receiver_operating_characteristic

Inflow SixHour Data Availability



HeadwaterSix Hour Data Availability



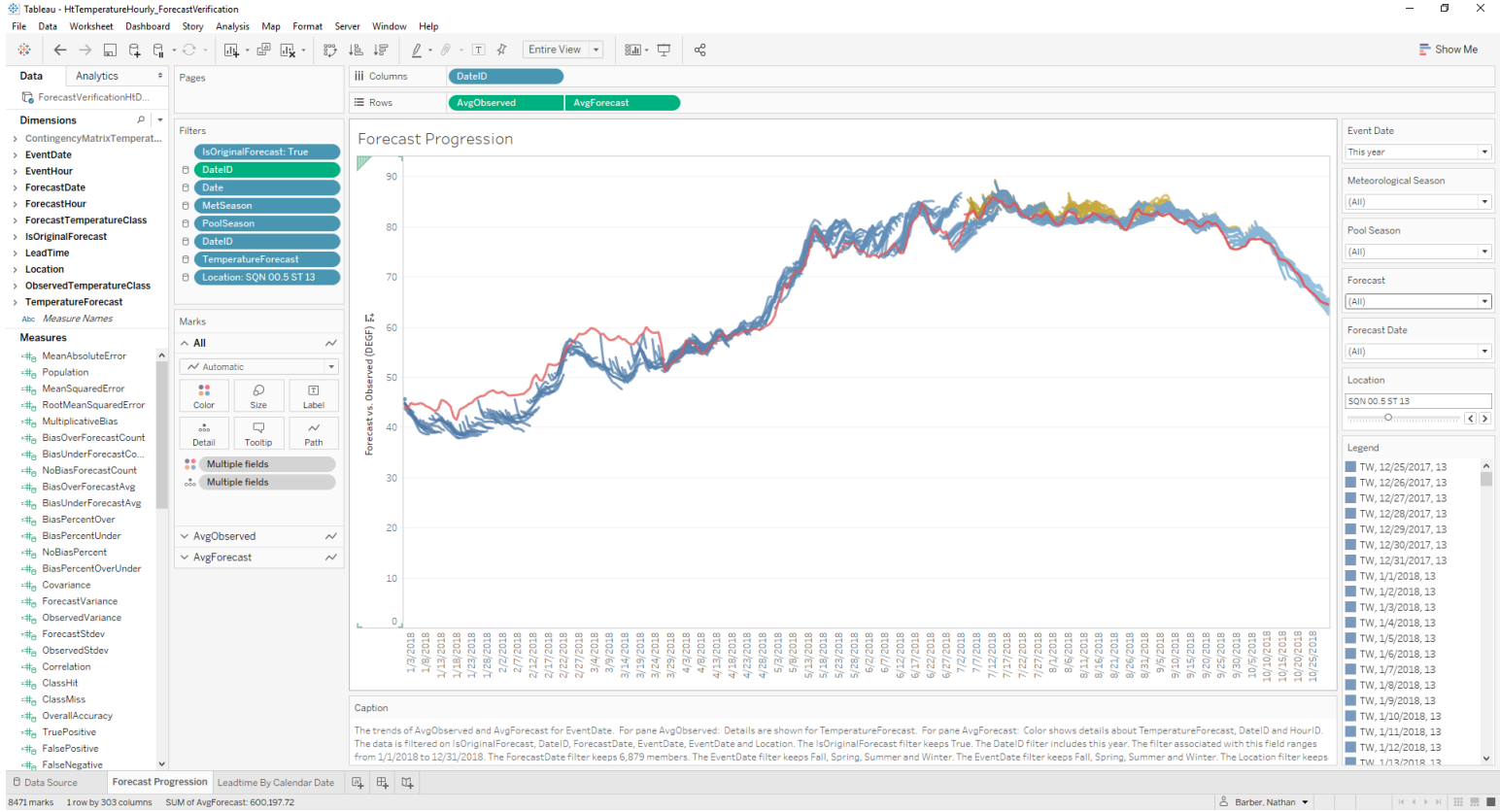
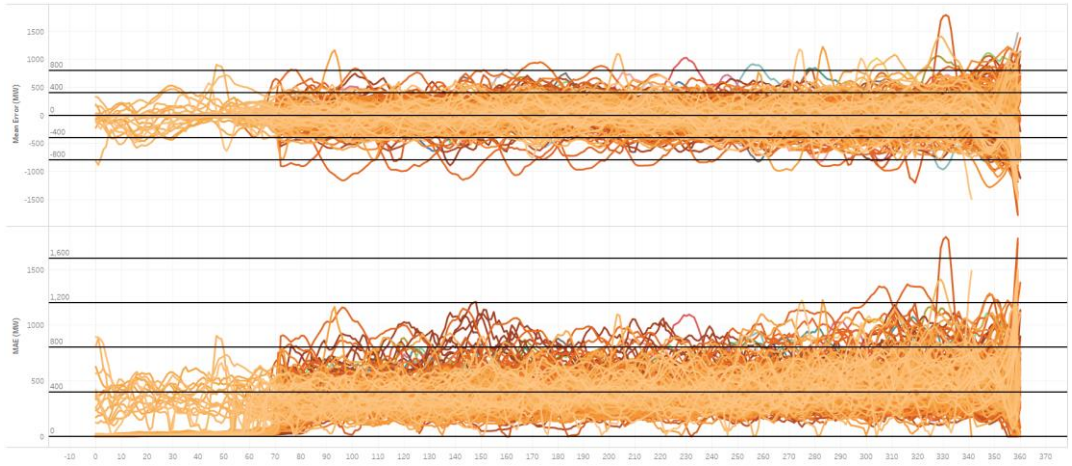
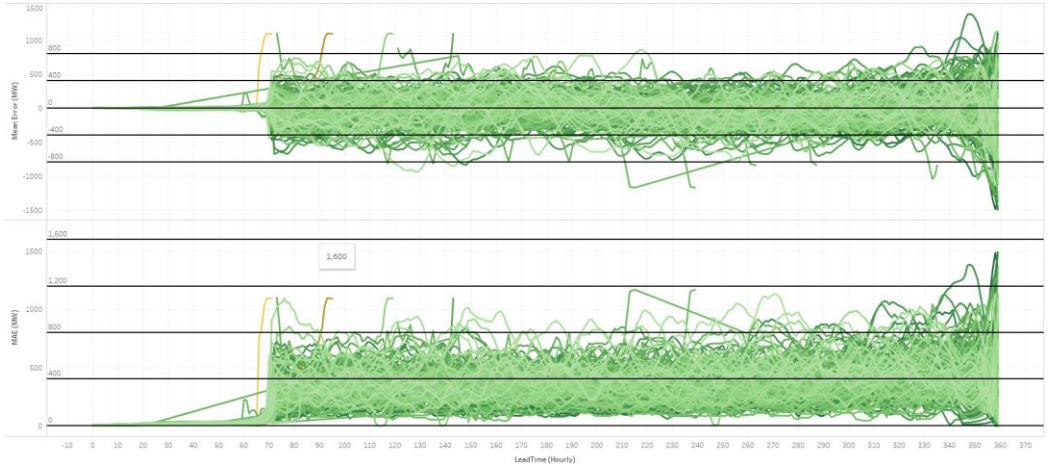
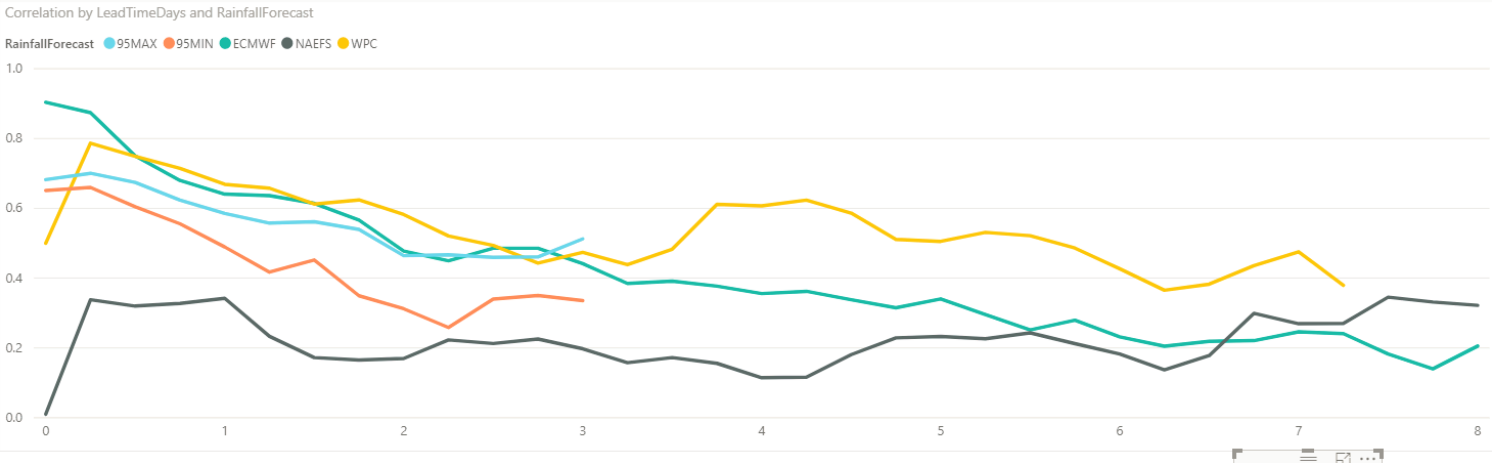
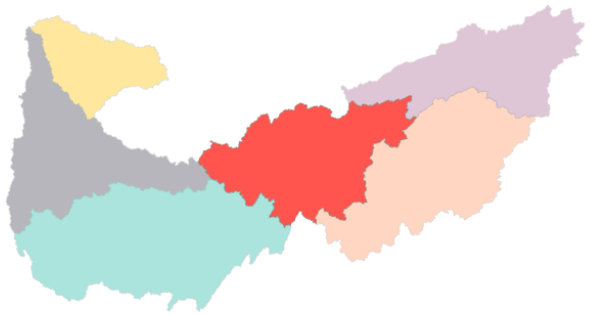


Tableau - HtTemperatureHourly_ForecastVerification
File Data Worksheet Dashboard Story Analysis Map Format Server Window Help
Entire View
Show Me
Data
Analytics
ForecastVerificationHtD...
Dimensions
ContingencyMatrixTemperat...
EventDate
EventHour
ForecastDate
ForecastHour
ForecastTemperatureClass
IsOriginalForecast
LeadTime
Location
ObservedTemperatureClass
TemperatureForecast
Measure Names
Measures
MeanAbsoluteError
Population
MeanSquaredError
RootMeanSquaredError
MultiplicativeBias
BiasOverForecastCount
BiasUnderForecastCo...
NoBiasForecastCount
BiasOverForecastAvg
BiasUnderForecastAvg
BiasPercentOver
BiasPercentUnder
NoBiasPercent
BiasPercentOverUnder
Covariance
ForecastVariance
ObservedVariance
ForecastStdev
ObservedStdev
Correlation
ClassHit
ClassMiss
OverallAccuracy
TruePositive
FalsePositive
FalseNegative
Pages
Filters
IsOriginalForecast: True
DateID
Date
MetSeason
PoolSeason
DateID
TemperatureForecast
Location: SQN 00.5 ST 13
Marks
All
Automatic
Color
Size
Label
Detail
Tooltip
Path
Multiple fields
Multiple fields
AvgObserved
AvgForecast
Columns
DateID
Rows
AvgObserved
AvgForecast
Forecast Progression
Forecast vs. Observed (DEGF)
Event Date
This year
Meteorological Season
(All)
Pool Season
(All)
Forecast
(All)
Forecast Date
(All)
Location
SQN 00.5 ST 13
Legend
TW, 12/25/2017, 13
TW, 12/26/2017, 13
TW, 12/27/2017, 13
TW, 12/28/2017, 13
TW, 12/29/2017, 13
TW, 12/30/2017, 13
TW, 12/31/2017, 13
TW, 1/1/2018, 13
TW, 1/2/2018, 13
TW, 1/3/2018, 13
TW, 1/4/2018, 13
TW, 1/5/2018, 13
TW, 1/6/2018, 13
TW, 1/7/2018, 13
TW, 1/8/2018, 13
TW, 1/9/2018, 13
TW, 1/10/2018, 13
TW, 1/11/2018, 13
TW, 1/12/2018, 13
Caption
The trends of AvgObserved and AvgForecast for EventDate. For pane AvgObserved: Details are shown for TemperatureForecast. For pane AvgForecast: Color shows details about TemperatureForecast, DateID and HourID. The data is filtered on IsOriginalForecast, DateID, ForecastDate, EventDate, EventDate and Location. The IsOriginalForecast filter keeps True. The DateID filter includes this year. The filter associated with this field ranges from 1/1/2018 to 12/31/2018. The ForecastDate filter keeps 6,879 members. The EventDate filter keeps Fall, Spring, Summer and Winter. The EventDate filter keeps Fall, Spring, Summer and Winter. The Location filter keeps
Data Source
Forecast Progression
Leadtime By Calendar Date
8471 marks
1 row by 303 columns
SUM of AvgForecast: 600,197.72
Barber, Nathan


Mean Error (MW)
1500
1000
800
500
400
0
-400
-500
-800
-1000
-1500
MAE (MW)
1,600
1500
1,200
1000
800
500
400
0
-10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370


Mean Error (MW)
1500
1000
800
500
400
0
-400
-500
-800
-1000
-1500
MAE (MW)
1,600
1,600
1500
1,200
1000
800
500
400
0
-10 0 10 20 30 40 50 60 70 80 90 100 110 120 130 140 150 160 170 180 190 200 210 220 230 240 250 260 270 280 290 300 310 320 330 340 350 360 370
LeadTime (Hourly)





RainfallClass...
- ZERO
- LOW
- LOW-MID
- MID-LOW
- MEDIUM
- MID-HIGH
- HIGH-MID
- HIGH

MetSeason
- Fall
- Spring
- Summer
- Winter

DateID
9/28/2014 10/31/2018

LeadTimeDays
0.00 8.00

RainfallForecast
- 95MAX
- 95MIN
- Avg95MinMax
- ECMWF
- Extended,HR...
- HRRR
- ML
- NAEFS
- Normal
- Selected
- WPC

http://riverwiki/wiki/Main_Page/Verification

Main Page/Verification - Riv... | Waiting for knxawtableaupri | LocalFlowHourlyScatterPlot - ... | MapHourlyLeadTime - Report ...

Page | Safety | Tools

Main page
Recent changes
Email Questions

- Checklists
  - Roles
  - Lead
  - Preschedule
  - Kentucky Barkley
  - Data Steward
- FEWS Basics
- River Forecast System
- Hydrothermal
- System Documentation
- Wiki Basics
- Tools

Main Page/Verification

Contents [hide]

Why Verify?

"*Forecasts are almost always made and used in the belief that having a forecast available is preferable to remaining in complete ignorance about the future event of interest. It is important [however] to test this belief a posteriori by assessing how skillful or valuable a given forecast is.*" - Ian T. Jolliffe/David B. Stephenson (Mathematics Research Institute, University of Exeter".

In the context of TVA's River Management, forecasts and decisions are an integral and essential component of the mission. In fact, many of the forecasts and subsequent decisions are based on other forecasts - These come from other federal agencies, private utilities, vendors, and TVA organizations. TVA River Management's River Forecast Center is then tasked with integrating all of these forecasts in a variety of models and then coming up with the best decision on when and where to move water within the Tennessee River Valley. While expert experience naturally lends itself to a qualitative understanding of uncertainty/performance/skill in these forecasts and decisions, much of this can be hidden just below the surface. This verification system will allow forecasters, managers, and hydrologists, the opportunity to look as shallow or deep as necessary with respect to forecasts and the corresponding "truth". This system can be used to

1. Track performance in forecasts
2. Help forecasters choose which forecast to use in the models (when many exist)
3. Identify bias or persistent errors in TVA forecasts
4. Identify bias or persistent errors in external forecasts
5. Understand data integrity
6. Identify process improvements/issues
7. Propose new projects and identify funding opportunities based on above results
8. Compare common forecasts between TVA and external agencies (NWS) for the purposes of mutually improving
9. Train forecasters on how modifications can affect performance

What is Verification?

Verification allows River Management the ability to quantify how well a given forecast or set of forecasts perform(s) with respect to a corresponding observation. Forecasts that do not have a corresponding observation are not part of this system. Since performance is subjective and depends on the question and measure, the roughly 156 pre-calculated statistics are available to the user to choose from. Many of these are common to all forecasts (such as Mean Error) while some are specific to the measure (such as skill scores - only relevant to MAP). Yet others are specific to the way the measures are sliced (Class-based performance measures). The variety of verification statistics allows the user to ask a specific question and choose the statistic(s) that best answer(s) or inform(s) that question. For instance, if I am interested in the magnitude of error for a given forecast, I would likely choose **mean absolute error** as my statistic. If I am interested in understanding how a given forecast performs over time with respect to an unskilled forecast, I would likely choose **skill score mean absolute error**. If I am interested in identifying how a given forecast represents reality, I would likely choose **correlation**. In some cases, several statistics are necessary to answer the question. In the rare case that none of the 156 pre-calculated statistics are sufficient, many of the access tools provide a means to generate specific calculations. An expert user could even bypass the verification system and manipulate the raw data via the tools, to inform a highly-specialized study.

Measures

3:53 PM
11/5/2018

Next Steps

- Error propagation
- Better quantification and visualization of statistical significance
- Broader community of users
- More focused training
- R/Python integration when necessary

Lessons Learned

- Start with a focused set of questions and data
- Start simple
- Verify your verification
- Consult with experts every step of the way
- Focus on maintainability, sustainability, and extensibility with respect to a verification system
- Use quantitative information to augment qualitative information

Acknowledgements

- Curt Jawdy – Lead Hydrologist (TVA)
- Gabriel Miller – Hydrologist (TVA)
- Edward Roush – Software Engineer/Data Scientist (Infisys)
- Kelsey Rector – Software Engineer (TVA)
- Caige Pirtle – Software Engineer (TVA)
- Andre Grijze – Java Developer (Deltares)
- Jan Verkade – Hydrologist (Deltares)


TVA