Model calibration objectives

Harpeth River water quality model

J. Davis

03/04/2021

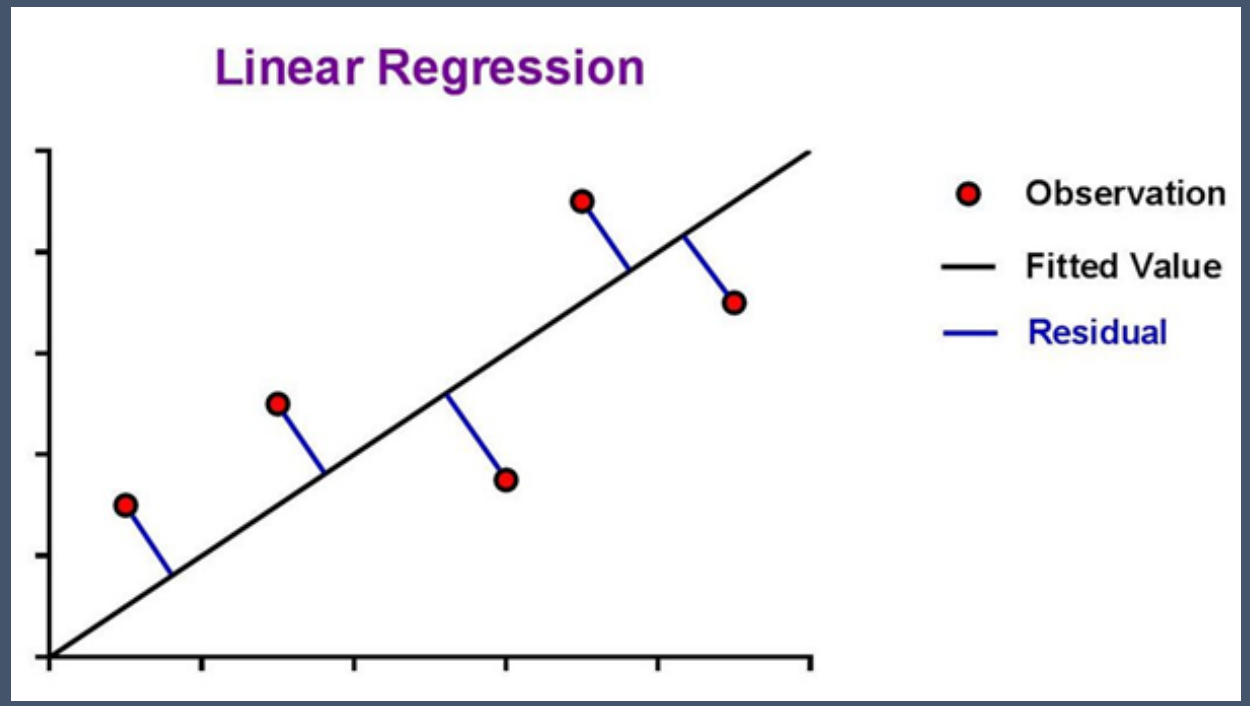
What is model calibration and how do we assess it?

Model error

- Total model error = Irreducible error + Reducible error

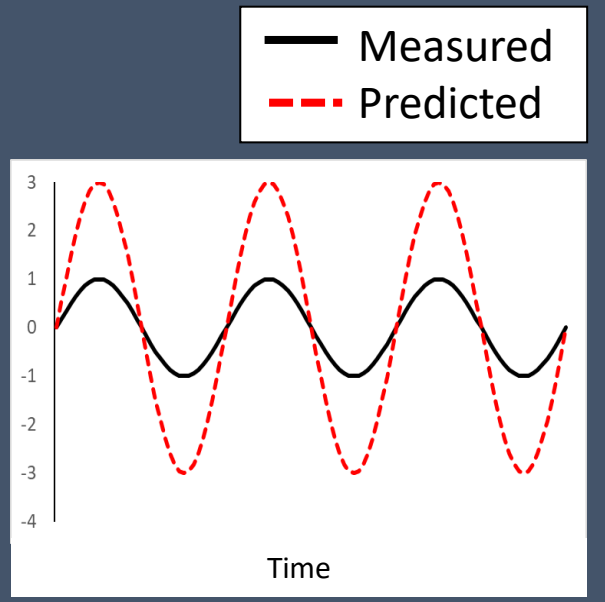
- Irreducible (inherent) error:
  - Sampling/analytical error
  - Natural variability
  - Random error (noise)
  - Difficult to control w/o ‘overfitting’

- Reducible error
  - Model bias
  - Model variance
  - Can be minimized w/ model calibration



Bias vs. variance vs. correlation

- Model bias (Accuracy)
  - How much do predicted values differ from true measured values?
- Model variance (Precision)
  - How much do predicted values vary from each other?
- Model correlation (Temporal dynamics)
  - How well do temporal dynamics of predicted values capture dynamics of measured values?



The optimal model

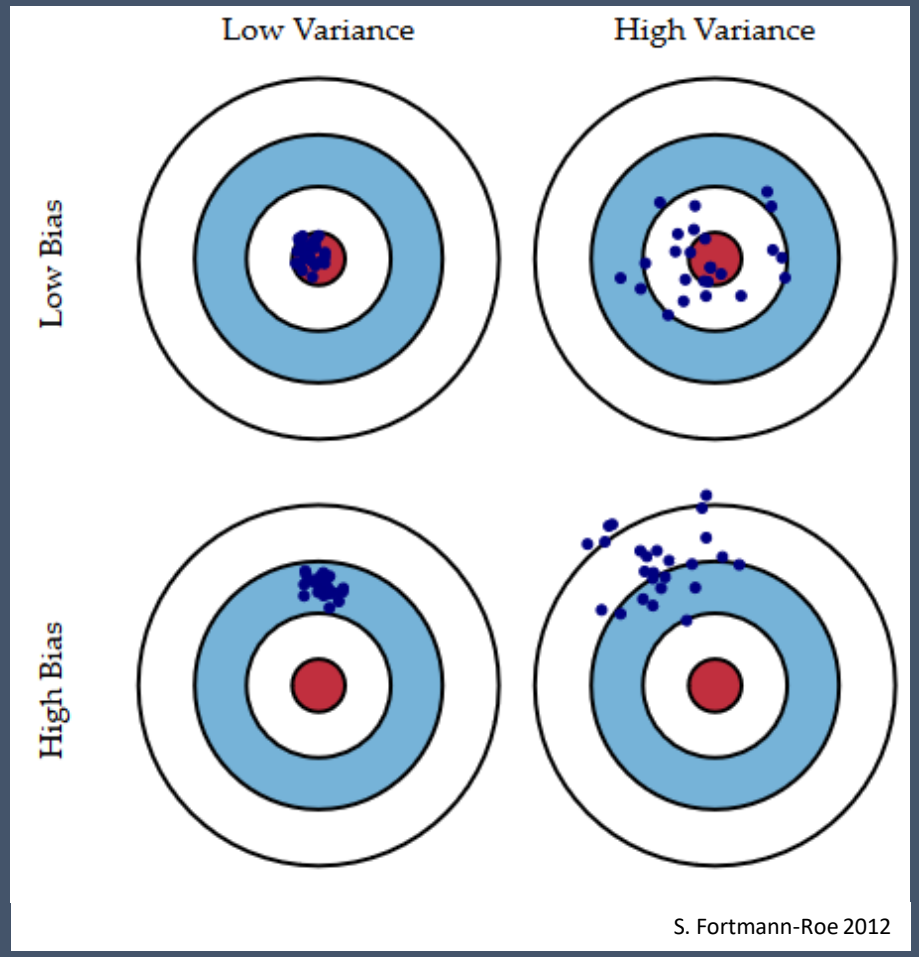
Optimal model =
Low variance & Low bias

Low variance & Low bias

Model consistent and accurately predicting true measure values

High bias & Low variance

Model consistent but not accurately predicting true measure values



High variance & Low bias

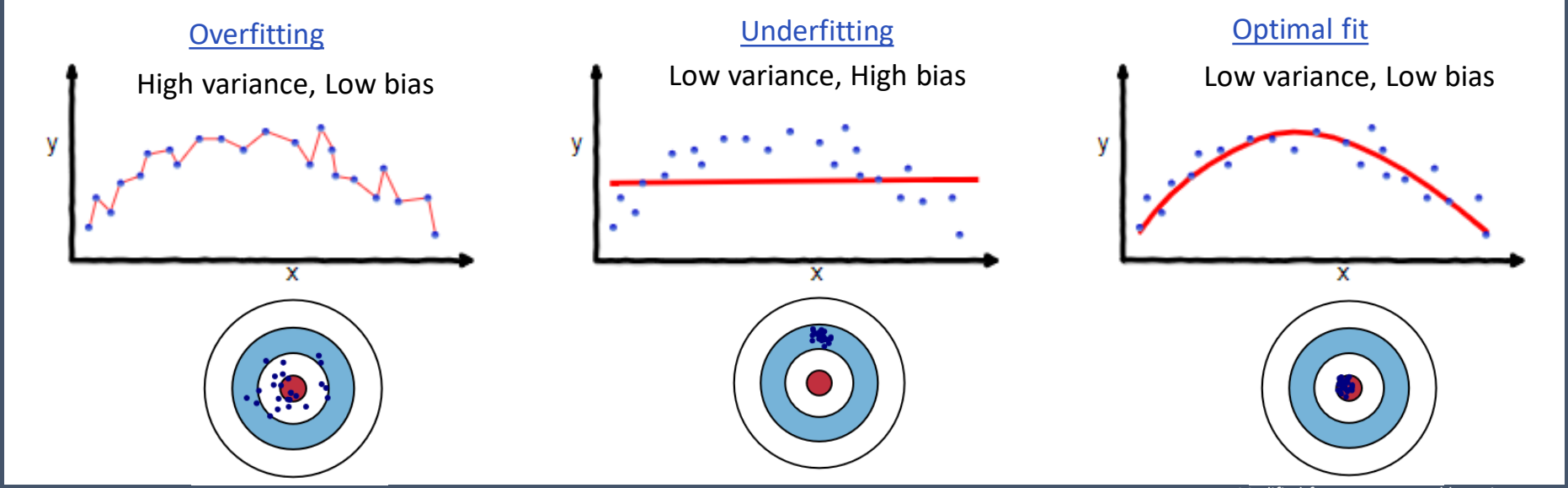
Model inconsistent but accurately predicting true measure values

High variance & High bias

Model inconsistent and not accurately predicting true measure values

Each blue dot represents a predicted value based on the model. Bullseye represents true measured value

Under vs. overfitting



Modified from www.towarddatascience.com

- Overfit model = model that has been overly trained on data
  - Predicts current data well, but performs poorly when applied to novel data or scenario
  - Captures random noise and may not generalize well

- Underfit model = overly simple model that does not capture true underlying trend
  - e.g., Applying linear regression to nonlinear data

- Both perform poorly when applied to new datasets/environmental scenarios

How to avoid overfitting

- Constrain parameters/constants used in the model based upon measured data
  - No single answer to how many parameters/constants to include in a model
- Definition of 'well-calibrated' may differ for measured data that are naturally variable
  - Depends on quantity/quality of measured data and modeling objective
- Assess model fit based on full range of measured data
  - Is the model missing extreme tails/outliers (99^{th} percentile) in measured data, but well-predicting bulk of measured data distribution (~90^{th} percentile)?
  - What if water quality standard is based on low-flow conditions?
- May give greater weight to stations that integrate larger spatial extent or longer time span
  - May exclude stations that don't meet a minimum number of samples

Reminder: models are only mathematical representations of the natural world, and don't have to explicitly include all processes/mechanisms

How are model bias, variance, and correlation assessed?

Two methods for assessing models

Quantitatively

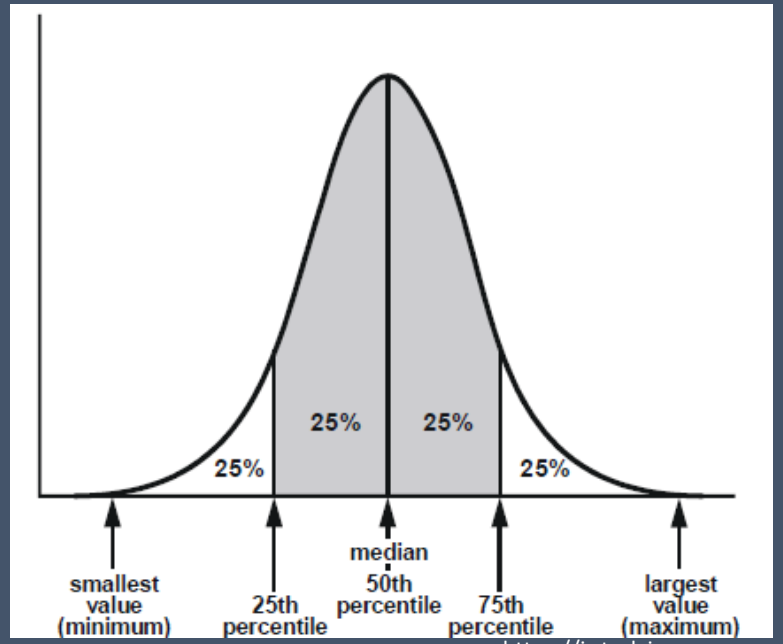
- Can be counted, measured, & expressed numerically
- Summary statistics
- Statistical tests (e.g., goodness of fit)

Qualitatively

- Can be descriptive, conceptual, & expressed categorically
- Time series plots
- Animations
- Frequency/probability

Quantitative analysis – statistical measures

- Descriptive statistics
  - Average, median, range
  - Percentiles, quantiles
- Measures of error (Predicted – Observed)
  - Mean error
  - Mean absolute error (MAE)
  - Root mean square error (RMSE)
  - Normalized RMSE (nRMSE)
- Regression/correlation analysis
  - Standard deviation (σ)
  - Spearman rank coefficient (rho)
  - Coefficient of determination (r^2)
- Efficiency metrics
  - Percent bias (PBIAS)
  - Index of agreement (d)
  - Nash-Sutcliffe efficiency (NSE)
  - Modified Kling-Gupta efficiency (KGE')



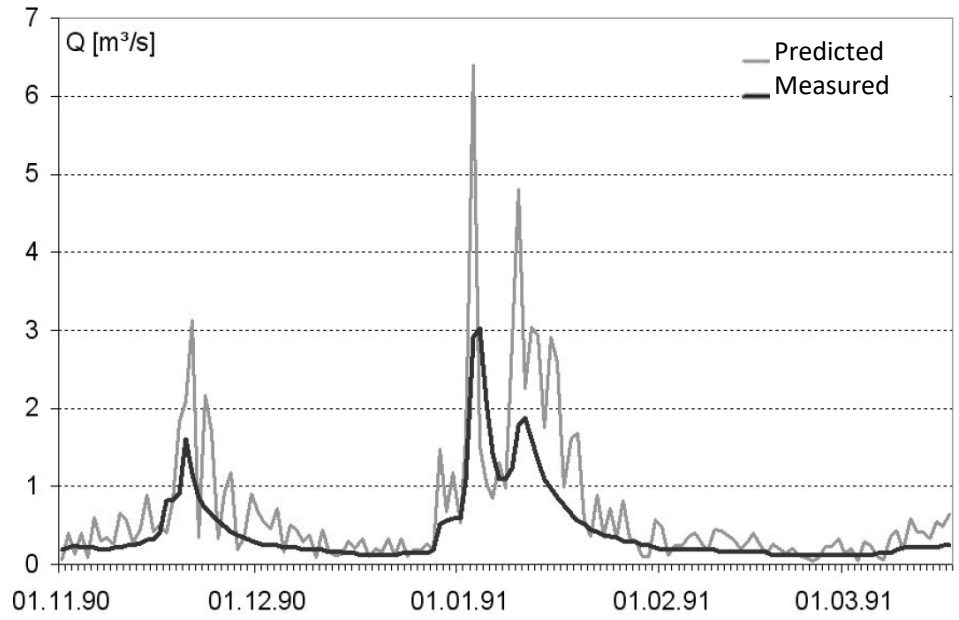
https://i.stack.imgur.com

Examples

Dataset	Average	10%	20%	50%	80%	90%
OBS	26.420	1.53	2.65	8.72	36.47	62.52
SIM	26.193	2.99	4.80	11.04	31.84	57.21

GoF Metric	Value
Num Obs	3652.0000
R2	0.7655
NSE	0.7601
RMSE	28.6905
d	0.9330

Limitations of r^2 – effects of covariance

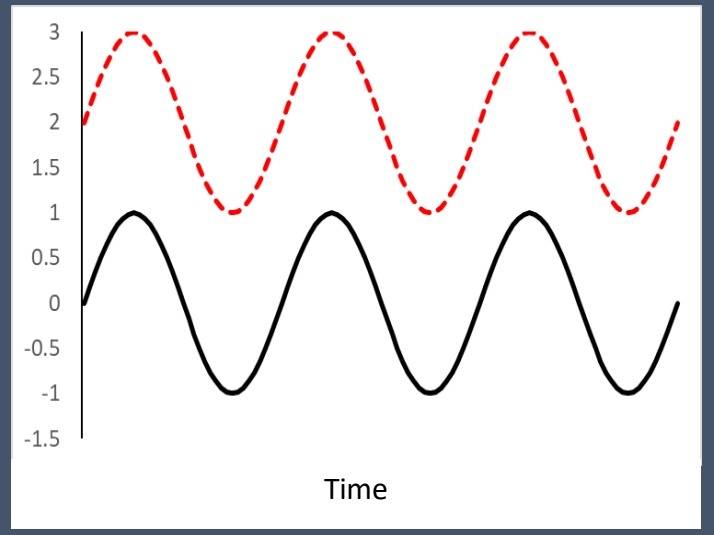


Predicted vs. measured discharge

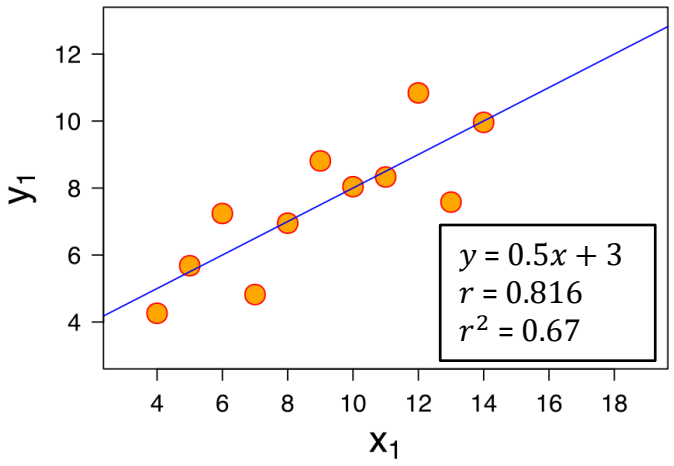
Krause et al. 2005

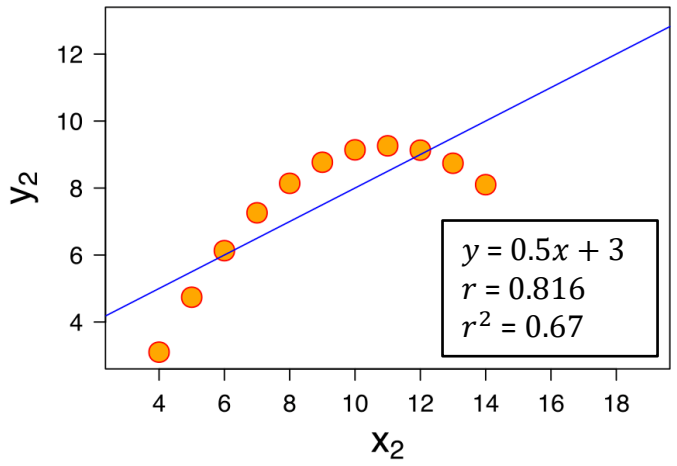
- r^2 can increase w/ larger sample size

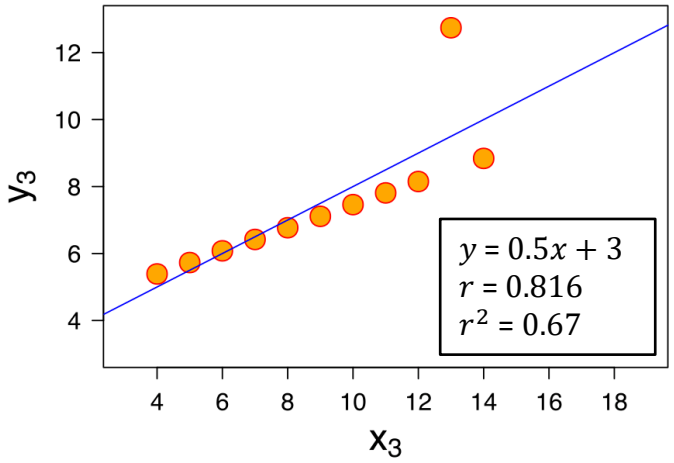
- Model consistently over/underpredicts measured values
  - Measured and simulated data are covarying
  - High r^2
  - Avg. value may still be poorly predicted

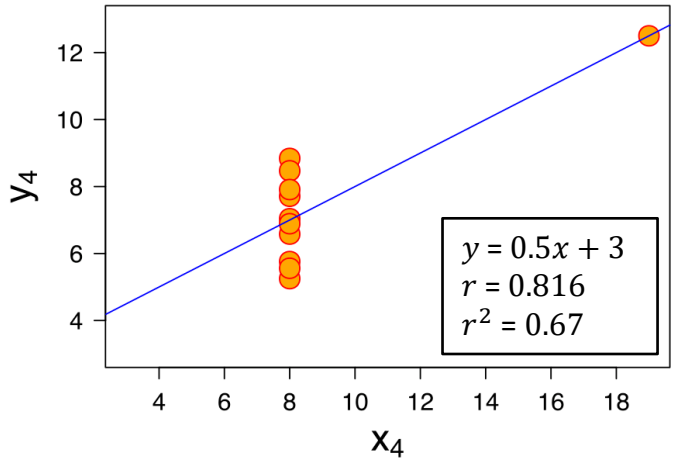


Limitations of r^2 – Anscombe's quartet







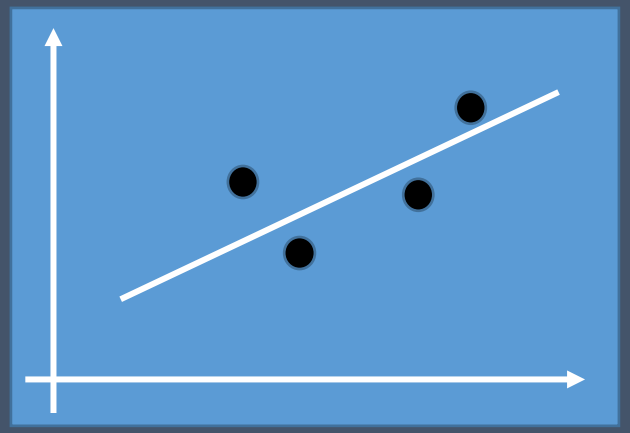


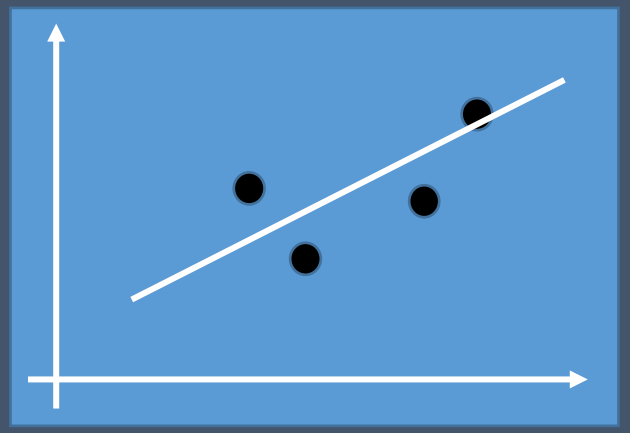
Modified from www.wikipedia.com

- Based on r^2, which is the best-fitting model?
- Statistics
  - Aid with discovering bias
  - Assess central tendency
  - Should be included in analysis
  - May require additional information to interpret

Quantitative & qualitative

- Visually (Qualitatively), these seem like okay fits, but…





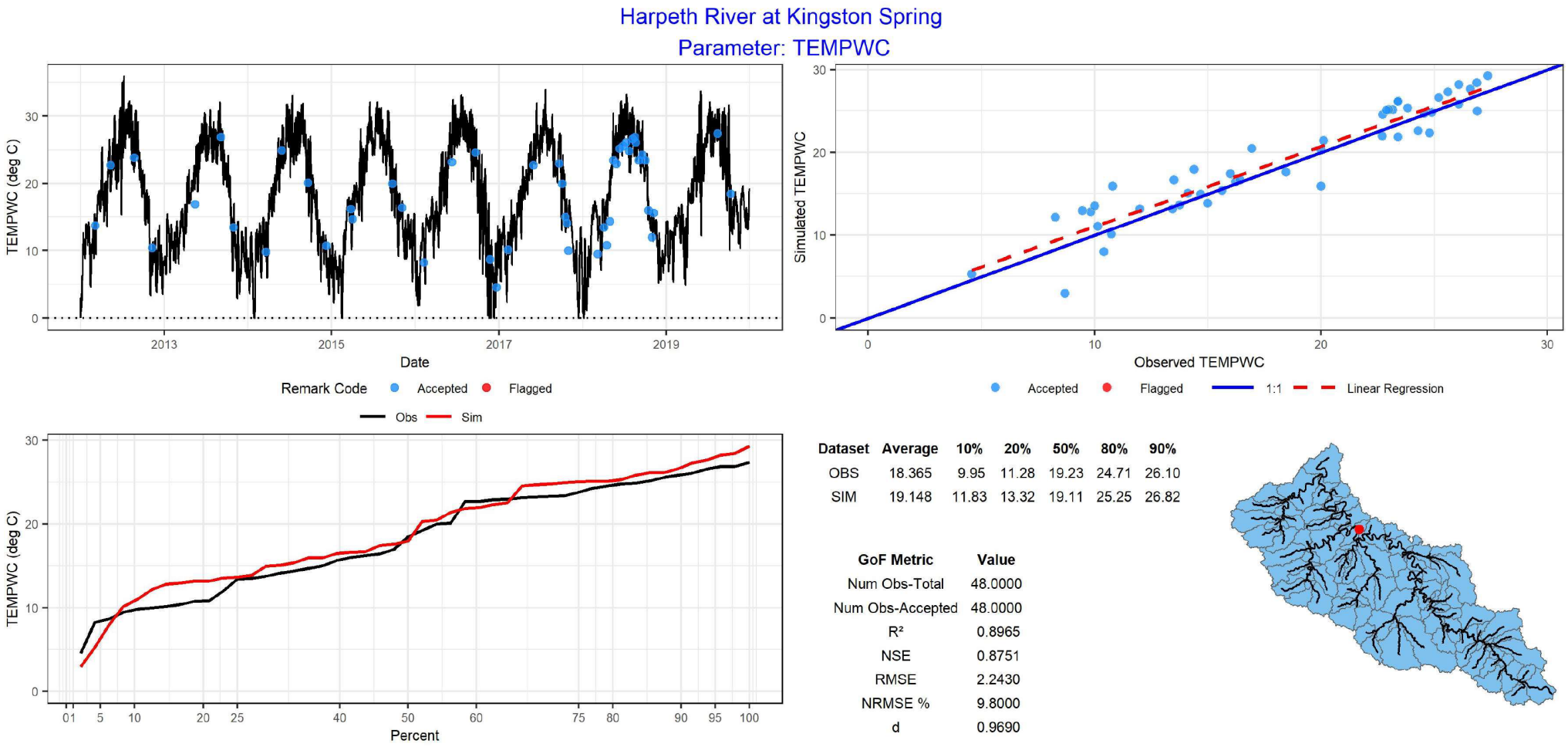
Which one's better??

- Need statistical (Quantitative) measures

- Weight of evidence approach

Time series

1:1 Simulated vs. measured



Dataset	Average	10%	20%	50%	80%	90%
OBS	18.365	9.95	11.28	19.23	24.71	26.10
SIM	19.148	11.83	13.32	19.11	25.25	26.82

GoF Metric	Value
Num Obs-Total	48.0000
Num Obs-Accepted	48.0000
R^2	0.8965
NSE	0.8751
RMSE	2.2430
NRMSE %	9.8000
d	0.9690

(Calib Station: FRANK-KINGSPR / HRC-KINGSPR / TNW000002793; WASP Seg: 14)

Cumulative probability distribution

Statistical summary & map

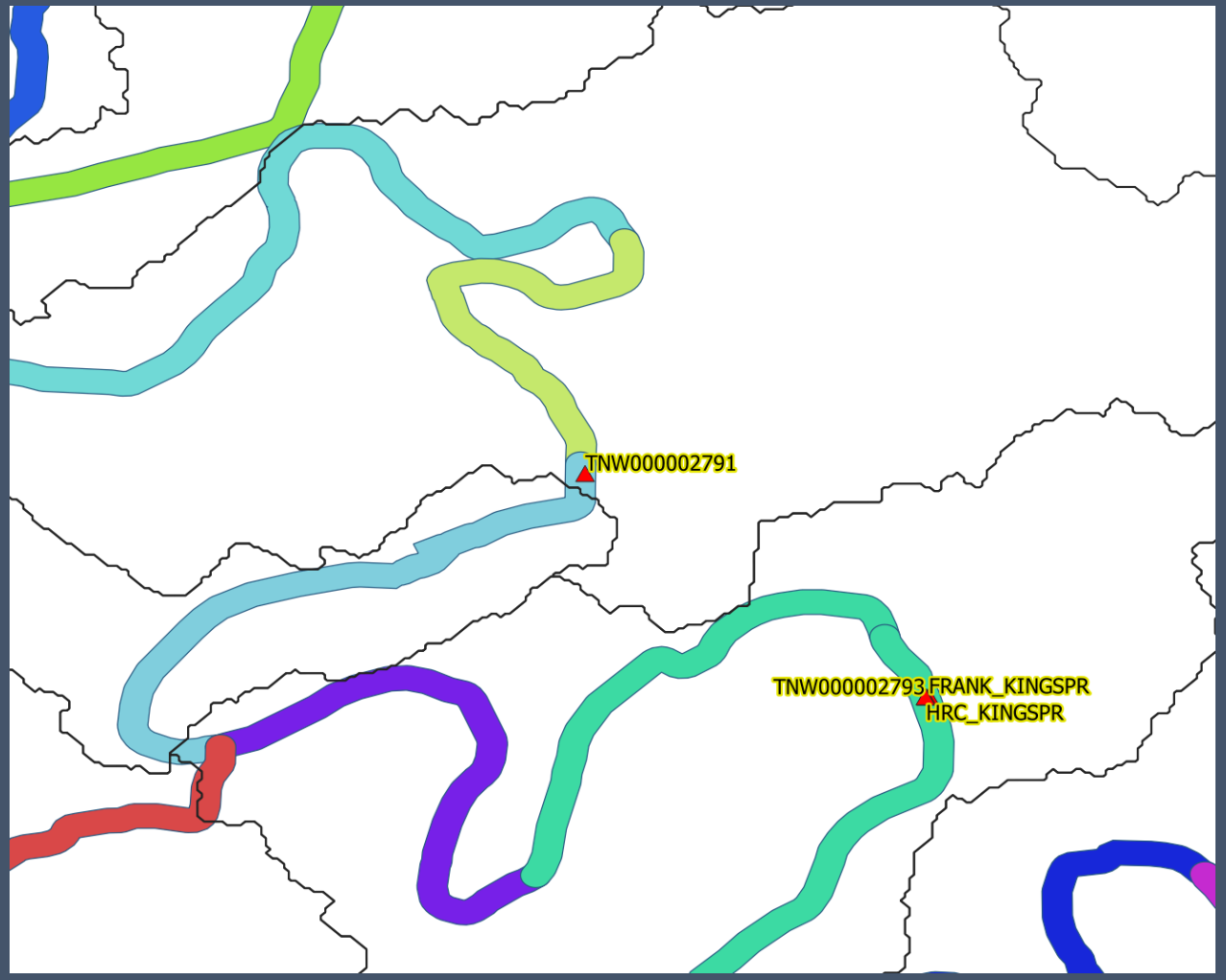
How well does the model predict temporal dynamics of measured values (correlation), and do those predictions also capture the measured data's average (bias) and variability (variance)?

Additional factors to consider when assessing model fit

Spatial considerations

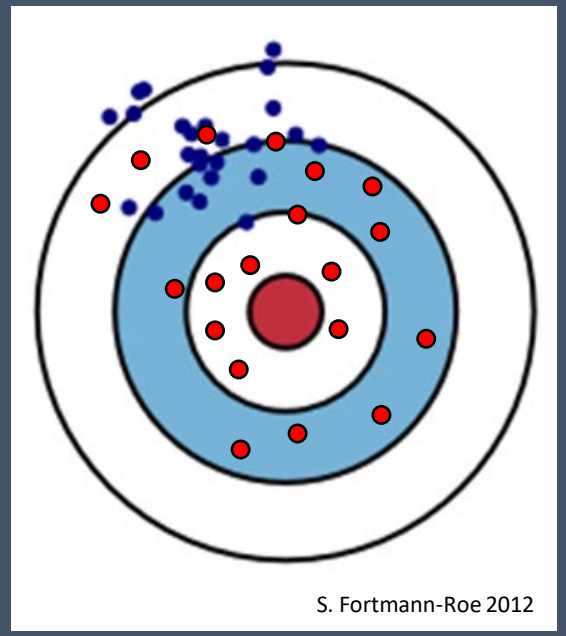
- Simulated model results represent average condition across entire model segment

- Monitoring data represent measured value at specific location at specific time (snapshot)



Monitoring data variation

- Bullseye (i.e., average) treated as ‘true’ measured value
- Variability in measured data not explicitly considered in many model-fit statistics
- Measured data may not even fall within bullseye
- Need to assess model fit and calibration relative to monitoring data quality/quantity



S. Fortmann-Roe 2012

Monitoring data quality—USGS flow gages

Characterization	Description
Excellent	95% of daily discharge within 5% of true values
Good	95% of daily discharge within 10% of true values
Fair	95% of daily discharge within 15% of true values

- USGS characterizes quality of flow data

- If USGS measurements are only within 15% of true values, is it realistic to expect a watershed/water quality model to perform better than that?

Questions?