A Novel Use of R to Couple Hierarchical Bayesian Methods with a Spatially Explicit Hydrological Model across Regional and Continental Scales

Richard B. Alexander and Gregory E. Schwarz

National Water Quality Assessment Project

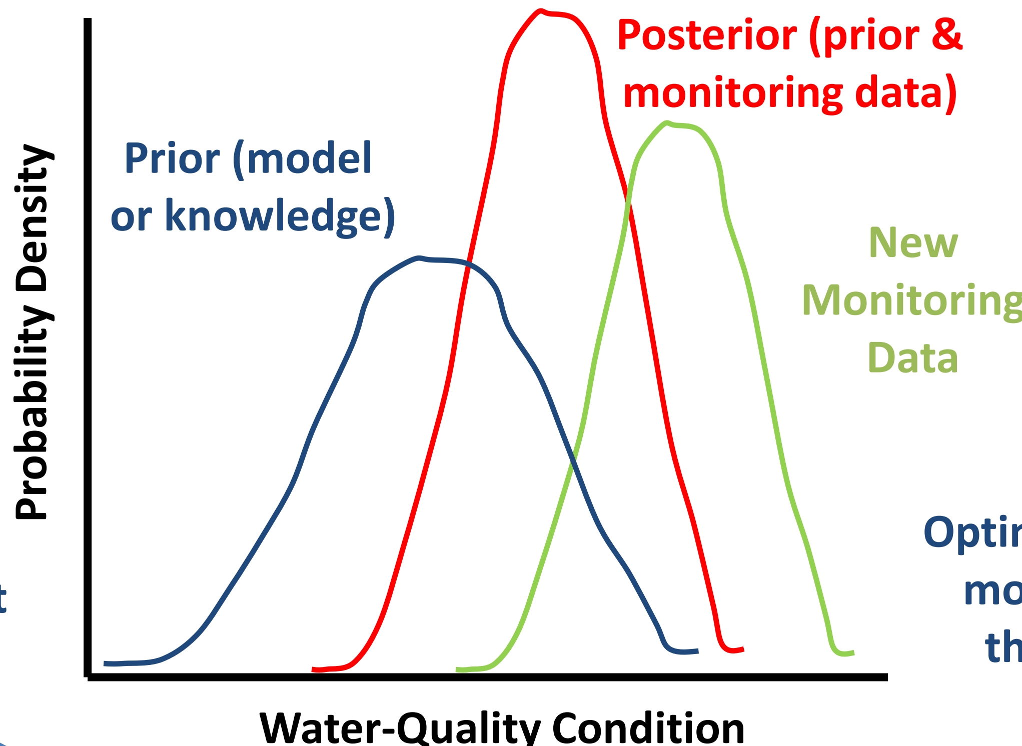
U.S. Geological Survey

Reston, VA

USGS science for a changing world

NWQMC 2016 Conference, Tampa, FL, May 2-6, 2016

Growing Interest in Bayesian Methods



User decisions about data sharing over space and time

Optimized watershed models maximize this probability

Employs conditional probability theory expressed by Bayes' Theorem:

$$P[\text{ model} | \text{ data }] \propto P[\text{ model }] * P[\text{ data } | \text{ model }]$$

Updated knowledge of model (posterior distribution)

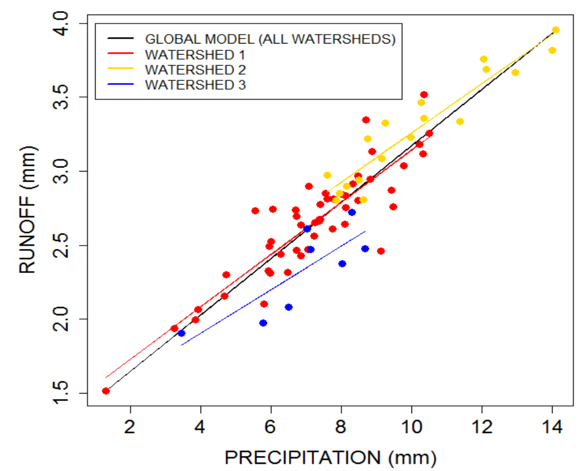
Prior model

New data and the model that best describes the data

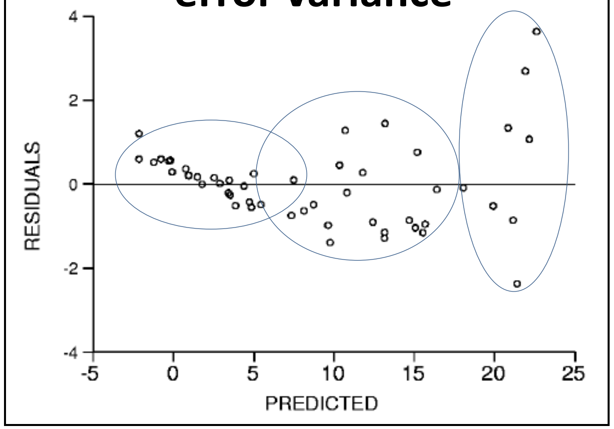
Hierarchical Bayesian Methods

- **Provide flexibility in describing complexities in data over space and time**
  - Model coefficients (and error variance) treated as random variables, with values that vary in time/space
  - Hierarchical structure nests higher-order models of parameters (e.g., region) within lower-order structural models (e.g., monitoring stations)
  - Partially pooling ("data borrowing") can improve prediction accuracy (bias, precision)
  - Handles small samples well

Allow partial pooling of site data across watersheds



Allow modeling of error variance



Hierarchical Bayesian Methods

- **Improve quantification of model uncertainties (*state-space* methods)**
  - Explore spatial and temporal patterns in the uncertainties
  - Identify causes of uncertainties
  - Allows separate estimation of ***measurement*** errors in observations vs. real ***process-related*** errors associated with the model structure
  - Produces more accurate estimates of uncertainties and ***updated predictions*** of "true" WQ conditions
- **Provide probabilistic model outcomes**
  - Inclusive of uncertainties
  - Avoid prediction biases related to transformed response variables

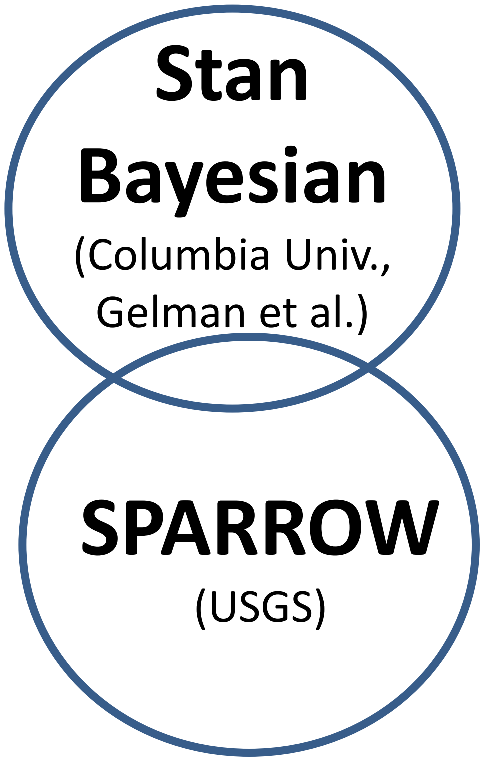
Challenges in Bayesian Hydrological and Watershed Modeling

- Previous Bayesian applications:
  - evaluated small catchments with little diversity (e.g., soils, climate, land use)
  - few systematic comparisons of a wide range of hierarchical structures

- Reported computational limitations of the available software for hydrological models (e.g., WinBugs, RJAGS)

- *Declarative* programming style (e.g., WinBugs) inconsistent with *imperative* procedural style (*order matters*) of many hydrologic models >> *complicates use of the software*

A New-Generation Bayesian Method



Stan

- Hamiltonian Monte Carlo (HMC) methods – uses a more strategic sampling of posterior parameter distributions than Markov Chain Monte Carlo methods
- More efficient (~10x faster) and robust
- R interface (**Rstan**) uses *imperative* procedural language with C++ translator

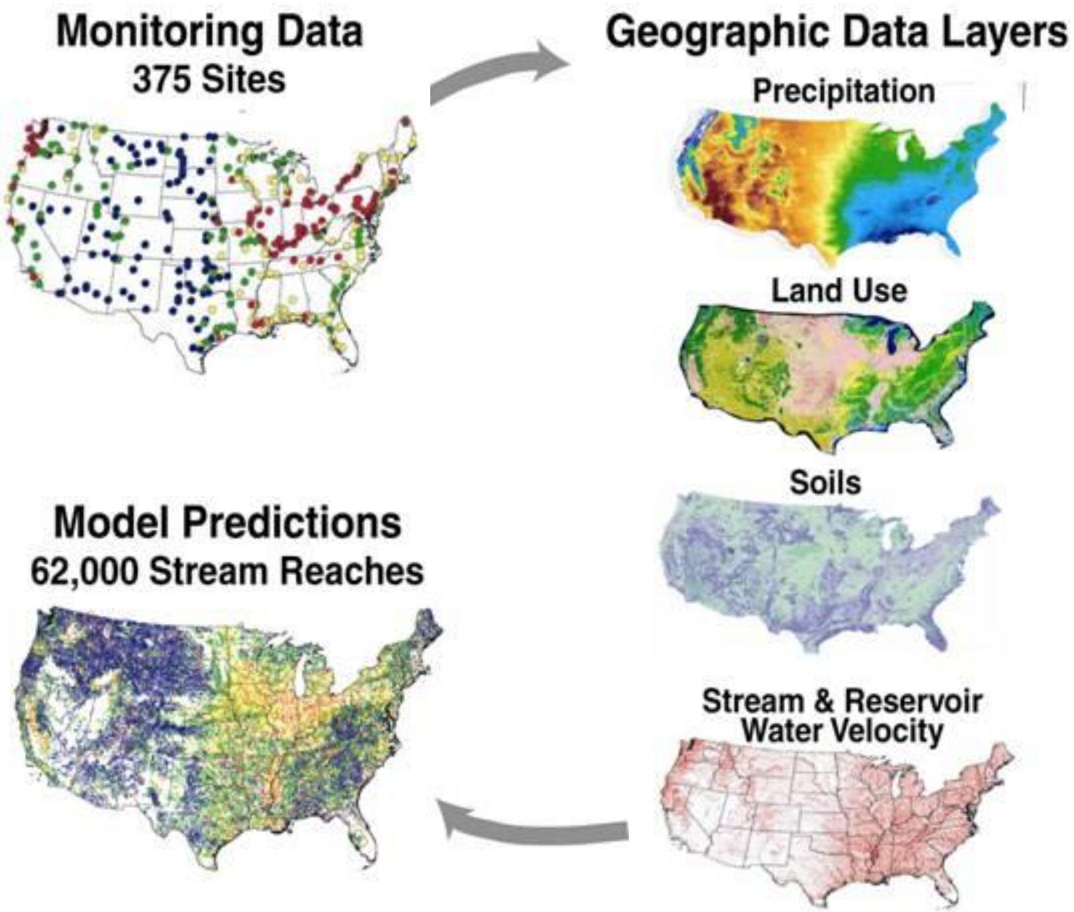
USGS SPARROW water-quality model

- R scripts, with R libraries under development
- Novel automated construction and execution of standard user-selected models:

  Excel parameter table settings >> Stan script
- Evaluated multi-spatial/temporal scale performance
- **Results:** Improved quantification of uncertainties over space/time, with wide-ranging improvements in prediction accuracy (small to ~50%)

USGS SPARROW Water-Quality Model

SPAtially Referenced Regression on Watershed Attributes (Smith et al., 1997)



Hybrid mechanistic and statistical features:

- Spatially explicit (land/water)
- Mass-balance constraints
- Non-conservative transport
- Parameter estimation using least squares optimization

Capabilities:

- Predict annual water-quality load, yield, and concentration for unmonitored stream reaches
- Assess effects of hydrological and biogeochemical processes on transport and fate in watersheds
- Apportion stream loads to major pollution sources and upstream watersheds

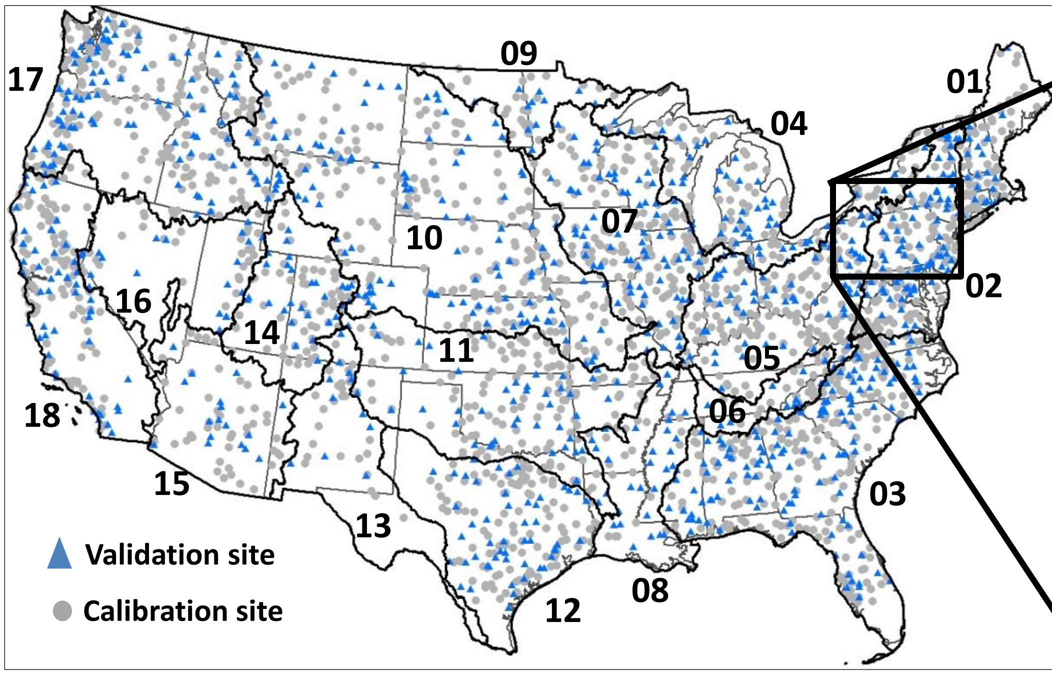
Home page: http://water.usgs.gov/nawqa/sparrow

SPARROW Streamflow Models

Mean Annual Streamflow (1997-2007)

1:500,000 RF1 streams (18 HUC-2 regions)

(1,778 calibration sites; 890 validation sites)

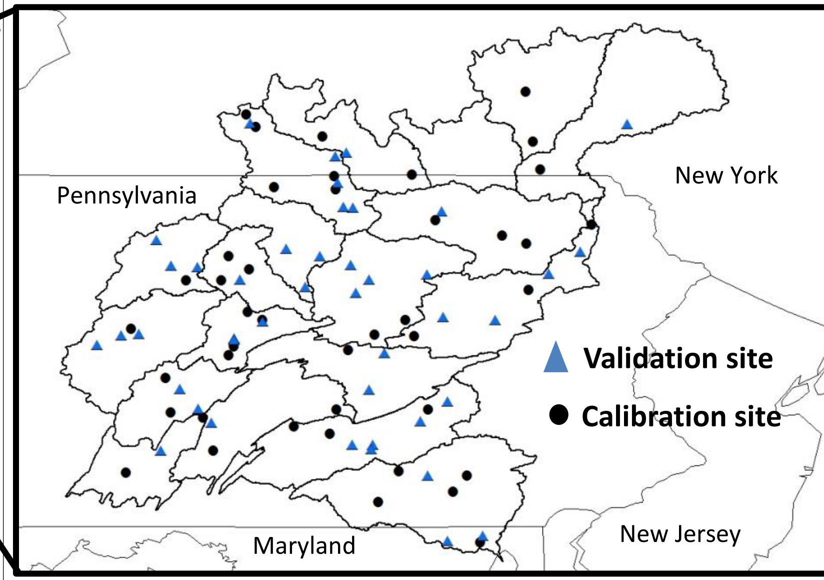


Susquehanna River Basin

Mean Seasonal Streamflow, 2001-08

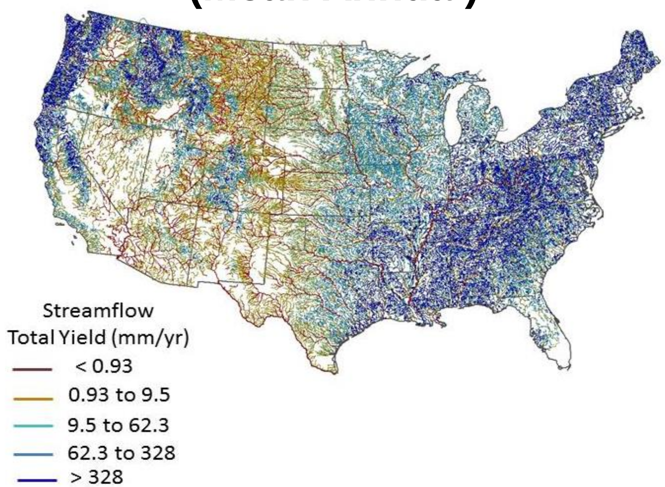
1:100,000 NHD streams

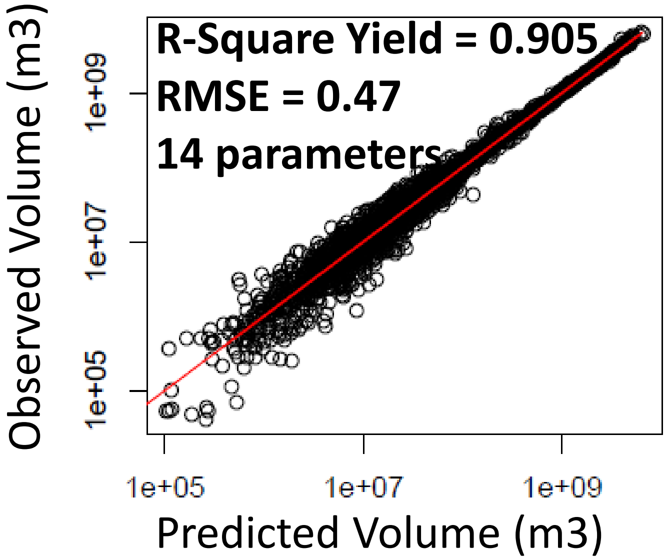
(19 HUC-8 watersheds) (85 sites)



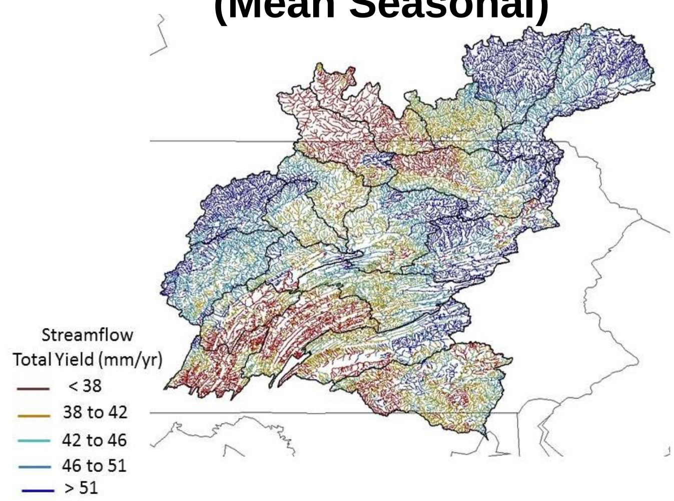
SPARROW Non-Hierarchical Bayesian Streamflow Models ("baseline" models)

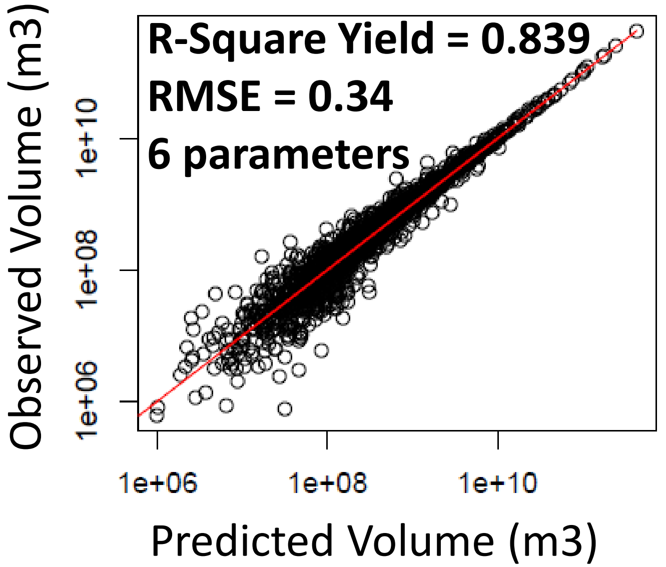
National – RF1
(Mean Annual)





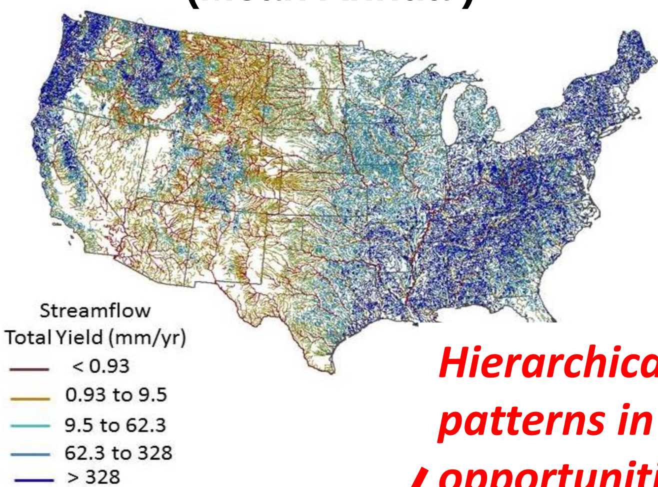
Susquehanna River Basin - NHD
(Mean Seasonal)



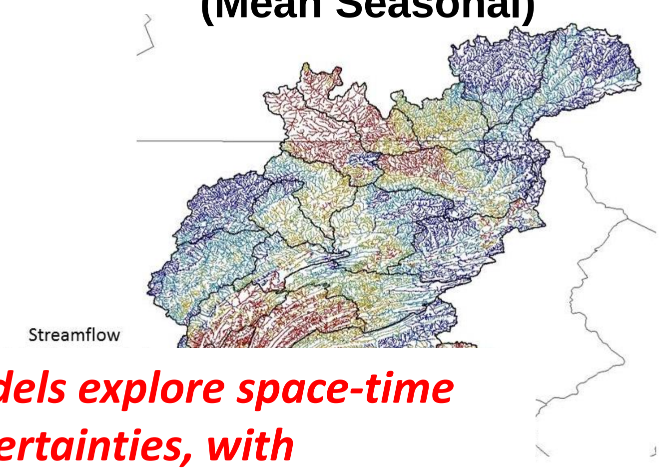


SPARROW Non-Hierarchical Bayesian Streamflow Models (“baseline” models)

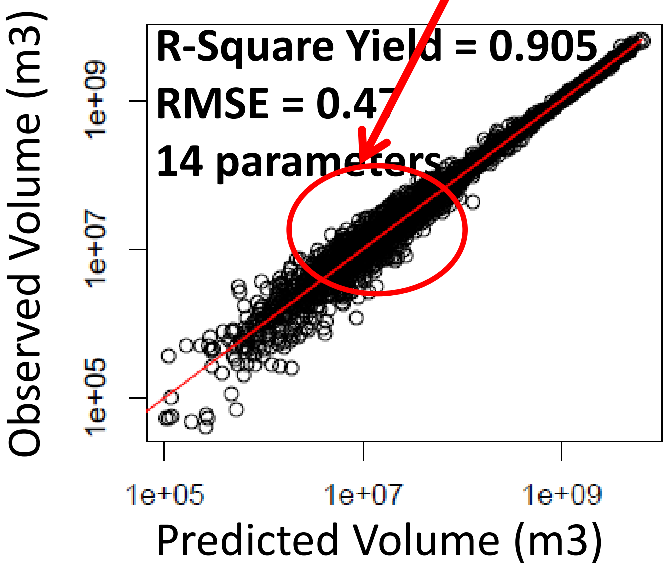
National – RF1
(Mean Annual)

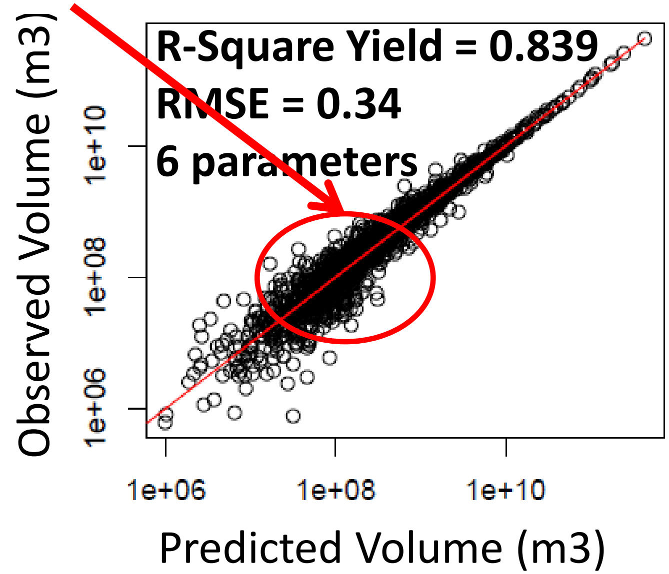


Susquehanna River Basin - NHD
(Mean Seasonal)



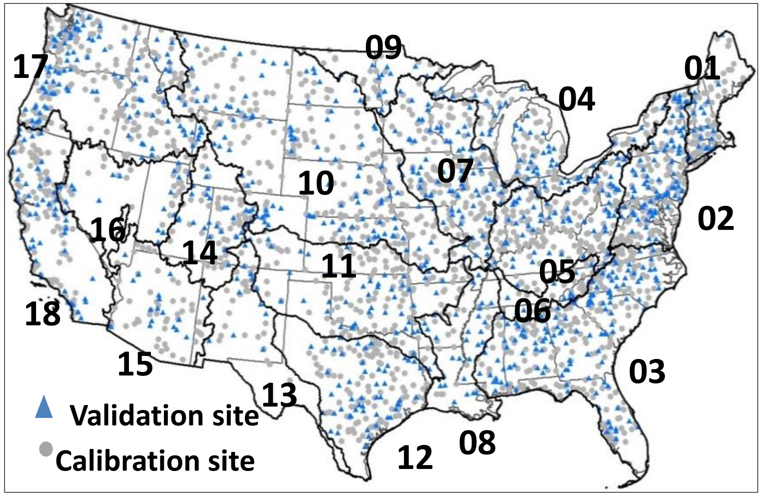
Hierarchical Bayesian models explore space-time patterns in the model uncertainties, with opportunities to reduce bias and increase precision



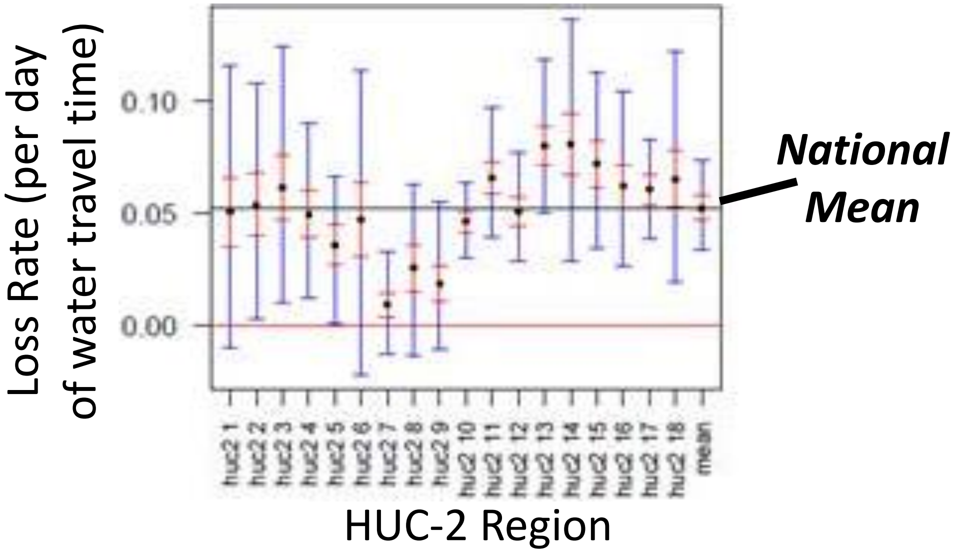


Bayesian SPARROW R Components

- Hierarchical rate coefficients for user-specified regions (discrete)

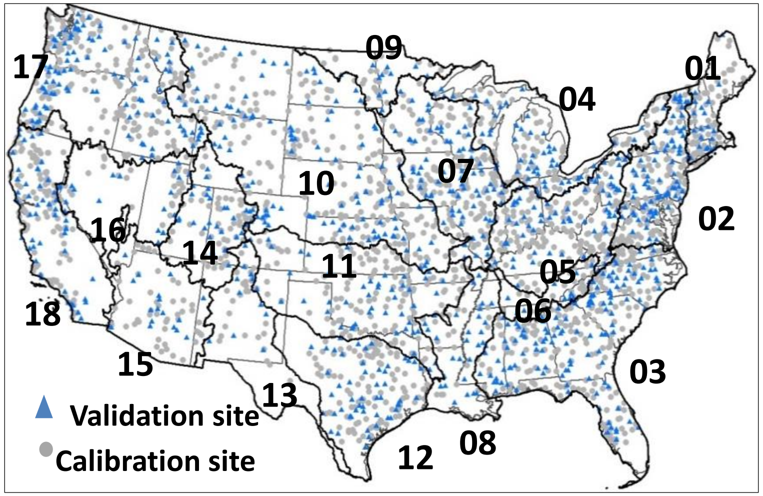


In-Stream Loss Rate Coefficient

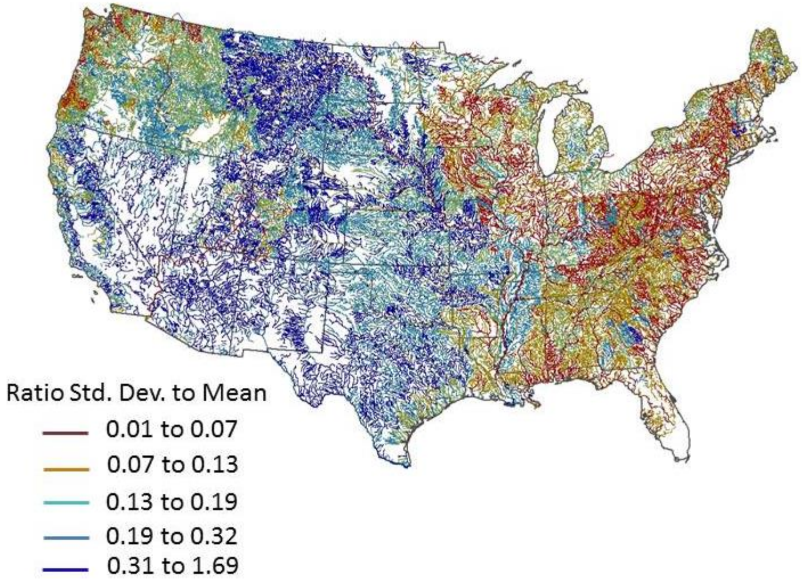


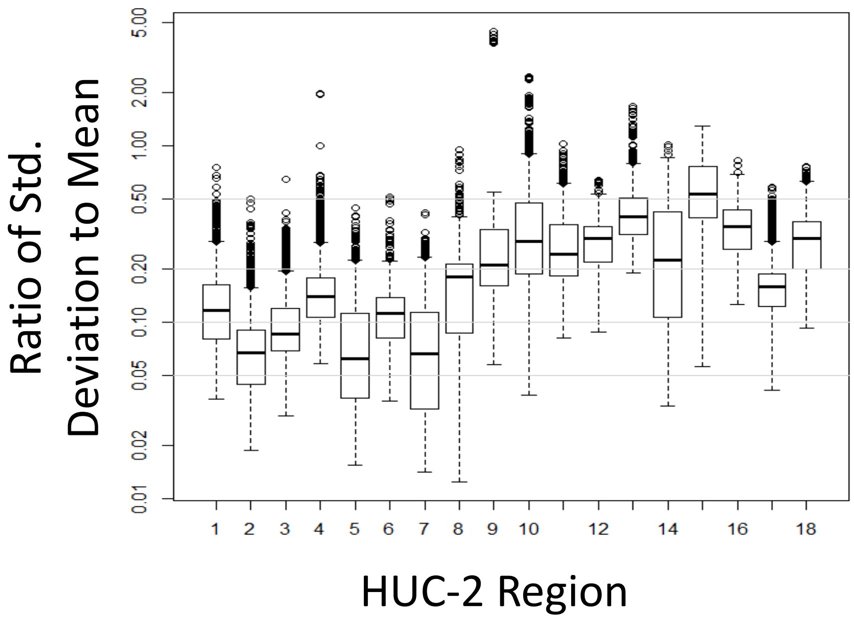
Bayesian SPARROW R Components

- Hierarchical rate coefficients for user-specified regions (discrete)
- Hierarchical error variance for user-defined regions (discrete, continuous)



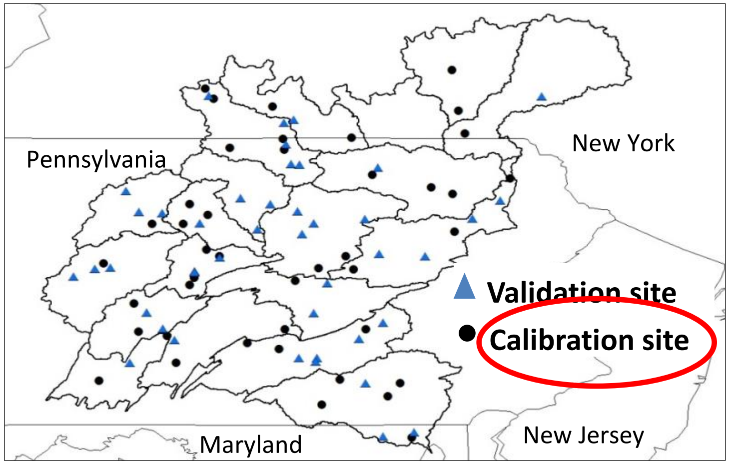
Regional Differences in Prediction Error



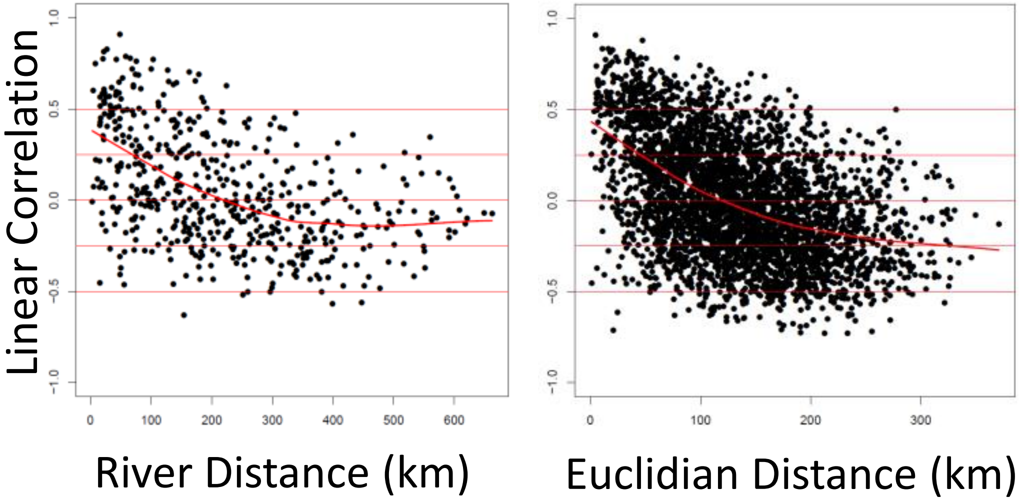


Bayesian SPARROW R Components

- Hierarchical rate coefficients for user-specified regions (discrete)
- Hierarchical error variance for user-defined regions (discrete, continuous)
- State-space option to quantify measurement and process uncertainties over space/time

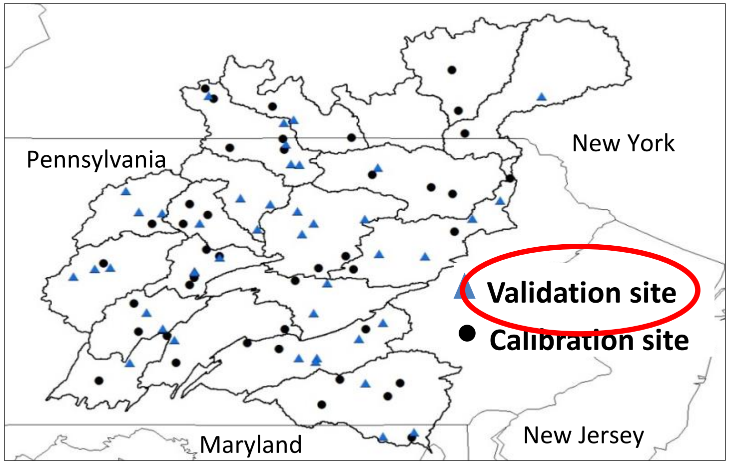


Spatial correlation evident in the process uncertainties among calibration station pairs



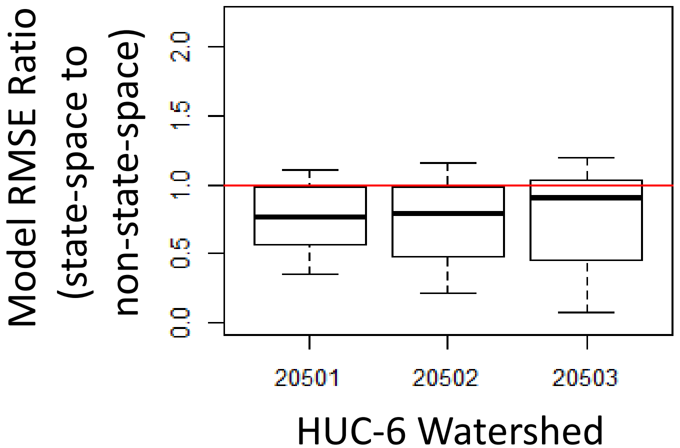
Bayesian SPARROW R Components

- Hierarchical rate coefficients for user-specified regions (discrete)
- Hierarchical error variance for user-defined regions (discrete, continuous)
- State-space option to quantify measurement and process uncertainties over space/time



Updated predictions that include process uncertainties show improved accuracy

Highest Accuracy for Summer



Process uncertainties include prominent effects of baseflow/groundwater

Bayesian SPARROW R Components

- Hierarchical rate coefficients for user-specified regions (discrete)
- Hierarchical error variance for user-defined regions (discrete, continuous)
- State-space option to quantify measurement and process uncertainties over space/time
- Additional model refinements under development:
  - Seasonal spatial covariance terms for adjacent catchments
  - Spatial smoothing of process errors for updated predictions (Kalman filtering terms)

Conclusions: A Novel Use of R with SPARROW across Regional and Continental Scales

Use of hierarchical Bayesian methods with SPARROW:

- Improved understanding of space-time variability in model coefficients and prediction uncertainties
  - Wide-ranging improvements in prediction accuracy (small to ~50%)
  - Enhanced understanding of spatial and temporal variability in process components
- Improved understanding of large-scale RStan performance across variable time-space scales: *Results suggest that seasonal models of >~70,000 km^2 and >10 yrs. will require NHD river network thinning*

Planned library releases of SPARROW R (GitHub, CRAN):

- Classic SPARROW - steady state non-Bayesian version (end 2016)
  SAS to R translation plus additional enhancements
- Bayesian version of classic SPARROW (2017)