



# National Guidelines for Developing and Documenting Surrogate Regression Models to Compute Continuous Water-Quality Concentrations

**Teresa Rasmussen,  
Patrick Rasmussen,  
US Geological Survey,  
Kansas Water Science Center**

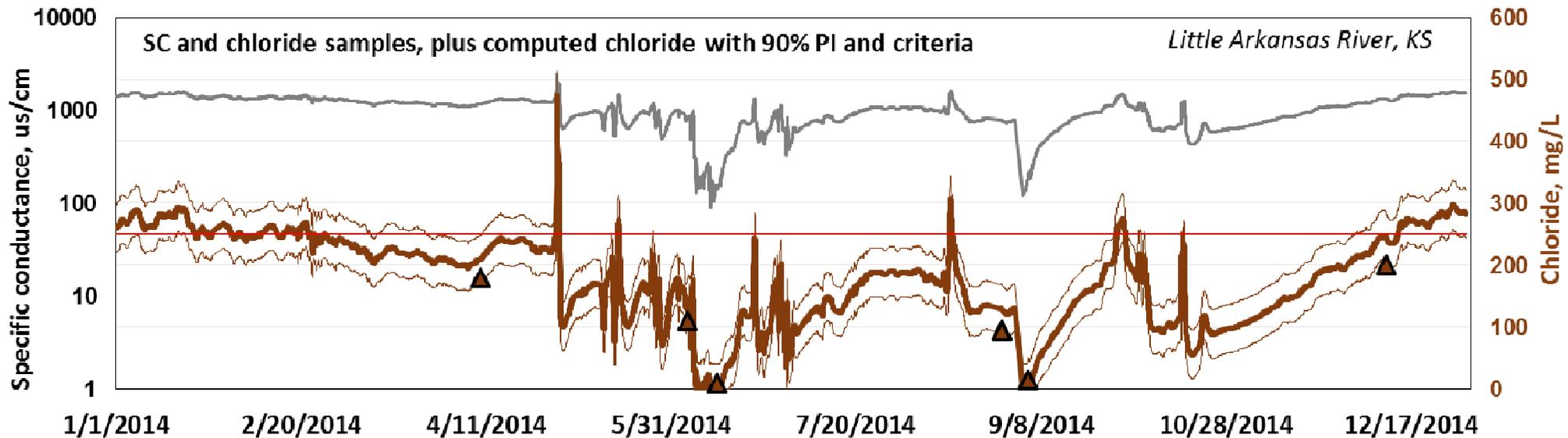
**Charles Crawford,  
US Geological Survey,  
National Water Quality Program**

With substantial contributions from Donna Myers, Robert Mason, Andy Ziegler, Bob Hirsch, Greg Schwarz, Steve Corsi, Mark Landers, Cherie Miller, Dale Robertson, and others

# Value of real-time continuous data

Direct measurements of water-quality (specific conductance, turbidity, dissolved oxygen, pH, water temperature, and others)

- Instant information
- Better characterization of variability – hourly, daily, monthly, annually
- Improved scientific understanding

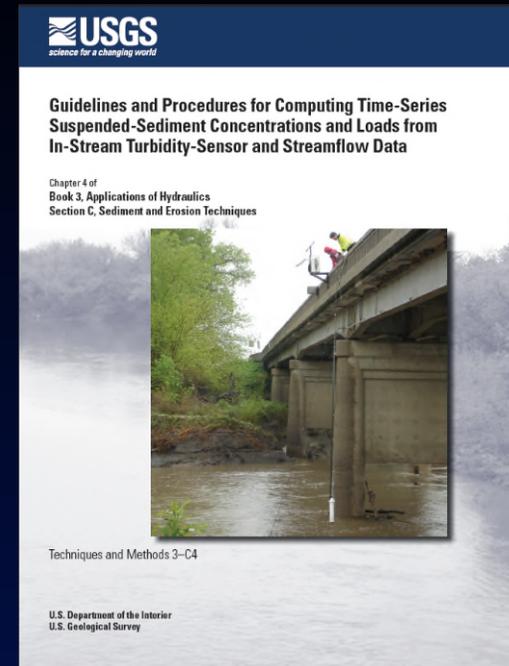


As surrogate for constituents of particular interest that cannot easily be measured directly (sediment, dissolved solids, nutrients, indicator bacteria)

# Turbidity and suspended sediment surrogate methods

Rasmussen and others, 2009

- Methods for using continuous in-stream turbidity and streamflow to compute continuous suspended sediment concentrations (SSCs) and loads
- Guidelines for collecting data and developing regression models



Landers and others, in press

- Methods for using acoustic indices from acoustic Doppler velocity meters (ADVMS) backscatter data to compute SSCs and loads
- Extends utility of ADVMs used for streamflow velocity; higher temporal resolution



From Wood and Teasdale, 2013

# Examples of water-quality surrogates

Surrogate	Water-quality constituent
Turbidity	Total suspended solids, total nitrogen, organic nitrogen, total phosphorus, total organic carbon, indicator bacteria
Specific conductance	Dissolved solids, alkalinity, chloride, calcium, magnesium, sodium, potassium, orthophosphate, atrazine

Baldwin and others, 2012; Chanat and others, 2013; Christensen and others, 2000; Galloway, 2014; Miller and others, 2007; Miller and others, 2013; Rasmussen and others, 2005; Ryberg, 2006; Schaepe and others, 2014; Stone and others, 2013; Stone and Graham, 2014; Wood and Etheridge, 2011.

# **Effort is underway to publish USGS methods report and release policy guidance for water-quality surrogates**

## **Purpose**

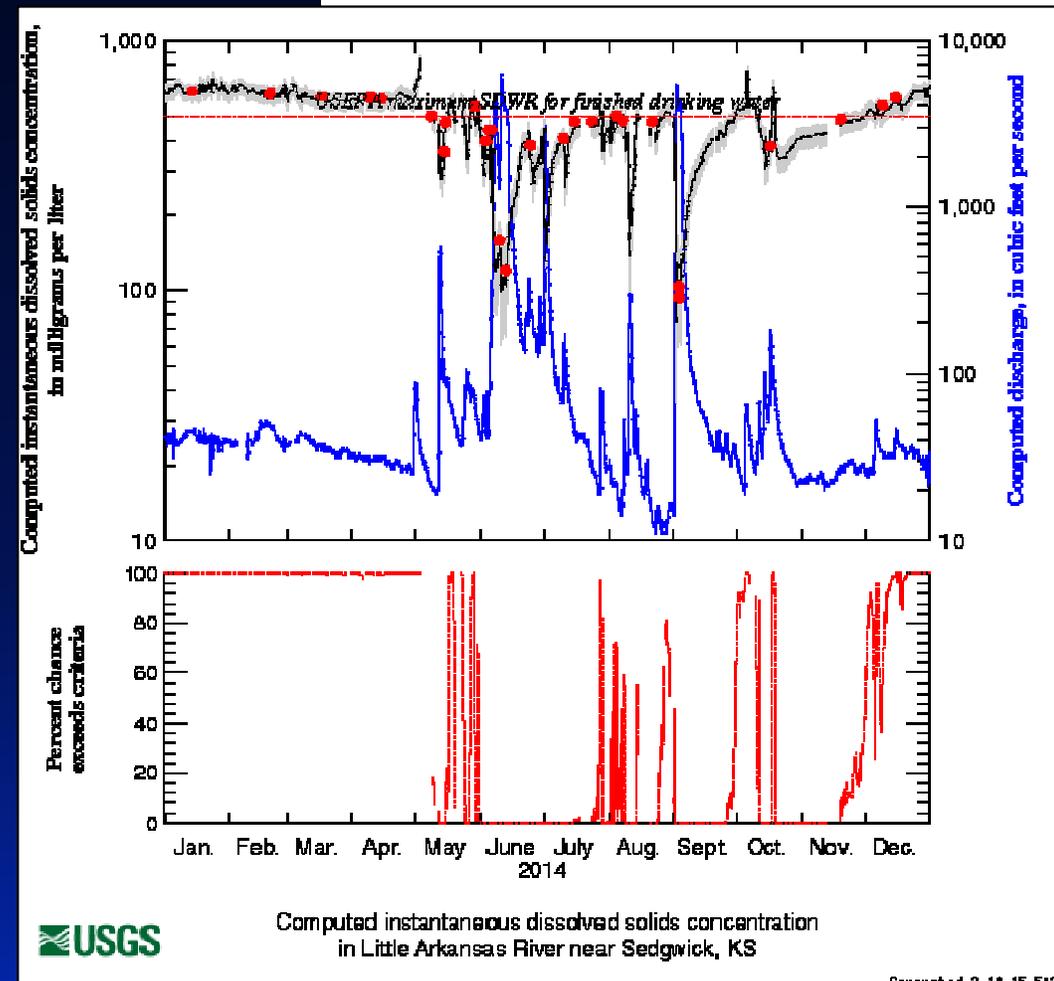
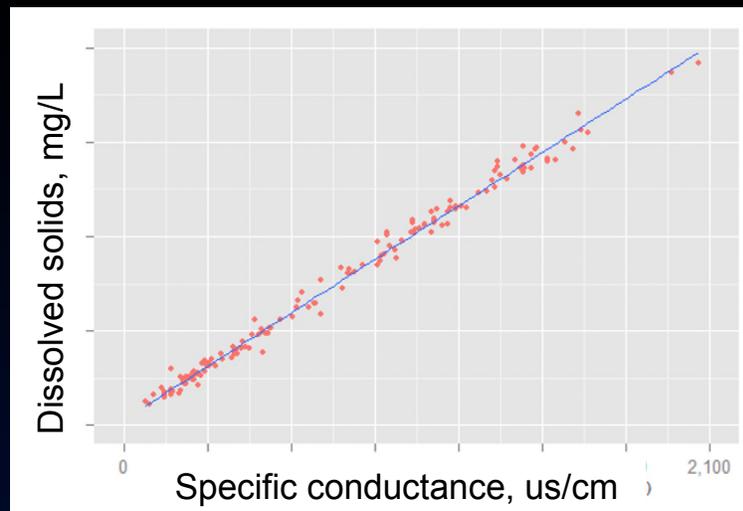
- **Provide consistent approach to model development**
- **Describe documentation and archiving requirements**
- **Meet Fundamental Science Practices without an interpretive report**

## **Benefits to USGS and others**

- **Consistent, clear, and streamlined approach**
- **Documentation of methods**
- **Better characterization of variability for important water-quality constituents**
- **Make models and data easily available to the public**
- **More timely release of information because publication not needed if methods are followed**

# General regression model approach

- Install water-quality monitors at streamflow gages and transmit data in real time
- Collect water samples
- Develop regression models using samples and monitor data
- Display computed data on the Web
- Continue sampling to verify models



## **Sounds easy enough, but...**

- **How many samples are needed?**
- **Over what period of time, and at what frequency should samples be collected?**
- **During what hydrologic conditions should samples be collected?**
- **What surrogates might be important for the constituent of interest?**
- **Is it okay to collect multiple samples during 1 runoff event?**
- **Where to start when building a model?**
- **How should non-detects be handled?**
- **How should outliers be treated?**
- **How to choose between the 2 best models?**
- **What statistical tests and plots are best for evaluating models?**
- **Is it okay to estimate values outside the model calibration range?**
- **How should an existing model be evaluated for performance?**
- **How should an existing model be updated as new samples are collected?**

# USGS Techniques and methods report

## Applications:

- Linear (and log-linear) regression models using instantaneous-value surrogate data to compute instantaneous-value water-quality concentration data
- Waters for which methods have been published for continuous (surrogate) data collection - rivers, streams, beaches

## Does not apply when:

- Non-linear or non-parametric statistical approaches are used
- Standard methods for data collection have not been published
- Data sets include more than 5% non-detections

# Methods – collection of continuous (surrogate) data

1. Follow published methods including
  - USGS National Field Manual – field methods
  - Wagner and others – multiparameter monitors
  - Pellerin and others – optical sensors
2. Follow requirements for timely review of data
3. Document operating range of sensors



# Methods – collection of discrete data



**1. Follow field methods described in USGS National Field Manual**

**2. Minimum of 48 samples recommended for model with 2 explanatory variables; more samples needed for additional explanatory variables**

- Half fixed interval, half runoff (or targeting other source of variability)
- Sample collection over 3 years preferred; 2-6 years okay
- Sample range of conditions

**3. Ensure data meet quality assurance objectives**

# Methods – regression model development

## 1. Use linear regression analysis; understand the basic assumptions

- Samples representative of population
- Residual errors have constant variance and are normally distributed

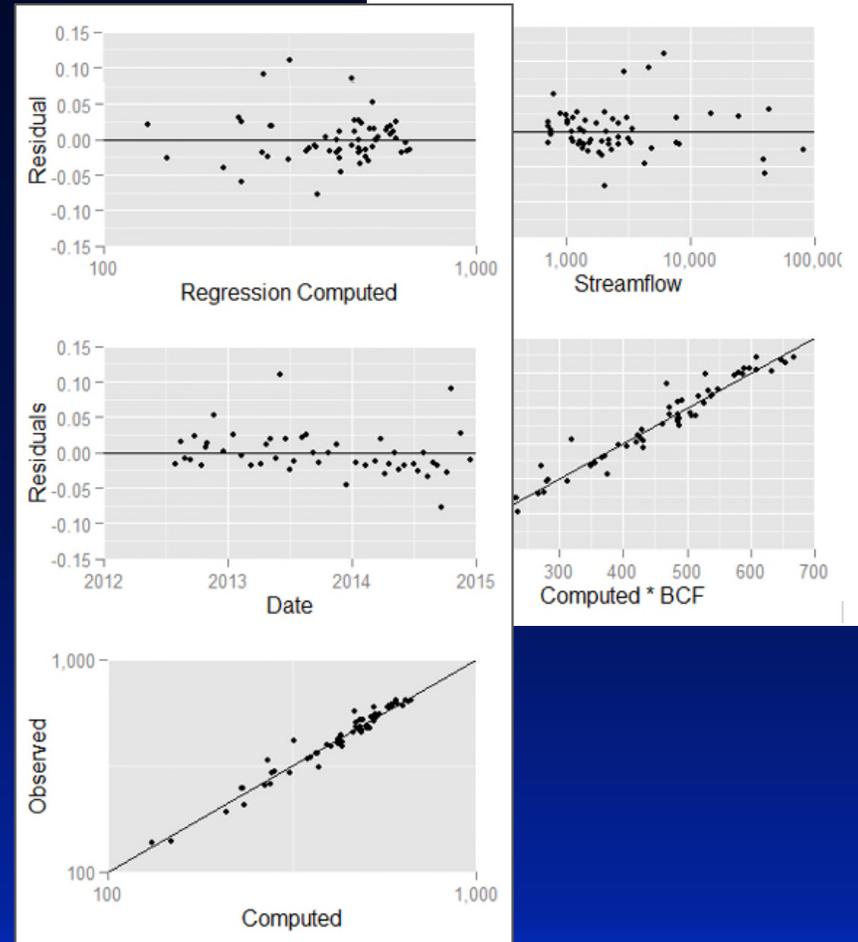
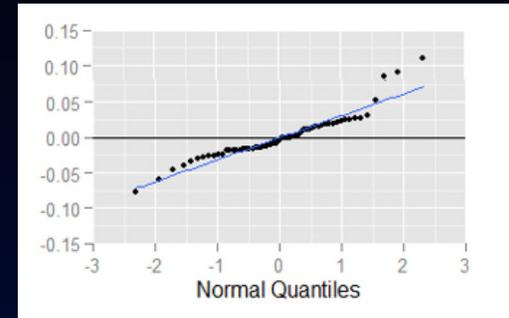
## 2. Use recommended statistical diagnostics and graphs to evaluate models

- Residual plots, time series, boxplots, model standard percentage error (MSPE), prediction error sum of squares (PRESS), Mallow's C, others
- R package will be available

## 3. Exclude outliers only when shown to be errors (or rare circumstance)

## 4. If more than 2-5% of data are non-detects, use different approach (Tobit)

## 5. Include transformation bias correction



# Methods – model documentation

1. Model Archive Summary (MAS) includes:
  - Description of model and model-building decisions
  - Model diagnostics and graphs
  - Link to dataset
2. MAS is submitted to 2 qualified (experienced in regression models) colleague reviewers
3. After approval, model is archived in an official data repository
4. R package available to produce standard MAS

# MA Exa

## Models Considered

Co	Model Formula	Number of Variables	Standard Error	R2	Adjusted R2	Cp	PRESS	VIF	MSPE
	logSS ~ logTURB	1	0.1595	92.76	92.65	21.06	1.884	<NA>	± 38
	logSS ~ TURB	1	0.3693	61.18	60.62	404.9	10.44	<NA>	± 96
	logSS ~ logQ	1	0.4443	43.82	43.01	615.9	14.44	<NA>	± 120
	logSS ~ logTURB + logSC	2	0.1548	93.28	93.08	16.69	1.802	1.742	± 36
	logSS ~ logTURB + SC	2	0.1548	93.28	93.08	16.75	1.843	2.008	± 36
	logSS ~ TURB + logTURB	2	0.156	93.17	92.97	17.98	1.875	2.449	± 37
	logSS ~ logTURB + logQ + logSC	3	0.1434	94.32	94.07	6.046	1.541	1.947	± 34
	logSS ~ logTURB + logQ + SC	3	0.1491	93.86	93.58	11.68	1.706	2.268	± 35
	logSS ~ logTURB + Q + logSC	3	0.1519	93.62	93.34	14.55	1.736	1.886	± 36
	logSS ~ TURB + logTURB + logQ + logSC	4	0.1401	94.66	94.33	3.944	1.509	2.558	± 33
	logSS ~ logTURB + logQ + SC + logSC	4	0.1441	94.35	94	7.743	1.61	2.43	± 34
	logSS ~ logTURB + Q + logQ + logSC	4	0.1443	94.33	93.99	7.877	1.573	2.757	± 34
	logSS ~ TURB + logTURB + logQ + SC + logSC	5	0.1402	94.73	94.33	5.044	1.563	2.652	± 33
	logSS ~ TURB + logTURB + Q + logQ + logSC	5	0.1411	94.66	94.25	5.896	1.541	2.582	± 33
	logSS ~ logTURB + Q + logQ + SC + logSC	5	0.1452	94.35	93.91	9.7	1.639	2.889	± 34
	logSS ~ TURB + logTURB + Q + logQ + SC + logSC	6	0.1412	94.74	94.24	7	1.592	2.752	± 33

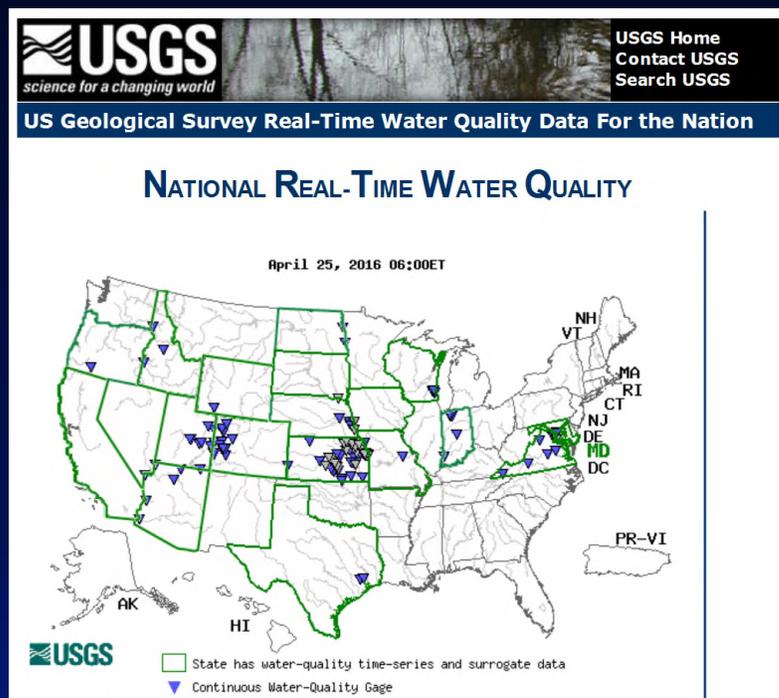
## Model-Calibration Data Set

	Date	logSS	logTURB	logQ	SS	TURB	Q	Computed logSS	Computed SS	Residual	Normal Quantiles	Censored Values
0												
1	2004-07-26	2.489	2.301	3.8	308	200	6310	2.608	434	-0.119	-0.983	--
2	2005-01-28	1.94	1.806	1.828	87	64	67.25	2.12	141	-0.18	-1.64	--
3	2005-03-25	2.417	2.204	2.843	261	160	696	2.493	333.3	-0.0767	-0.597	--
4	2005-05-11	1.914	1.623	1.404	82	42	25.33	1.949	95.27	-0.0357	-0.106	--
5	2005-05-26	3.161	2.568	2.822	1450	370	664	2.804	680.9	0.358	1.79	--
6	2005-06-07	2.55	2.255	2.694	355	180	493.9	2.532	364.4	0.0181	0.436	--
7	2005-06-14	2.433	2.079	3.587	271	120	3863	2.411	276	0.0215	0.475	--
8	2007-01-10	1	0.8195	0.7317	10	6.6	5.392	1.241	18.63	-0.241	-1.79	--
9	2007-03-12	1.813	1.633	1.041	65	42.96	11	1.946	94.48	-0.133	-1.17	--
10	2007-03-21	2.303	2.111	1.146	201	129.1	14	2.357	243.8	-0.0543	-0.436	--
11	2007-03-27	2.489	2.38	1.896	308	240	78.75	2.612	438.3	-0.124	-1.1	--
12	2007-04-02	2.819	2.477	3.43	659	300	2690	2.746	596.4	0.0729	0.824	--
13	2007-04-18	2.389	2.146	2.611	245	140	408.7	2.436	292.2	-0.047	-0.213	--
14	2007-05-08	2.459	2.301	3.719	288	200	5238	2.605	431.3	-0.146	-1.32	--
15	2007-05-10	2.114	2.079	3.686	130	120	4852	2.415	278.1	-0.301	-2.01	--
16	2007-05-24	2.609	2.419	3.936	406	262.5	8637	2.713	553.1	-0.105	-0.824	--
17	2007-05-25	2.486	2.204	4.037	306	160	10900	2.533	365.2	-0.0474	-0.249	--
18	2007-07-11	2.623	2.204	3.129	420	160	1345	2.503	340.7	0.12	0.983	--
19	2007-08-13	1.716	1.477	1.699	52	30	50	1.835	73.11	-0.119	-0.874	--
20	2007-09-05	1.785	1.431	1.204	61	27	16	1.779	64.34	0.00634	0.249	--
21	2007-12-12	2.709	2.322	3.219	512	210	1656	2.607	432.7	0.103	0.927	--
22	2008-03-04	3.276	2.982	3.409	1890	960	2566	3.177	1608	0.0997	0.874	--
23	2008-04-14	2.52	2.362	2.476	331	230	299	2.616	441.8	-0.0959	-0.683	--
24	2008-05-29	2.13	2.347	3.165	135	222.5	1463	2.626	452.8	-0.496	-2.39	--



# Methods – data computation and dissemination

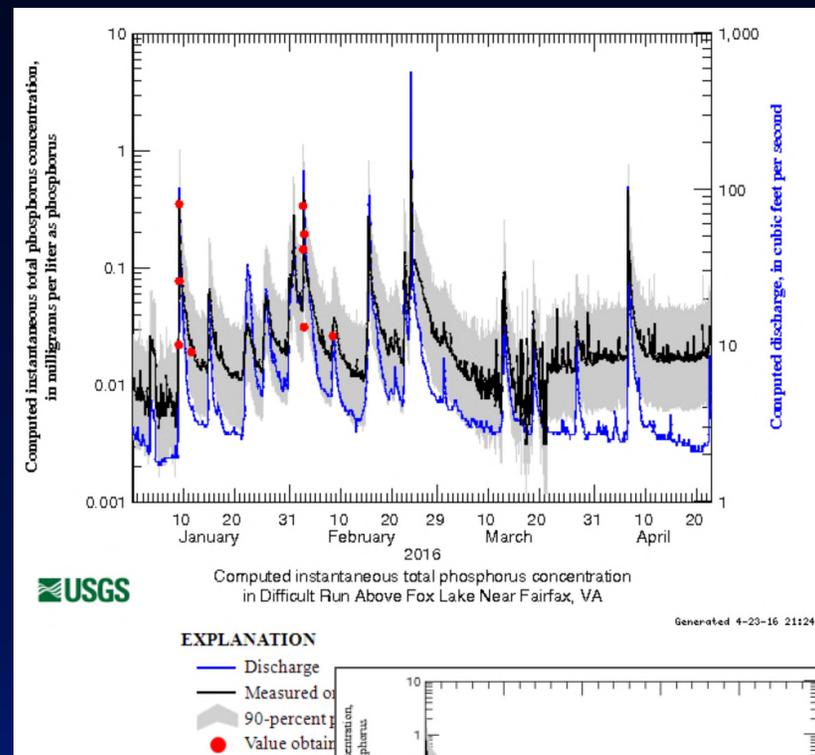
1. Via NRTWQ, project and other websites, reports
2. Compute load by multiplying computed unit-value concentration by the concurrent unit-value of streamflow, and summing over desired time period
3. No interpolation and limited extrapolation



<http://nrtwq.usgs.gov>

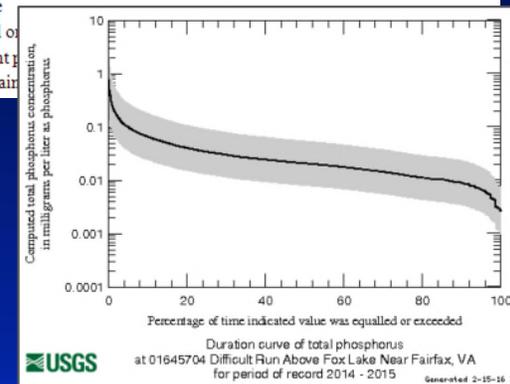


NRTWQ page displays water-quality concentrations and loads from surrogates in 19 states



Example: Difficult Run near Fairfax, VA

$$\ln TP = -4.813 + 0.547 \ln TURB + 0.107 \ln Q$$



Jastram, 2014

# Methods – reviewing and updating established models

1. Ongoing sample collection needed to verify model performance as long as model is being used
2. Minimum of 8 samples/year (4 fixed interval, 4 targeting variability)
2. Annual model review and documentation
3. Update model at least every 3 years
4. Timely evaluation of new data
5. Recommendations for evaluating values that fall outside model uncertainty bands
  - Evaluate samples with residuals of more than 2-3 sigmas; investigate cause, collect extra samples
  - Stop predictions if model not representing data

# Plan for water-quality surrogates

**Publish USGS T&M report to describe standard surrogate methods for water-quality constituents**

**Simultaneously release USGS technical memo describing policy**

- **Data computed following methods described in the T&M are considered non-interpretive data and can be disseminated without publishing a report**
- **Models must be documented and archived in accordance with requirements for model archiving**
- **Models and data must meet requirements for scientific data release (IM OSQI 2015-03)**

**Until these documents are released, published reports are required for using surrogate models other than suspended sediment (if published methods were followed).**



# References

- Baldwin, A.K., Graczyk, D.J., Robertson, D.M., Saad, D.A., and Magruder, Christopher, 2012, Use of real-time monitoring to predict concentrations of select constituents in the Menomonee River drainage basin, Southeast Wisconsin, 2008–9: U.S. Geological Survey Scientific Investigations Report 2012–5064, 18 p., plus six appendixes.
- Chanat, J.G., Miller, C.V., Bell, J.M., Majedi, B.F., and Brower, D.P., 2013, Summary and interpretation of discrete and continuous water-quality monitoring data, Mattawoman Creek, Charles County, Maryland, 2000–11: U.S. Geological Survey Scientific Investigations Report 2012–5265, 42 p.
- Christensen, V.G., Jian, Xiaodong, and Ziegler, A.C., 2000, Regression analysis and real-time water-quality monitoring to estimate constituent concentrations, loads, and yields in the Little Arkansas River, South-Central Kansas, 1995–99: U.S. Geological Survey Water-Resources Investigations Report 00–4126, 36 p.
- Galloway, J.M., 2014, Continuous water-quality monitoring and regression analysis to estimate constituent concentrations and loads in the Red River of the North at Fargo and Grand Forks, North Dakota, 2003–12: U.S. Geological Survey Scientific Investigations Report 2014–5064, 37 p.,
- Jastram, J.D., 2014, Streamflow, Water Quality, and Aquatic Macroinvertebrates of Selected Streams in Fairfax County, Virginia, 2007–12: U.S. Geological Survey Scientific Investigations Report 2014–5073, 68 p.
- Landers, M.N., Straub, T.D., Wood, M.S., and Domanski, M.M., in press, Sediment acoustics index method for computing continuous suspended-sediment concentrations: U.S. Geological Survey Techniques and Methods book \_\_, chap. \_\_, p.
- Miller, C.V., Gutiérrez-Magness, A.L., Feit Majedi, B.L., and Foster, G.D., 2007, Water quality in the Upper Anacostia River, Maryland: Continuous and discrete monitoring with simulations to estimate concentrations and yields, 2003–05: U.S. Geological Survey Scientific Investigations Report 2007–5142, 43 p.
- Pellerin, B.A., Bergamaschi, B.A., Downing, B.D., Saraceno, J.F., Garrett, J.A., and Olsen, L.D., 2013, Optical techniques for the determination of nitrate in environmental waters: Guidelines for instrument selection, operation, deployment, maintenance, quality assurance, and data reporting: U.S. Geological Survey Techniques and Methods 1–D5, 37 p.
- Rasmussen, T.J., Ziegler, A.C., and Rasmussen, P.P., 2005, Estimation of Constituent Concentrations, Densities, Loads, and Yields in Lower Kansas River, Northeast Kansas, Using Regression Models and Continuous Water-Quality Monitoring, January 2000 Through December 2003: U.S. Geological Survey Scientific Investigations Report 2005-5165, 117 p
- Rasmussen, P.P., Gray, J.R., Glysson, G.D., Ziegler, A.C., 2009, Guidelines and procedures for computing time-series suspended-sediment concentrations and loads from in-stream turbidity-sensor and streamflow data: U.S. Geological Survey Techniques and Methods book 3, chap. C4, 53 p.
- Ryberg, K.R., 2007, Continuous water-quality monitoring and regression analysis to estimate constituent concentrations and loads in the Sheyenne River, North Dakota, 1980–2006: U.S. Geological Survey Scientific Investigations Report 2007–5153, 22 p.
- Schaepe, N.J., Soenksen, P.J., and Rus, D.L., 2014, Relations of water-quality constituent concentrations to surrogate measurements in the lower Platte River corridor, Nebraska, 2007 through 2011: U.S. Geological Survey Open-File Report 2014–1149, 16 p.,
- Stone, M.L., Graham, J.L., and Gatotho, J.W., 2013, Model documentation for relations between continuous real-time and discrete water-quality constituents in the North Fork Ninnescah River upstream from Cheney Reservoir, south-central Kansas, 1999–2009: U.S. Geological Survey Open-File Report 2013–1014, 101 p.
- Stone, M.L., and Graham, J.L., 2014, Model documentation for relations between continuous real-time and discrete water-quality constituents in Indian Creek, Johnson County, Kansas, June 2004– May 2013: U.S. Geological Survey Open-File Report 2014–1170, 70 p.
- Wagner, R.J., Boulger, R.W., Jr., Oblinger, C.J., and Smith, B.A., 2006, Guidelines and standard procedures for continuous water-quality monitors—Station operation, record computation, and data reporting: U.S. Geological Survey Techniques and Methods 1–D3, 51 p. + 8 attachments.
- Wood, M.S., and Teasdale, G.N., 2013, Use of surrogate technologies to estimate suspended sediment in the Clearwater River, Idaho, and Snake River, Washington, 2008–10: U.S. Geological Survey Scientific Investigations Report 2013-5052, 30 p.



# Questions?

**Teresa Rasmussen**

**USGS Kansas Water Science Center**

**Lawrence, Kansas**

**[rasmuss@usgs.gov](mailto:rasmuss@usgs.gov)**

**785-832-3576**