Quantifying Uncertainty: Adding Value to USGS Time-Series Water-Quality Data

Stewart Rounds
USGS
Oregon Water Science Center
Portland, OR

with help & insights from many others:

USGS:	Stacey Archfield, Janice Fulford, David Holtschlag, Brian Pellerin, Pat Rasmussen, Susan Wherry
YSI / Xylem:	Rob Ellison
Aquatic Informatics:	Stu Hamilton, Brian Gouge

Measurement Errors

Error

- The difference between a measured value and a known reference or true value
- Composed of random and predictable components
  - Predictable components are repeatable & correctable
- Many sources of error
  - Avoidable Errors

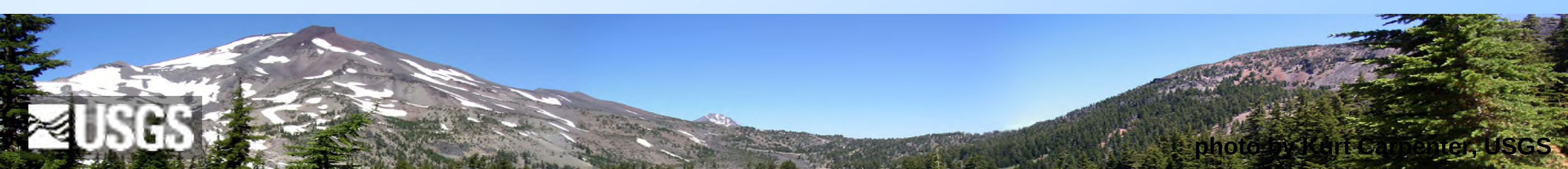
    User error, site-selection errors, site-installation errors
  - Unavoidable Errors

    Signal noise/variability, site variability, interferences

    Probe fouling, probe failure, calibration drift

    Deviation from lab verification samples or other references

    Surrogate model error



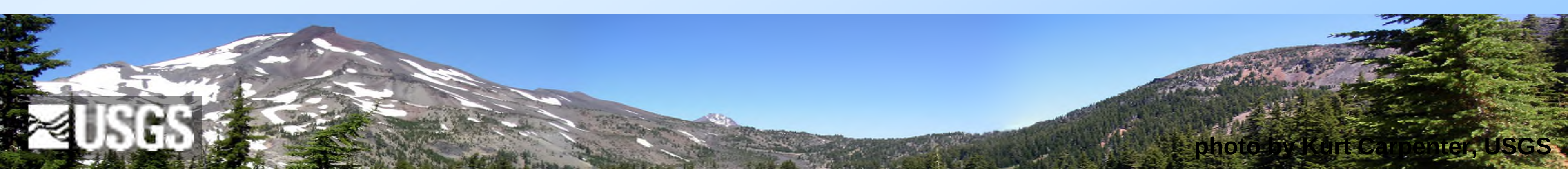
Error vs Uncertainty

Error

- The difference between a measured value and a known reference or true value
- Composed of random and predictable components
  - Predictable components are repeatable & correctable
- Many sources of error

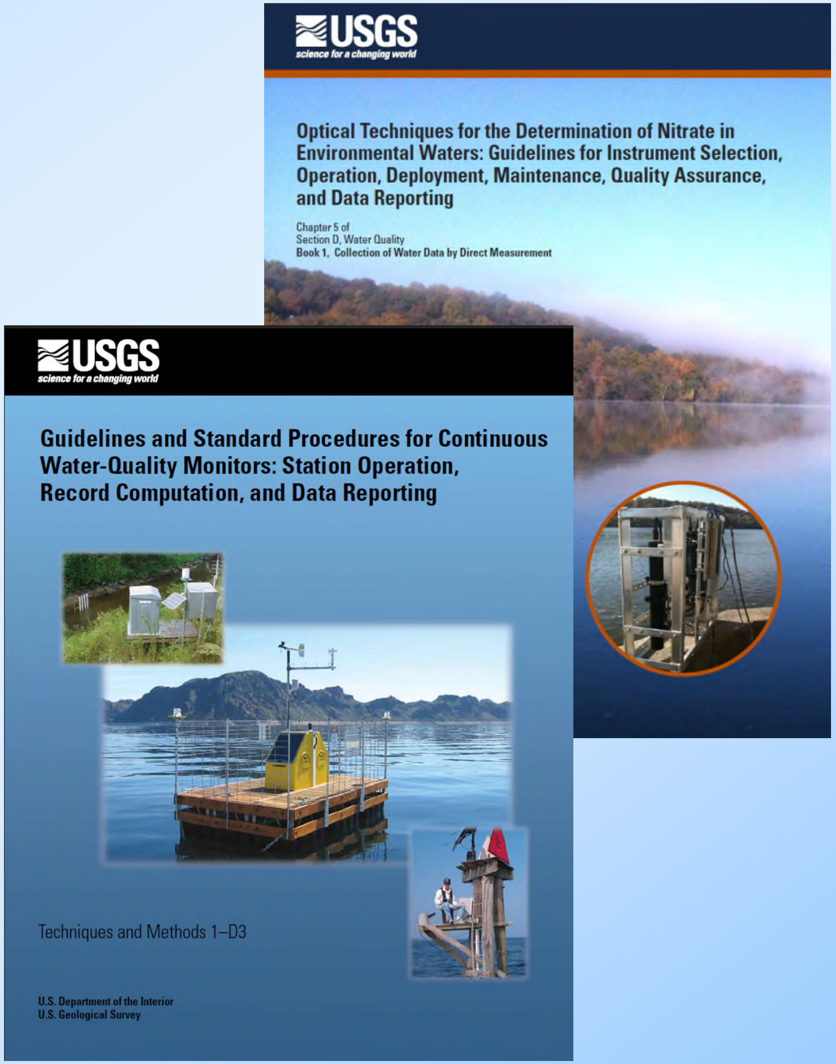
Uncertainty

- A description of the degree of accuracy of the final corrected data
  - Can be expressed in many ways, including statistical representations (std. dev.) or a simple half-width of a probable interval



Objectives

- Set forth general & usable methods for computing data uncertainty for water-quality time-series data, within a context that can be applied to other datasets.
- Produce a peer-reviewed report describing the methods (with examples) and recommendations for implementation and future work.



The Vision: Include Uncertainties with Data

Now:

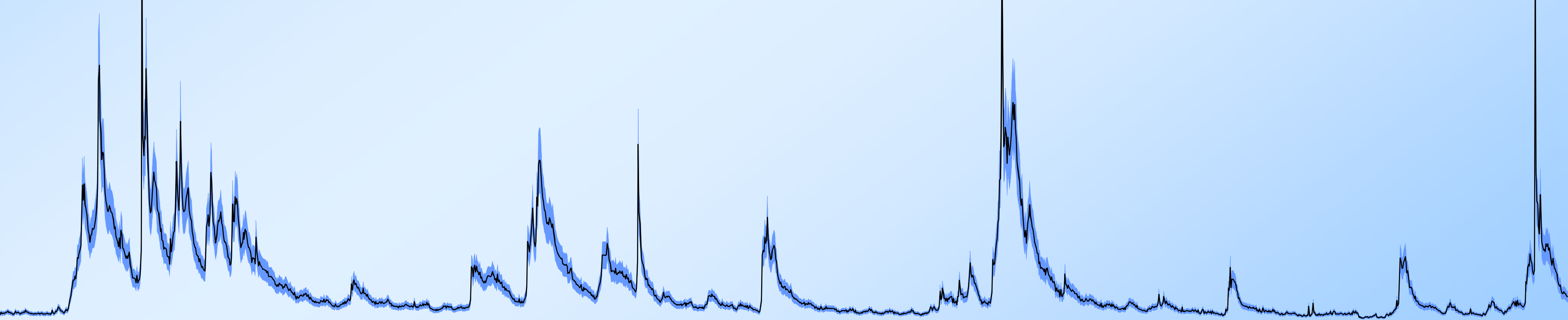
- **Time-series graphs from USGS show no uncertainty bands**
- **Data downloads are:**

  Date, Time, Time Zone, Value, Flag/Remark

Future:

- **Time-series graphs from USGS show uncertainty bands and comparison to standards w/ probability of exceedance**
- **Data downloads are:**

  Date, Time, Time Zone, Value, Uncertainty, Flag/Remark



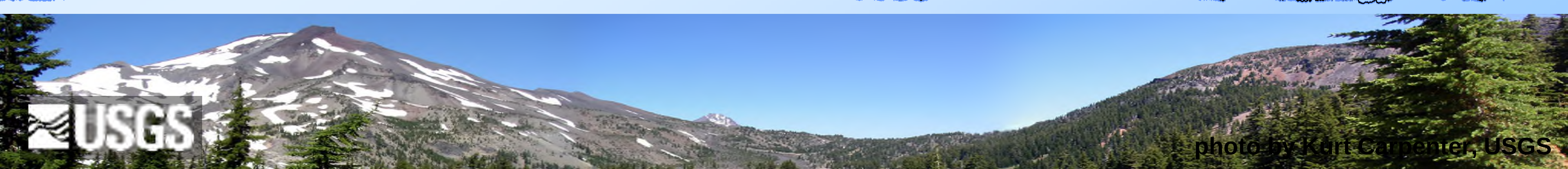
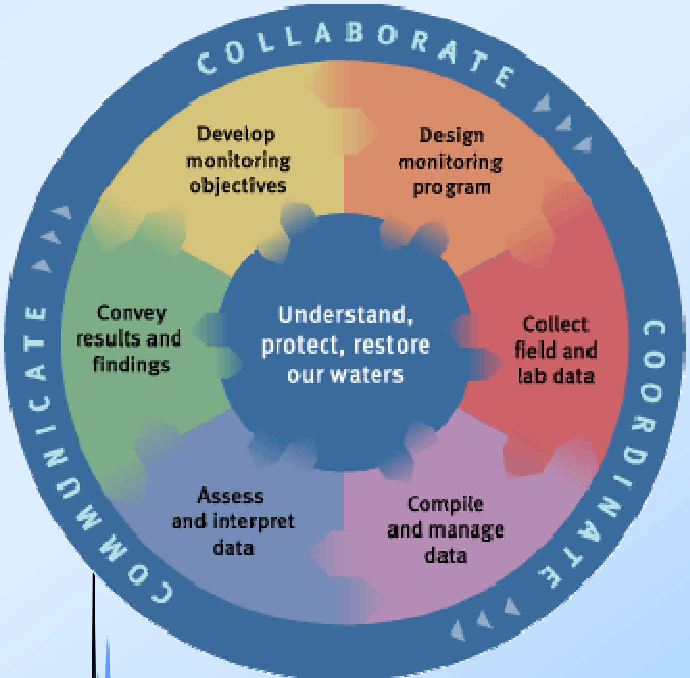


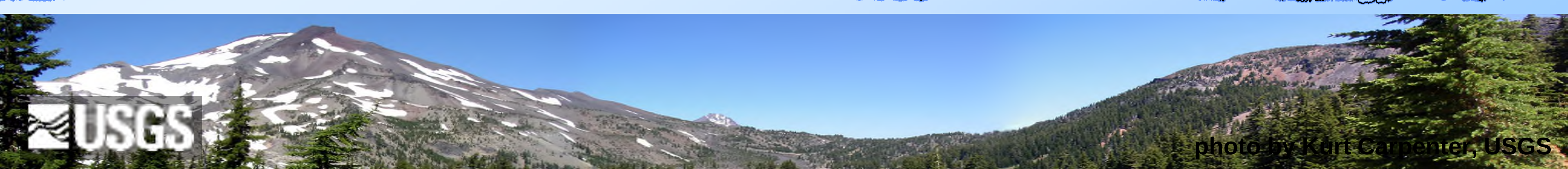
photo by Kurt Carpenter, USGS

Why Do We Care About Data Uncertainty?

1. Quantify uncertainty in order to manage it

- Identify & assess measurement errors in order to avoid, minimize, and correct for them
  (we already do this)
- Improve datasets and/or decrease costs by focusing resources where uncertainty is greatest
- Optimize data-collection and model building plans & protocols





Why Do We Care About Data Uncertainty?

2. Communicate the accuracy of datasets
→ Improves acceptance and appropriateness of use

Currently, USGS rates data:

- **Qualitatively**
  (Excellent / Good / Fair / Poor)
- **Based on:**
  1. **Sum of absolute values of all data corrections, and**
  2. **Assessment of hydrographer**
- **Ratings are not provided to the public with all data downloads**

Instead, data rounding is used as a poor substitute for a quantitative uncertainty.

```
Analysis performed Monday, 07-Jan-2013 08:44

Station Name:   FANNO CREEK AT DURHAM, OR
Station ID:     14206950
Parameter:      pH (std units)
PCODE:          00400
ADAPS DD:       6
Begin Date:     01-Oct-2009
End Date:       30-Sep-2010
Rating Method: Daily rating based on minimum unit-value rating for the day
DV Filter:      Rating provided regardless of existence of daily values

Verified 100.0% of 8731 computed data points, with a maximum
verification error of 0.02 std units.

SAC = Sum of the Absolute values of the Correct[illegible]
SAP = Sum of the Absolute values of the Pe[illegible] correcti[illegible]

The following data-qualit[illegible] were determ[illegible] on
the basis of the [illegible] values:

EXCE[illegible] 79.8% of the [illegible]es, 79.2% of [illegible]atings
 Crite[illegible]:  SAC <[illegible]d units
 Daily range[illegible]
   20[illegible] 2010/03/26
   201[illegible]/04/21 to 20[illegible]8
   2010/05/13 to 201[illegible]/08/24

GOOD -- 18.6% of the unit values, 18.6% of daily ratings
 Criteria:  0.2 std units < SAC <= 0.5 std units
 Daily ranges:
   2010/03/27 to 2010/04/20
   2010/04/29 to 2010/05/10
   2010/08/25 to 2010/09/24

FAIR -- 1.6% of the unit values, 2.2% of daily ratings
 Criteria:  0.5 std units < SAC <= 0.8 std units
 Daily ranges:
   2010/05/11 to 2010/05/12
   2010/09/25 to 2010/09/30

POOR -- 0.0% of the unit values, 0.0% of daily ratings
 Criteria:  0.8 std units < SAC <= 2 std units
 Daily ranges:
   none
```

The size of data corrections does not denote uncertainty. We can do better!

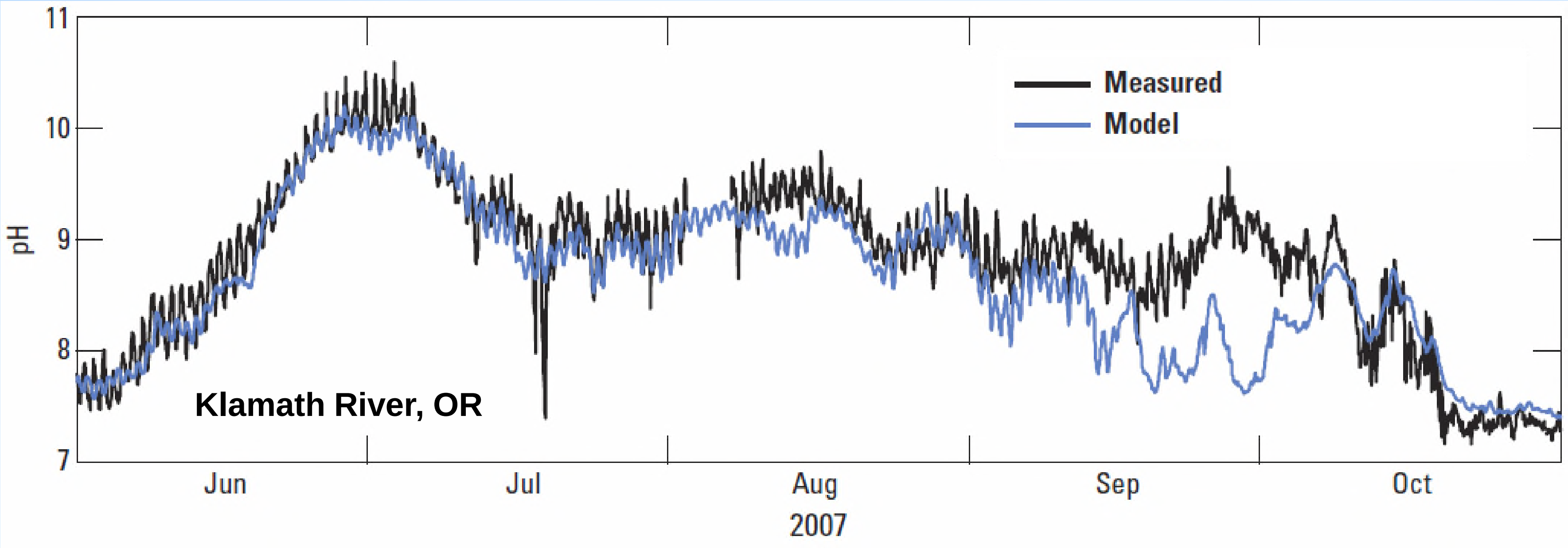
USGS

photo by Kurt Carpenter, USGS

Why Do We Care About Data Uncertainty?

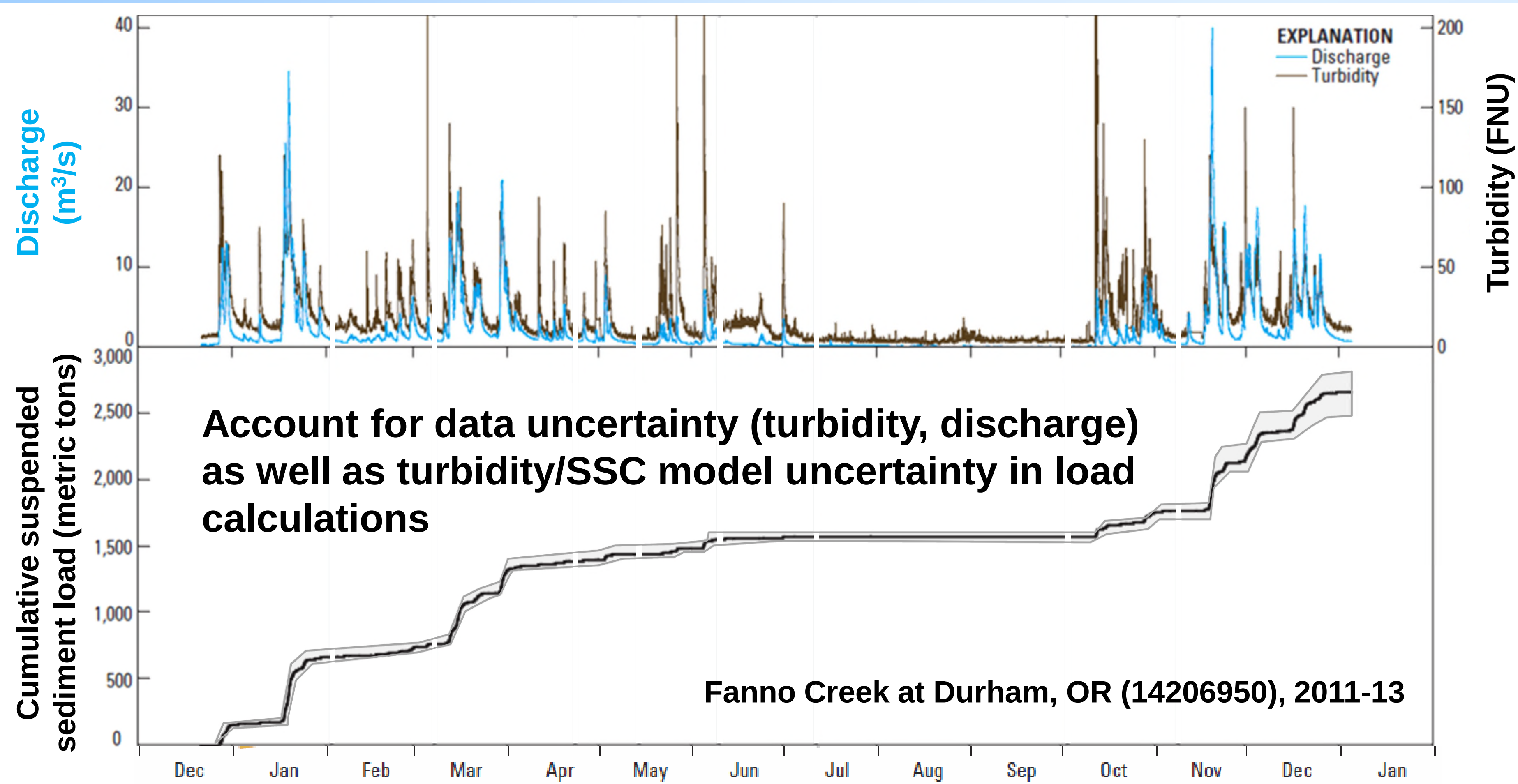
3. Improve & extend the value & applicability of data

- Model predictions can be compared to time-series data in a way that takes both model error and data uncertainty into account
  - Would help to determine the significance of model-data deviations



Why Do We Care About Data Uncertainty?

3. Improve & extend the value & applicability of data



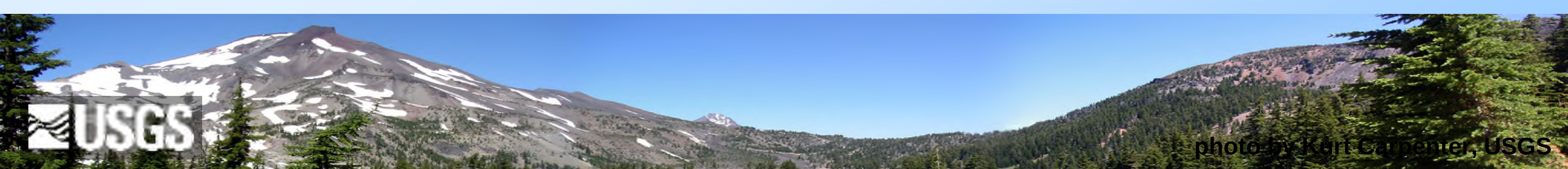
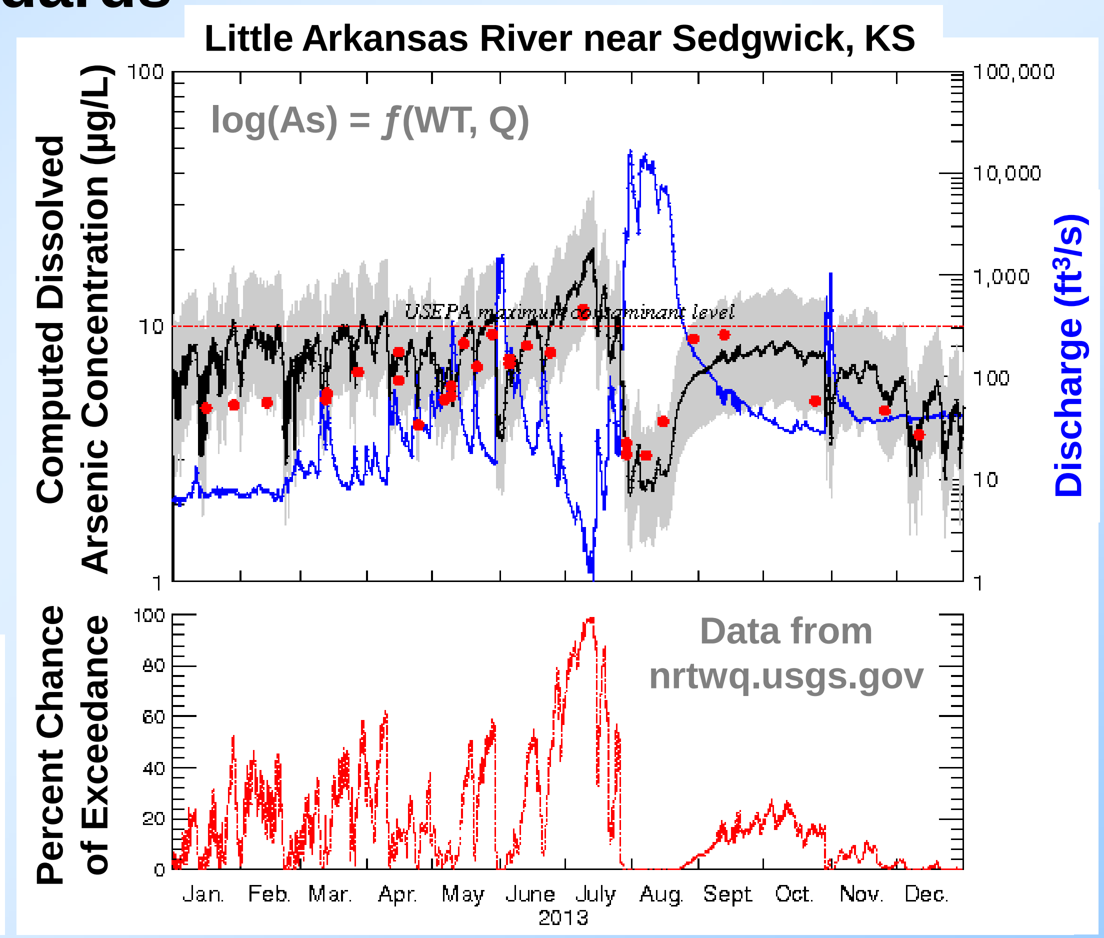


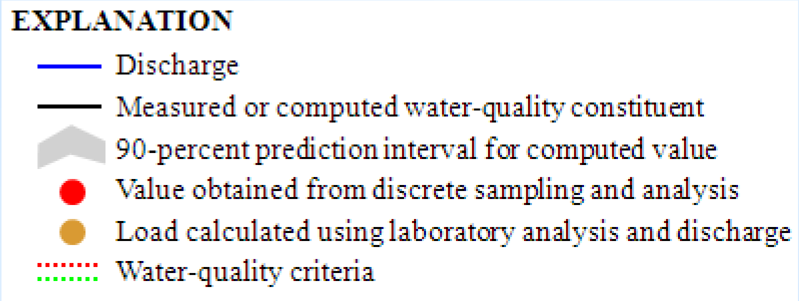
photo by Kurt Carpenter, USGS

Why Do We Care About Data Uncertainty?

4. Create a better framework for comparisons to benchmarks & standards

- **If uncertainties are known, we can compute the probability that a criterion is exceeded.**
- **Provides the basis for a new and better way to assess impairment and apply regulations**





Example – DOC Time-Series from FDOM

Fluorescent Dissolved Organic Matter

- Turner Cyclops FDOM probe
- FDOM → highly correlated with DOC
- DOC sample collection & laboratory analysis

USGS

photo by Kurt Carpenter, USGS

Methods With Great Potential

Rigorous statistical approach

- "The GUM"
  - Guide to the Expression of Uncertainty in Measurement by the Joint Committee for Guides in Metrology
  - Applied by Janice Fulford in the next presentation

Root Mean Square Error approach

- Simpler method of accounting for all sources of error
- Estimates a most probable value of the cumulative error
- Assumes independence of error sources
  (note the lack of correlation coefficients and cross-terms)

$$E_P = \sqrt{E_1^2 + E_2^2 + \cdots + E_n^2} = \sqrt{\sum_i E_i^2}$$

E_P is the combined probable error; E_i are the component errors

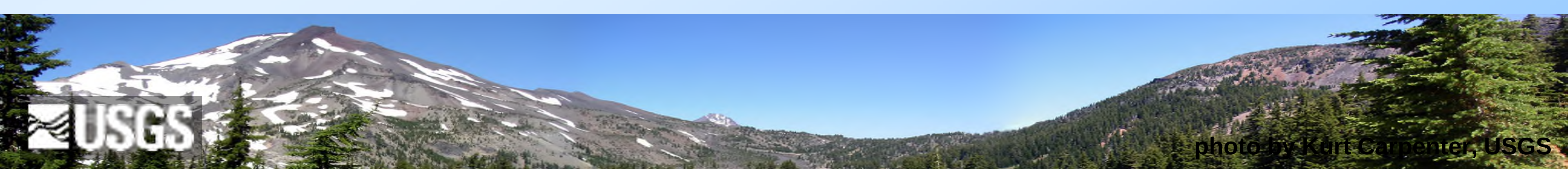


photo by Kurt Carpenter, USGS

FDOM Data – Sources of Error

Errors in FDOM data

- Sensor electronic noise	1%
- Instream variability during measurement period	1%
- Representativeness of sensor location in channel	1%
- Errors in the applied fouling and drift corrections	5%
- Errors in other data corrections	4%

$$E_P = \sqrt{(1\%)^2 + (1\%)^2 + (1\%)^2 + (5\%)^2 + (4\%)^2} = 6.6\%$$

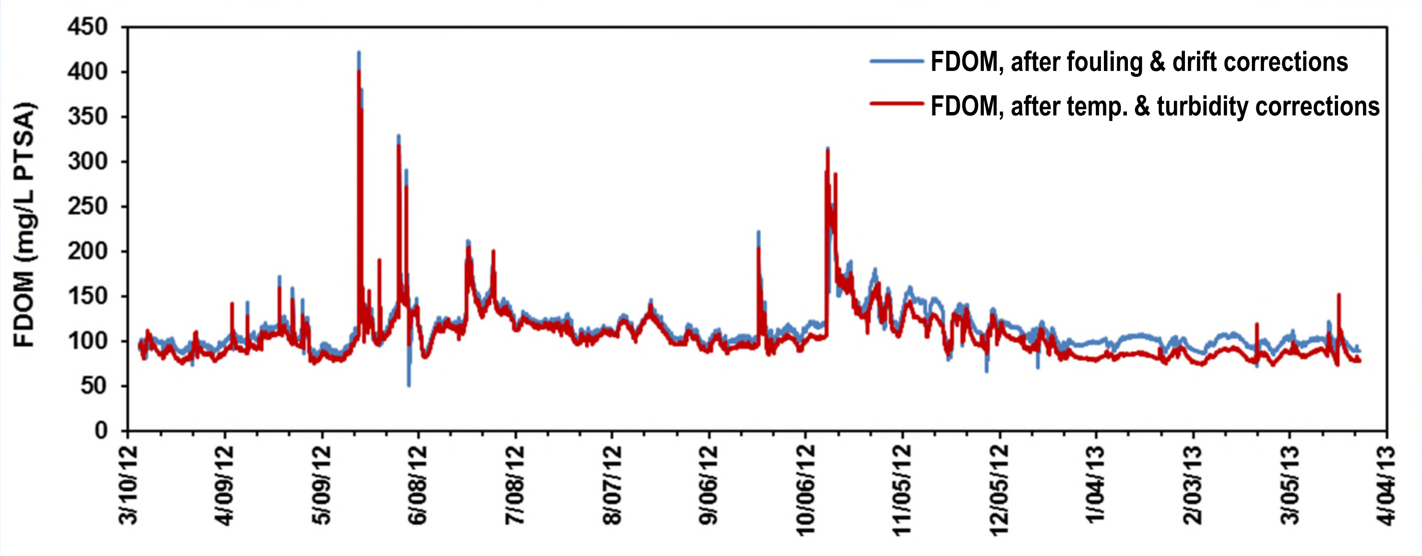


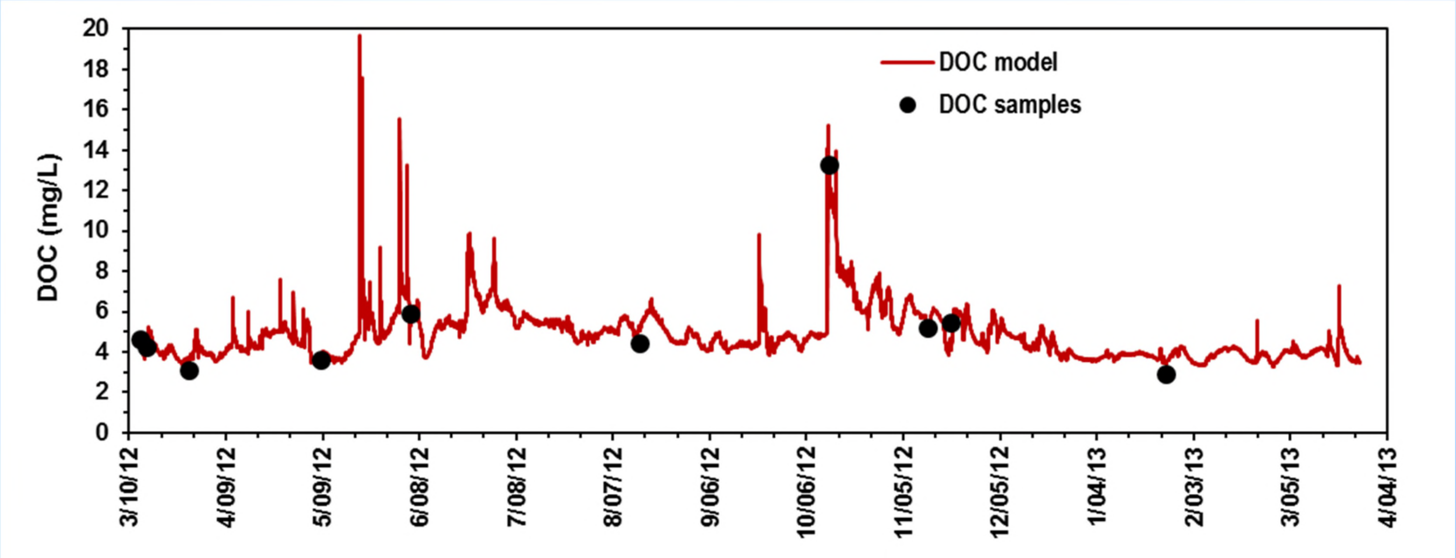
photo by Kurt Carpenter, USGS

FDOM to DOC – Sources of Error

Conversion to DOC

- Uncertainty in FDOM data 6.6%
- Analytical error in lab validation DOC samples 3.0%
- Model error 13.6%
- Total combined uncertainty for estimated DOC data → 15.4%

$$E_P = \sqrt{(6.6\%)^2 + (3\%)^2 + (13.6\%)^2} = 15.4\%$$



FDOM to DOC – Sources of Error

Calculation of Loads

- Uncertainty in DOC estimates 15.4%
- Error in streamflow data 5.0%
- Total combined uncertainty for hourly DOC loads → 16.2%

$$E_P = \sqrt{(5\%)^2 + (15.4\%)^2} = 16.2\%$$

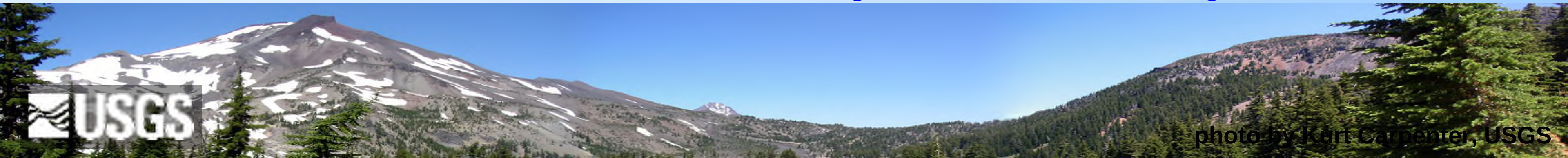
Calculation of Daily and Annual Loads

- Hourly loads are highly autocorrelated
  - → For daily loads: retained uncertainty of hourly loads (16.2%)
- Daily loads at this site are far less autocorrelated (flashy stream)
  - → For annual load: assumed daily loads were independent

$$E_{annual} = \sqrt{\sum E_{daily}^2} = 0.162\sqrt{\sum L_{daily}^2}$$

E_{annual} is the combined probable uncertainty of the annual load (1.8%)

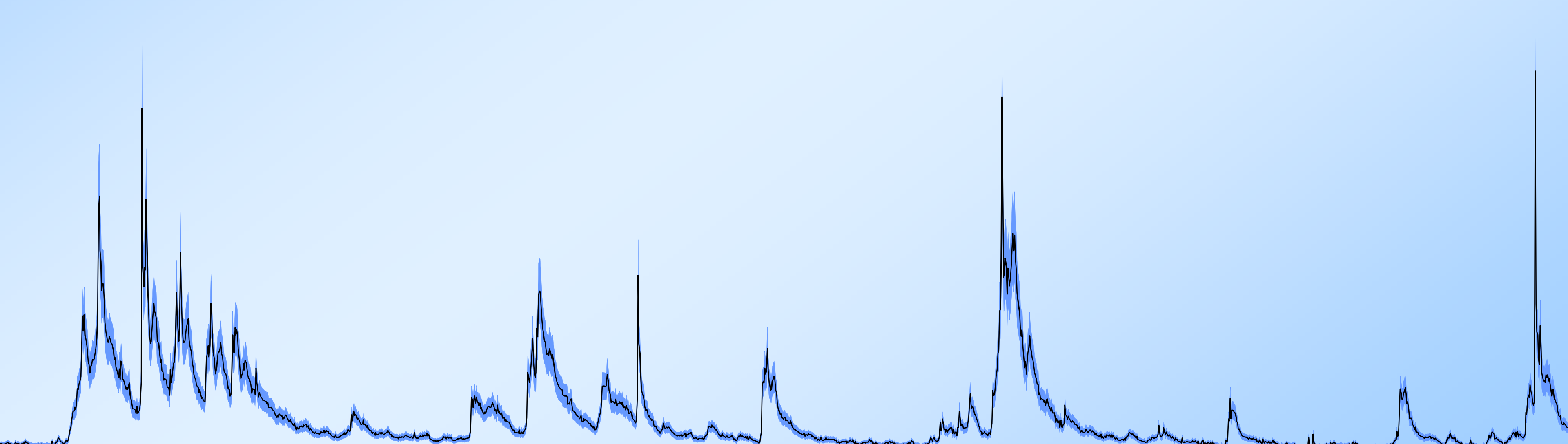
Annual load = 241,000 kg of DOC, ± 4,300 kg

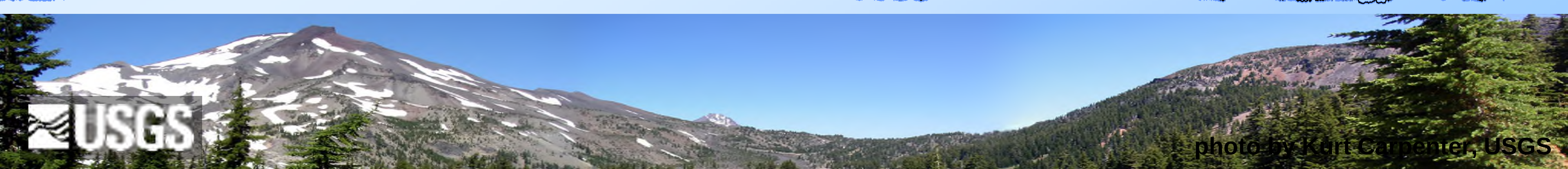


Now What?

Next Steps

- Refine the draft methods & guidelines, in consultation with various interested groups
- Write report, make recommendations
- Peer review & more discussions with interested groups
- Finalize report
- Implementation stage…





Contact, and Thanks

Stewart Rounds

USGS Oregon Water Science Center
sarounds@usgs.gov
503-251-3280

Data Uncertainty Team

USGS:	Stacey Archfield, Janice Fulford, David Holtschlag, Brian Pellerin, Pat Rasmussen, Susan Wherry
YSI / Xylem:	Rob Ellison
Aquatic Informatics:	Stu Hamilton, Brian Gouge