SAND2007-3031C

Event Detection from Water Quality Time Series

Sean A. McKenna, David Hart, Katherine Klise,
Victoria Cruz and Mark Wilson
Sandia National Laboratories,
Albuquerque, NM 87185-0735

ASCE-EWRI May 2007

Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Sandia National Laboratories

Goal and Approach

- On-line separation of anomalous water quality conditions, or outliers, from the normal background operating conditions
  - Work with multiple noisy and non-stationary water quality signals
  - Focus on classical statistical and time series estimation techniques
  - Rapid start up (on-line vs. data mining)

Basic Approach

- Two Components
  - State Estimation
  - Outlier Detection

- Different approach from data-mining based tools
  - Proposed approach has nearly immediate startup with no need for long-term data at a location
  - Data Mining approach may require considerable data base for training

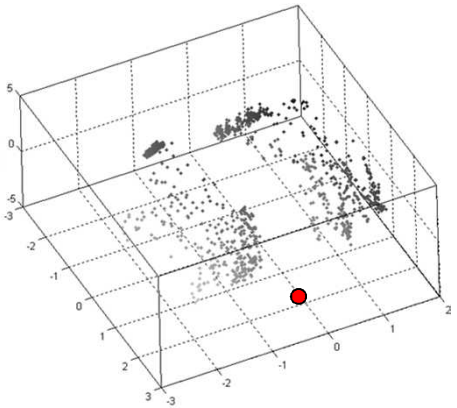
State Estimation

- Two approaches:
  - 1) Autoregressive model (linear filter)
    - Estimate: weighted linear comb. of previous obs.

$$\hat{z}(t) = \sum_{i=1}^{P} a_i z(t-i)$$

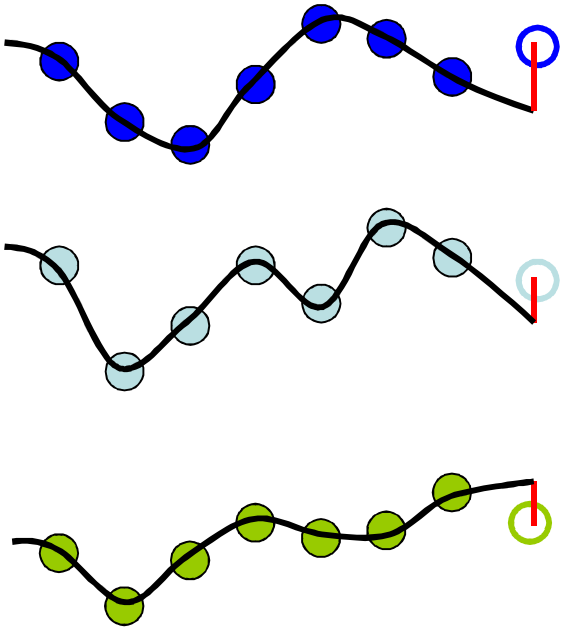
    - Residual: $d_{LPC}(t) = \hat{z}(t) - z_{obs}(t)$
  - 2) Multivariate Nearest Neighbor
    - Estimate: Current measurement in multivariate space
    - Residual: Euclidean distance to NN

$$d_{NN} = MIN\left[\sqrt[N_{dm}]{\sum_{i=1}^{N_{dm}} (x_i - x_0)^2} \right]$$



Sandia National Laboratories

State Estimation and Residuals



$$d^1_{LPC}(t) = \hat{z}^1(t) - z^1_{obs}(t)$$

$$d^2_{LPC}(t) = \hat{z}^2(t) - z^2_{obs}(t)$$

$$d^3_{LPC}(t) = \hat{z}^3(t) - z^3_{obs}(t)$$

Maximum of individual distances across all water quality signals at time step t (*residual, R(t)*) is compared against a threshold to identify an outlier

$$R(t) = MAX\left[ABS\left(d^j_{LPC}(t)\right)\right]$$

Prior normalization of the signals puts R(t) into units of standard deviation

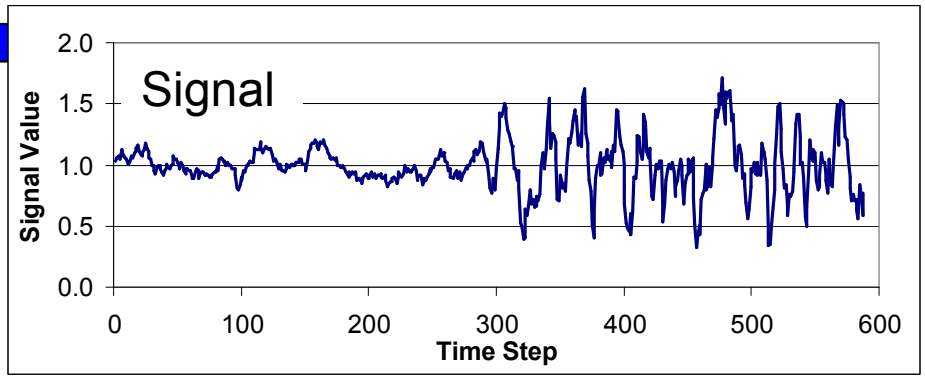
Sandia National Laboratories

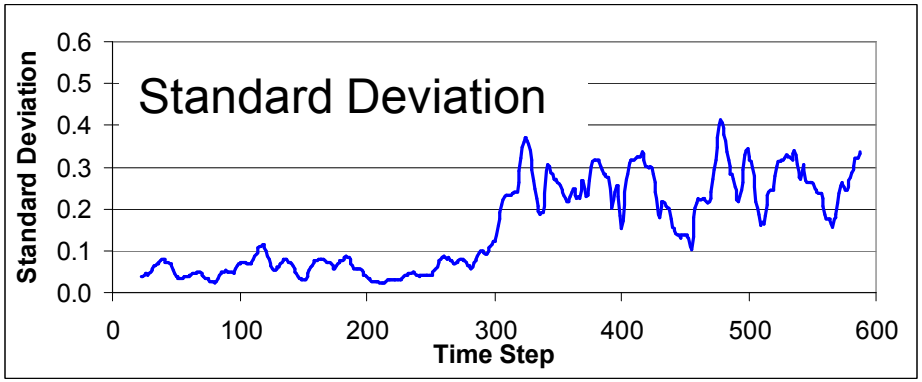
Residual Analysis

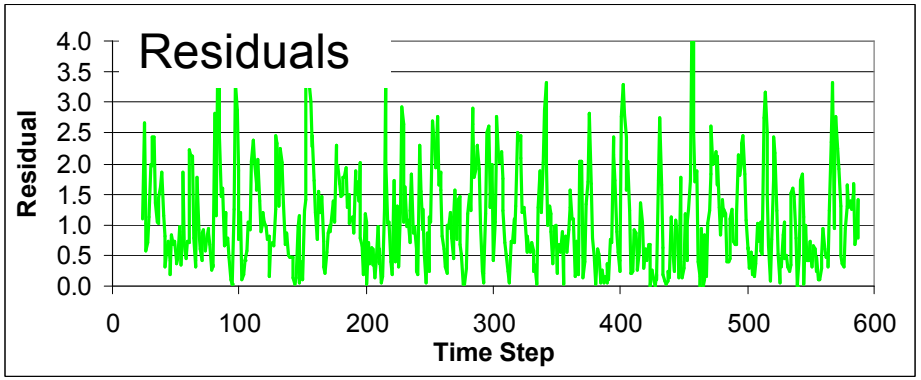
Residuals are defined in units of standard deviation away from the mean water quality.

A large change in the variance of the signal does not change the values of the residual in standard deviation space

Residuals are measured relative to the recent variability in the signal



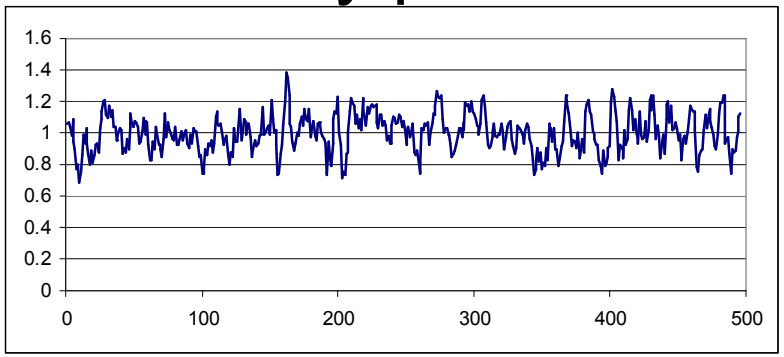




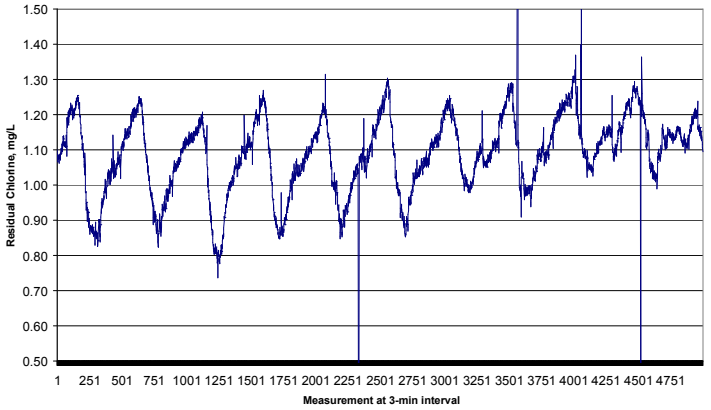
Sandia National Laboratories

Complications

- Water distribution systems are noisy places
  - $z(t) = m + e(t)$



- Non-stationary signals (often removed with detrending, but may not have that info in distribution systems)
  - $z(t) = m(t) + e(t)$
  - $z(t) = f(y(t)) + e(t)$

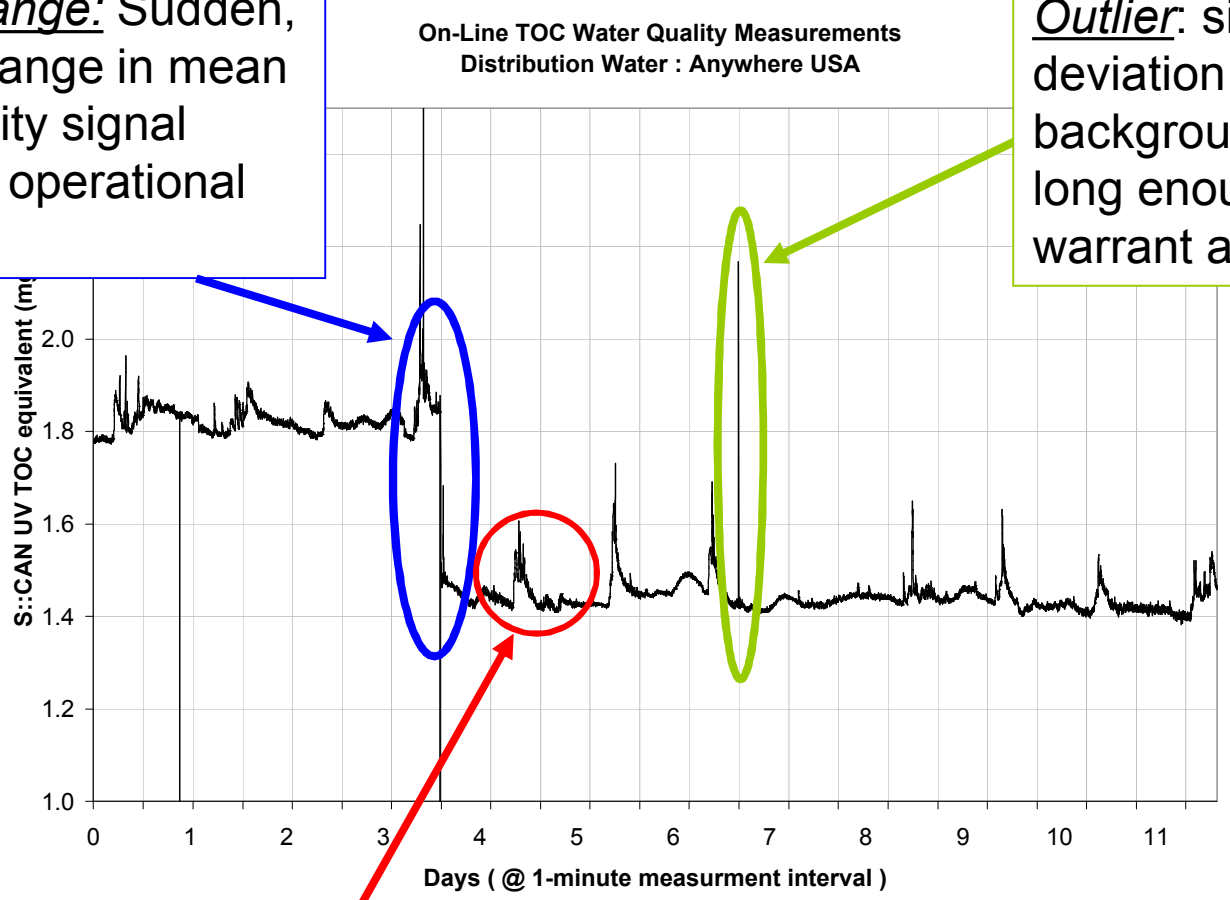


Sandia National Laboratories

Outliers, Events and Baseline Changes

Baseline Change: Sudden, persistent change in mean of water quality signal (often due to operational changes)

Outlier: significant deviation from background that is not long enough to warrant an alarm



Event: Multiple outliers within a specified period of time

Sandia National Laboratories

Binomial Event Discriminator

- How many outliers constitute a water quality event?
- Binomial distribution provides the probability of *r* outliers in *n* time steps for a given false positive rate

$$b(r;n,p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{(n-r)} = \frac{n!}{r!(n-r)!} p^r q^{(n-r)}$$

$$P(r \le z_c) = \sum_{i=1}^{N(r \le z_c)} b(r;n,p)$$

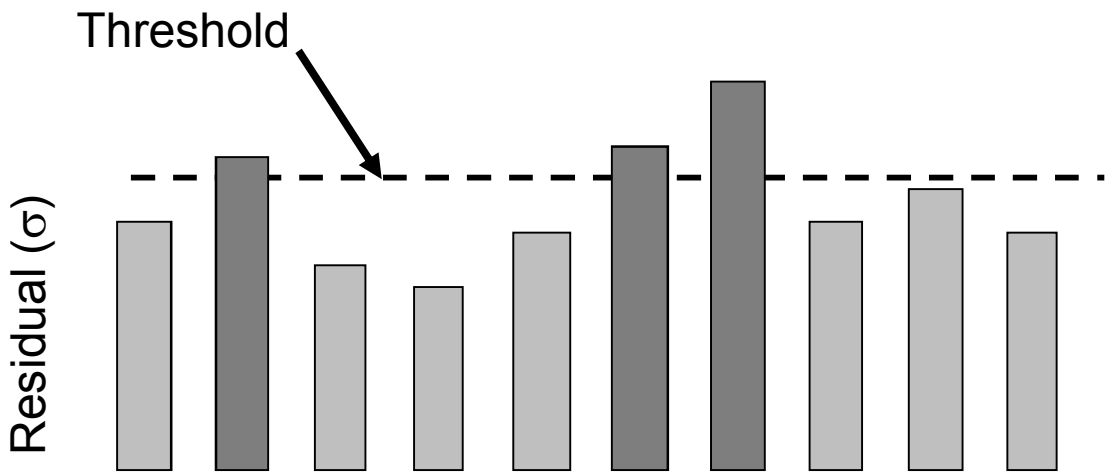
b(r;n,p) provides probability of observing *r* outliers in *n* time steps under background conditions

Define *1.0 - b(r;n,p)* as the probability of an event *P(event)*

Sandia National Laboratories

BED Example

- Simple Example:
  - *n* = 10 trials and *prob_thresh* = 0.950
  - *r* = 3 failures (outliers) causes event determination
  - *p* = individual probability of failure = 0.05



	P(event)
0	4.012631E-01
1	6.848753E-01
2	9.253652E-01
3	9.895249E-01
4	9.990352E-01
5	9.999391E-01
6	9.999973E-01
7	9.999999E-01
8	1.000000E+00
9	1.000000E+00
10	1.000000E+00

Note that outliers need not be consecutive

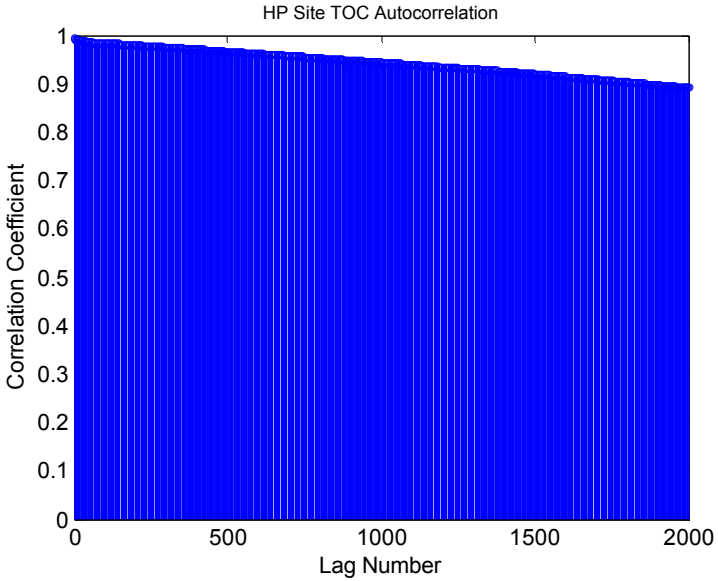
Binomial Assumptions

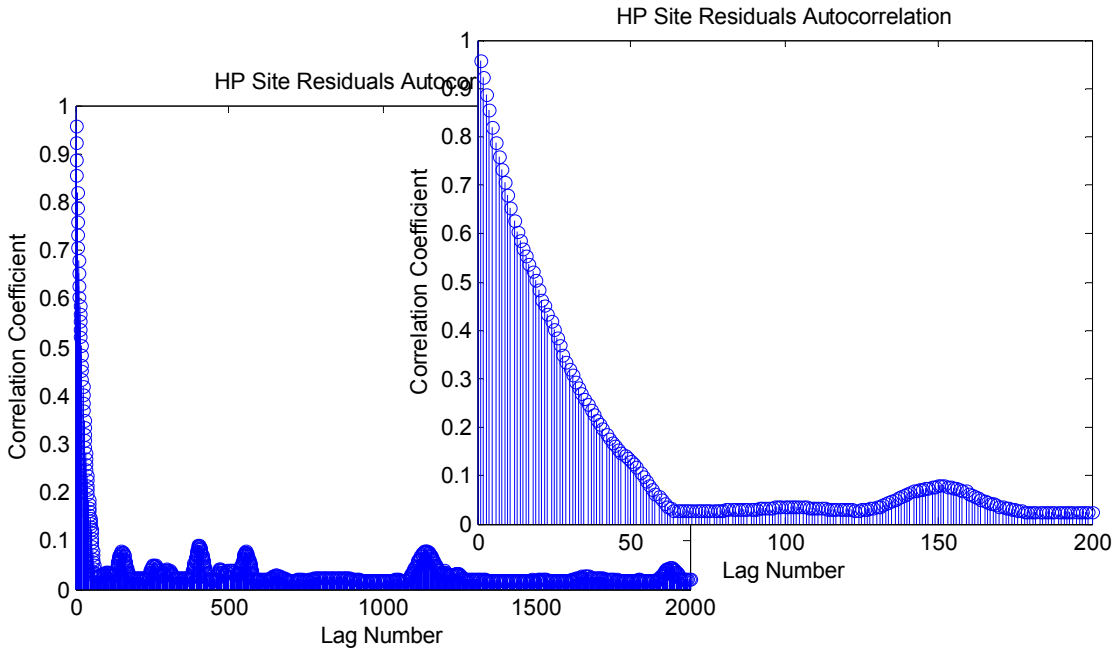
Use of Binomial model for event detection has inherent assumptions

- There are *n* repeated trials in the experiment
- Each trial can only have one of two outcomes: success or failure
- The probability of success, *p*, remains constant from one trial to the next.
- Repeated trials are independent of one another.

Independence of Trials

- Water quality signals display auto-correlation
- Residuals of predictions do not (HP Site Example)



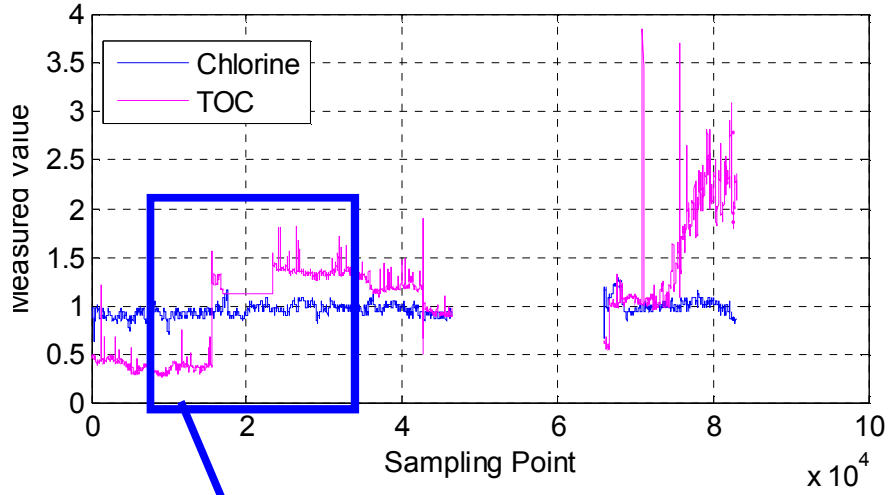


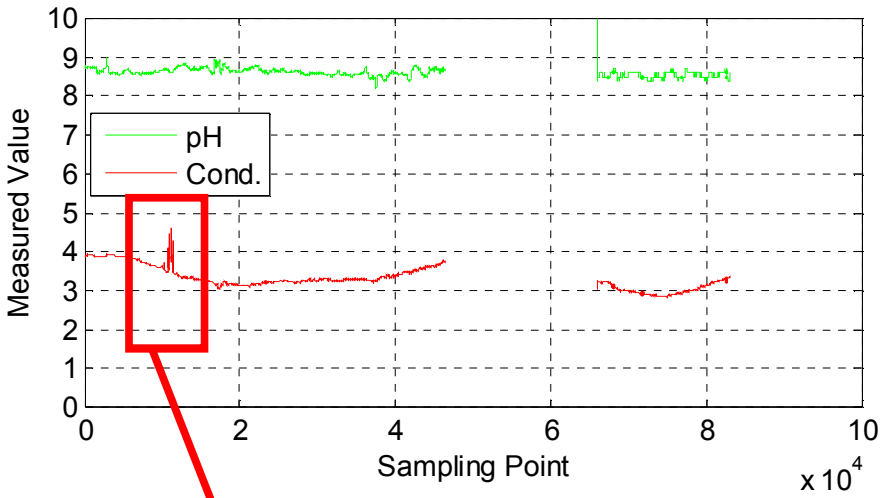
Ideally, distribution of residuals should be uncorrelated with mean zero (unbiased), and have minimum variance

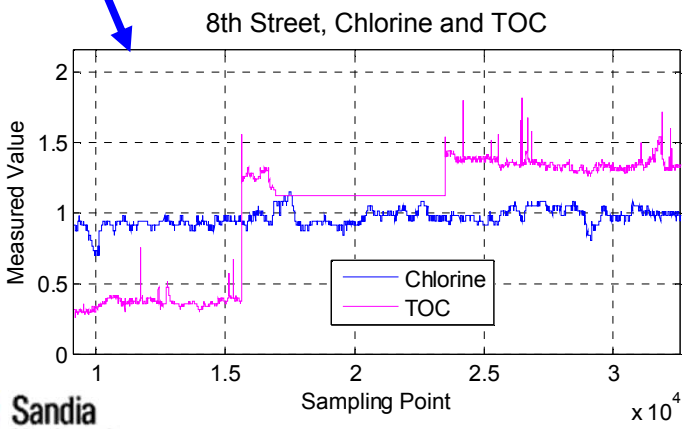
Sandia National Laboratories

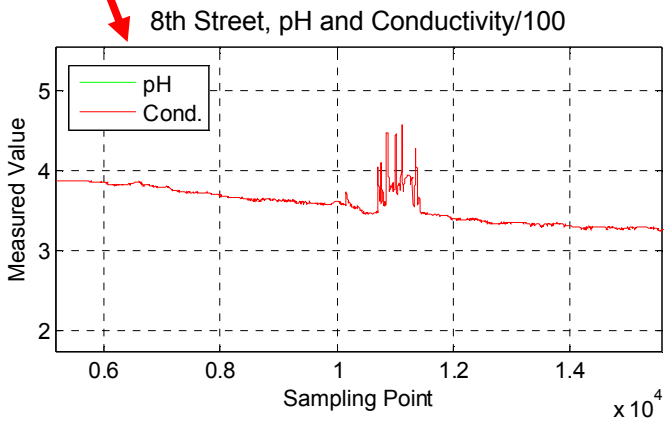
Example Application: 8S site

Four signals, Conductivity is scaled by 100, large gap, conductivity event, 2 minute sampling interval, approximately 16 weeks of data









Sandia National Laboratories

8S Site Results

LPC algorithm used to estimate water quality values. Window of 360 (12 hrs) and threshold of 1σ

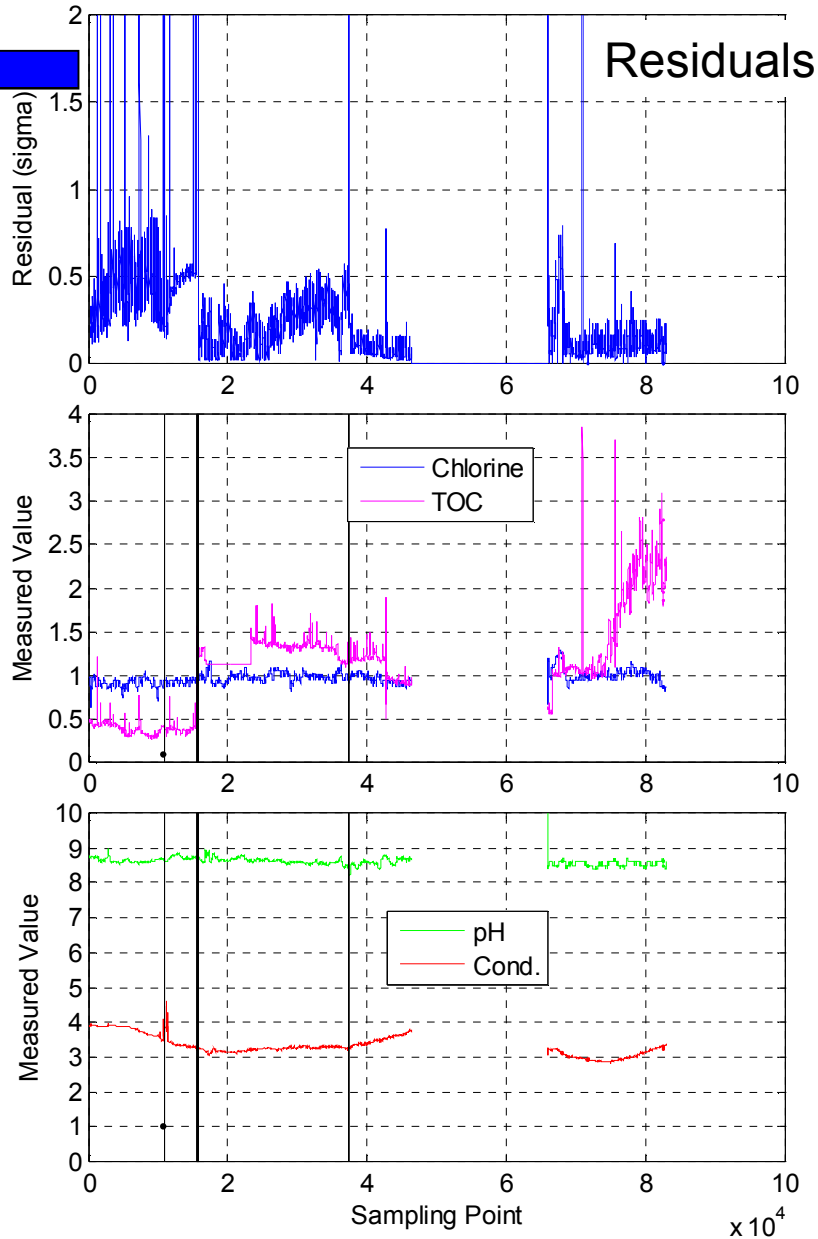
30 trials for event, 50 for baseline change. *P(outlier)* = 0.50 and *prob_thresh* = 0.995 (23/30 outliers required for an event)

Results:

3 baseline changes and 9 other events detected. Two baseline changes correspond to times highlighted on previous slide. Final baseline change is more subtle.

Strong trend in TOC over final 18,000 time steps is recognized as background

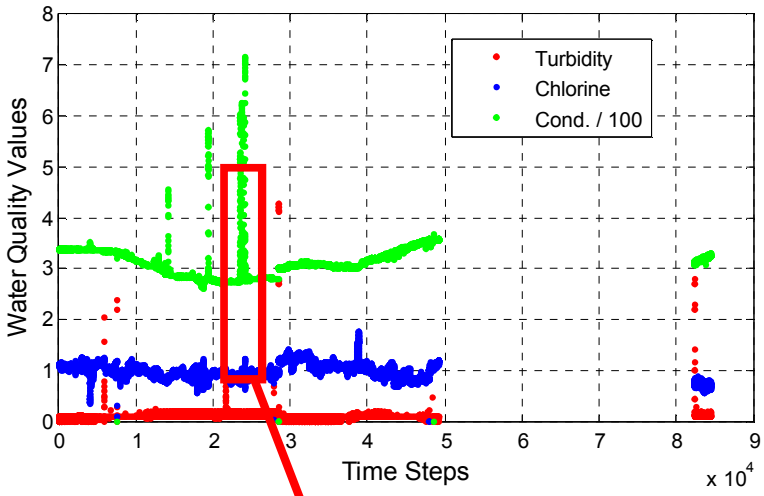
Average abs (*residual*) = 0.185σ

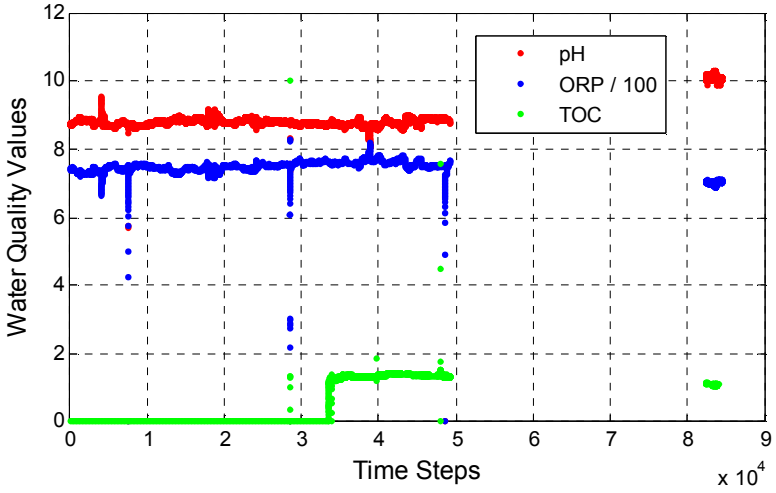


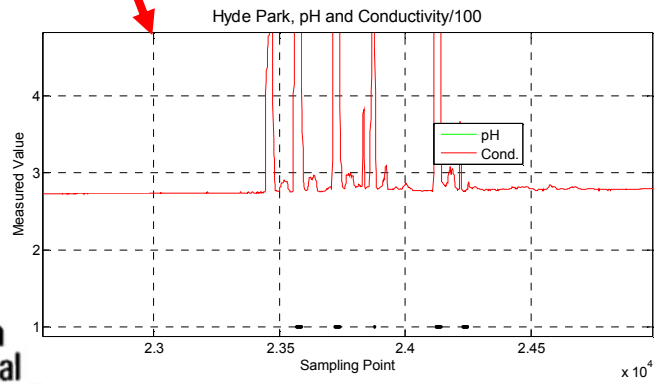
Sandia National Laboratories

Example Application: HP site

Six signals, Ignore turbidity and TOC, large gap – only work with first 50,000 time steps (approx. 10 weeks of 2 minute data)







Sandia National Laboratories

HP Site Results

LPC algorithm used to estimate water quality values. Window of 240 (8 hrs) and threshold of 0.9σ

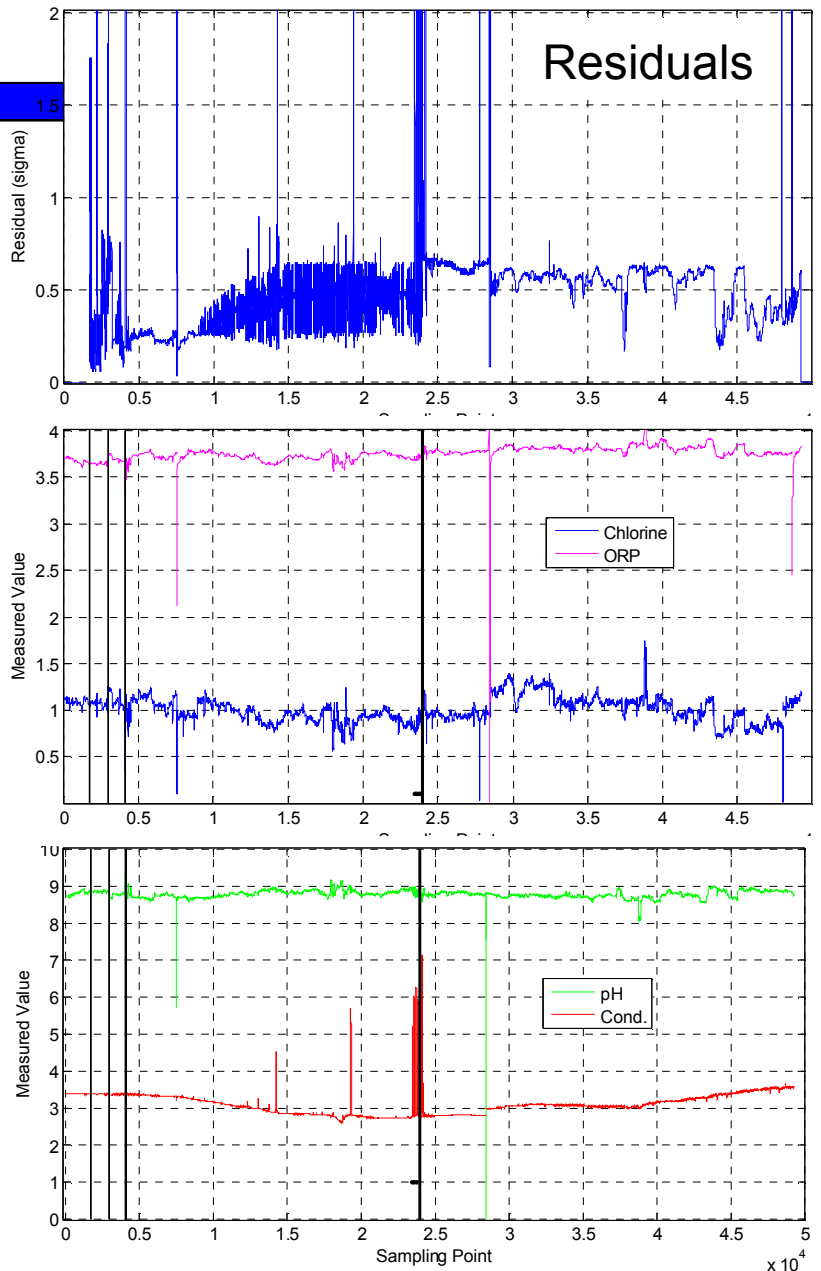
50 trials for event, 70 for baseline change. *P(outlier)* = 0.50 and *prob_thresh* = 0.995 (34/50 outliers required for an event)

Results:

6 baseline changes and 48 other events detected. Majority of baseline changes at early times. Final baseline change is due to Conductivity event

Large number of short spikes are ignored

Average abs (*prediction error*) = 0.35σ



Sandia National Laboratories

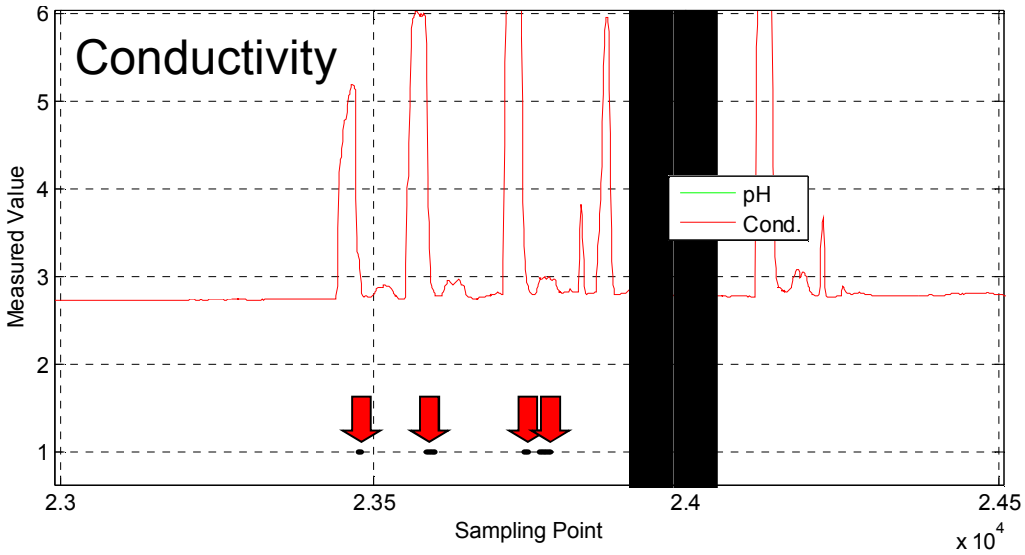
HP Site Results: Zoomed View

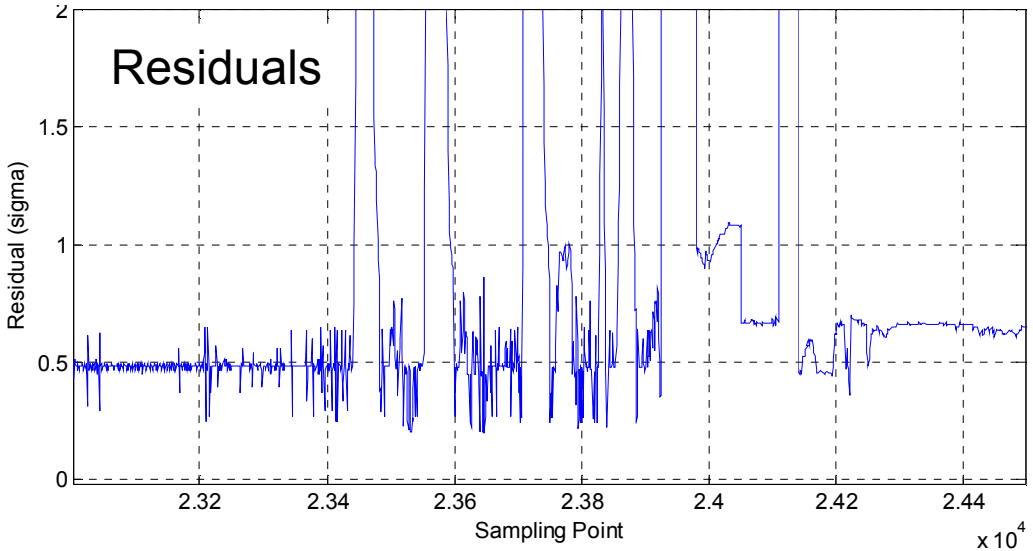
Expanded view of conductivity events from 23-25,000 time steps

Threshold is 0.9σ

Initial modes of conductivity are identified as events (black dots under red arrows)

Finally there is a sustained event that meets the baseline change criterion (70 time steps) shown as black band in upper figure)





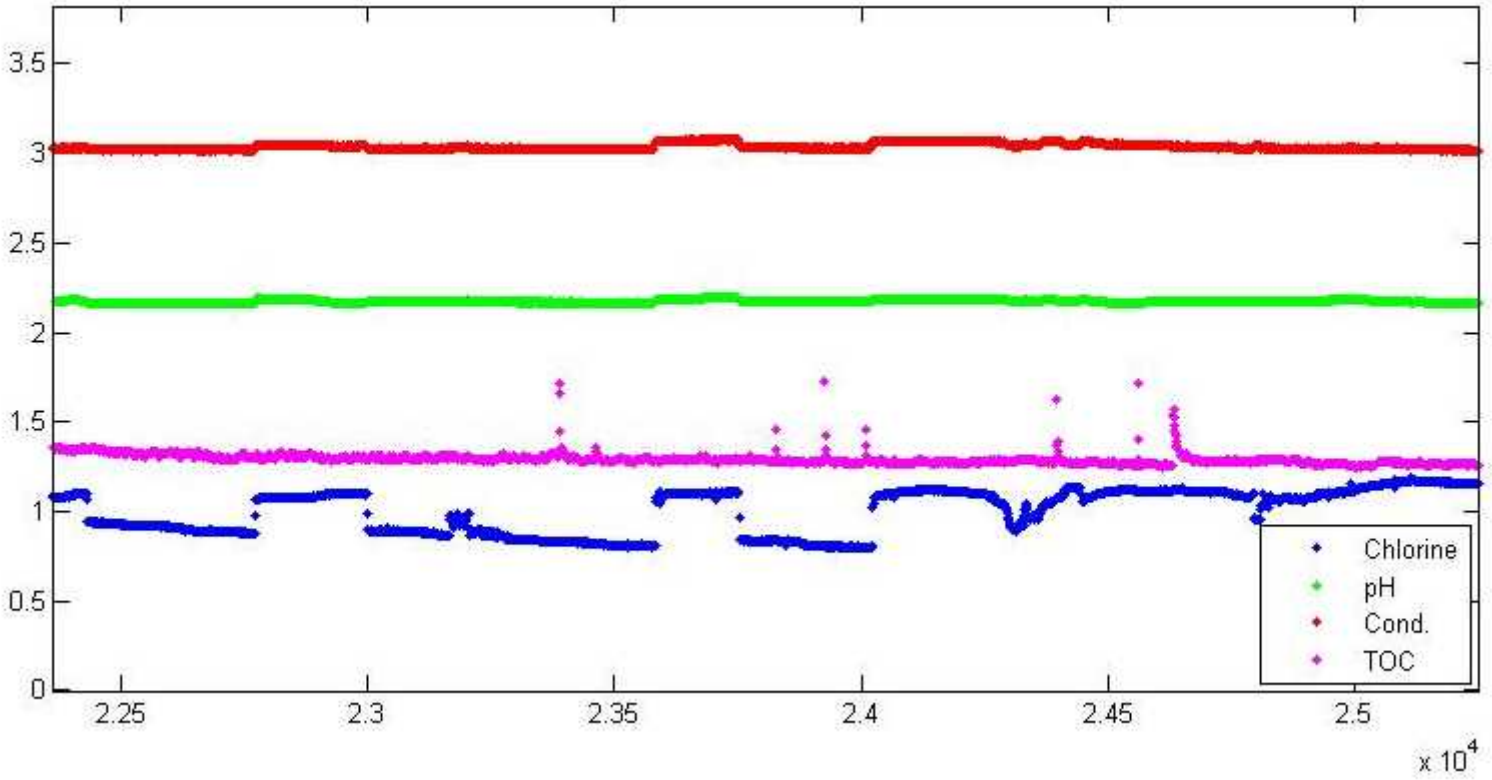
Sandia National Laboratories

Summary

- Developed two-phase approach for event detection
  - State estimation and residual evaluation
- Binomial event discriminator
  - Reduce false positives caused by short-lived outliers
  - Binomial parameters customizable to find particular event sizes
  - Quantitative evaluation of baseline changes in non-stationary time series (can be tailored to location specific expected changes)
- Results at two locations
  - Identified known events and baseline changes (few of these)
  - Percent of total events identified as false positives: 0.016 and 0.096

Sharp Changes due to Hydraulics

Sharp, significant changes occurring at aperiodic times are an outstanding challenge for the on-line approaches discussed here



Sandia National Laboratories

K site example: Cl, pH/4 and Conductivity/100