SAND2006-5235C

Water Quality Change Detection

Sean A. McKenna[1], Katherine A. Klise[1], Mark Wilson[2], David Hart[1],
Victoria Cruz[1], Nnaemeka Madueke[3], Sanjay Srinivasan[3]

[1]Geohydrology Department, Sandia National Laboratories, Albuquerque, New Mexico
[2]Gram Inc., Albuquerque, New Mexico
[3]Department of Petroleum and Geosystems Engineering, University of Texas, Austin, Texas

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-AC04-94-AL-85000

Outline

- Background on 3 algorithms
- ROC Curves
- Simple comparison to set points
- Results of testing on EPA T&E facility data
- Results of testing on simulated events
- Results of T&E events superimposed on observed water quality

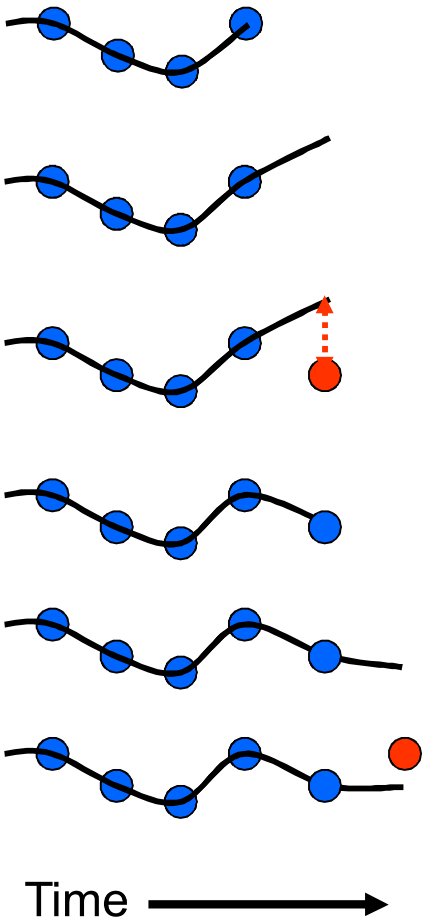
Motivation

- Is there something going on in the distribution system that is causing a significant change from normal operating conditions?
- Robust, in-situ, real-time, contaminant-specific sensors don't exist
- Develop "change detection" algorithms to identify anomalous water quality conditions

Key Points

- Water quality data are correlated in time
  - Future measurements are correlated with previous measurements
- Ability to detect a change is only half the answer
  - Ability to <u>not detect</u> false changes is key
- Testing of algorithms against known water quality events in systems is problematic

Water Quality Prediction



Model Fit to observations

Model predicts next value(s)

Next value is obtained and difference between model and observation is calculated

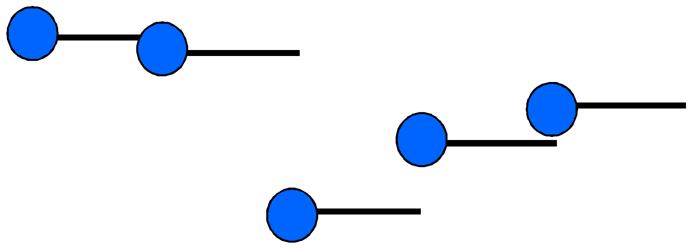
Most recent value is incorporated into model

Model predicts next value

Next value is obtained and difference between model and observation is calculated

Time Series Increments

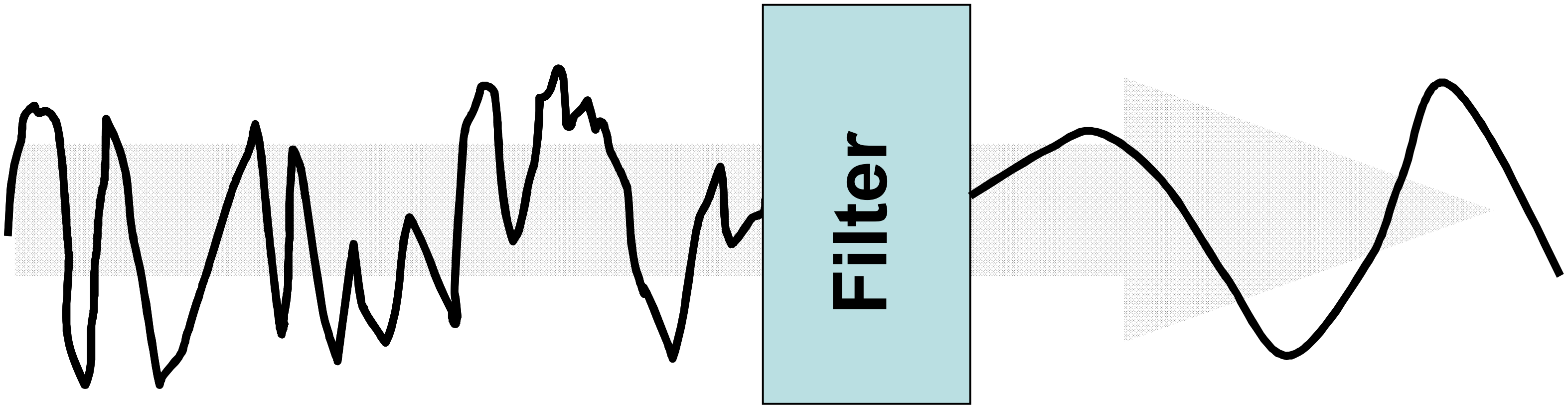
- The measured value at the previous time step serves as the prediction for the next time step



Simplest of the 3 algorithms. Only requires the previous measurement to be kept in the time history.

Linear Filter

- "Filter": means of removing noise to accentuate a signal
- Filter a time window of previously observed data to predict the signal at the next time step(s)

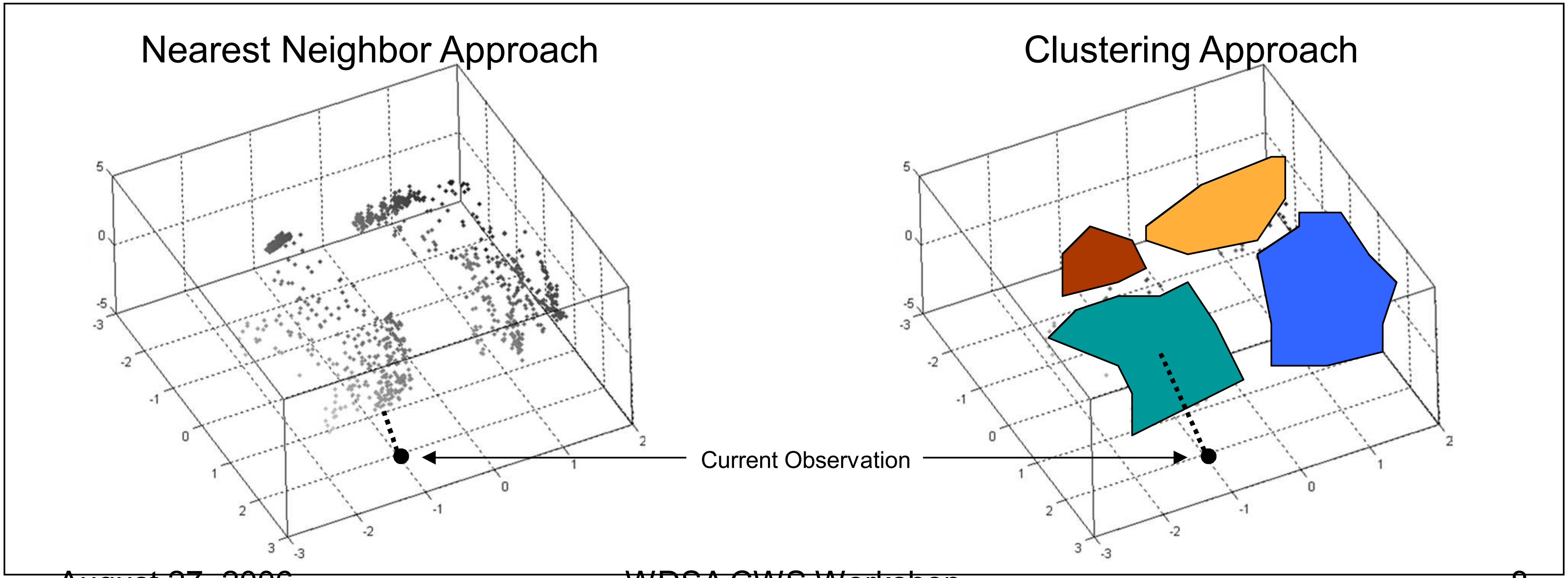


$$C(t) = \alpha_1 C(t-1) + \alpha_2 C(t-2) + \alpha_3 C(t-3) + \varepsilon(t)$$

Multivariate Nearest Neighbor

Have we seen this type of water recently?

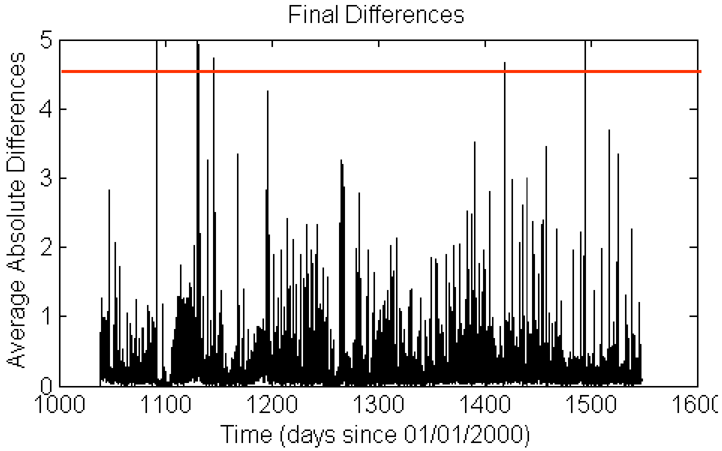
Measure distance from current measurement to nearest neighbor of recent measurements in multi-dimensional water quality space. Have also evaluated grouping previous measurements into clusters

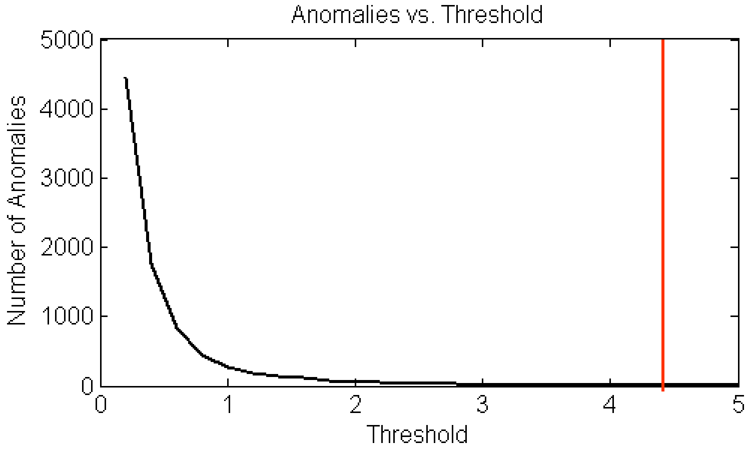


Where to Draw the Line?

Change detection approaches require some threshold above which the difference is significant

Absolute value of deviations from predicted water quality (σ)





What threshold best separates true detections from false positives? Need calibration events!

Possible Decision Results

	Actual Background Condition	Actual Anomalous Condition	
Estimated Anomalous Condition	*FP*	*TP*	*FP + TP*
Estimated Background Condition	*TN*	*FN*	*TN + FN*
	FP + TN	*TP + FN*	
	Specificity = TN / (TN + FP) FAR = FP / (TN + FP)	Sensitivity = TP / (TP + FN)	
	Specificity = 1 – FAR	Sensitivity = PD	

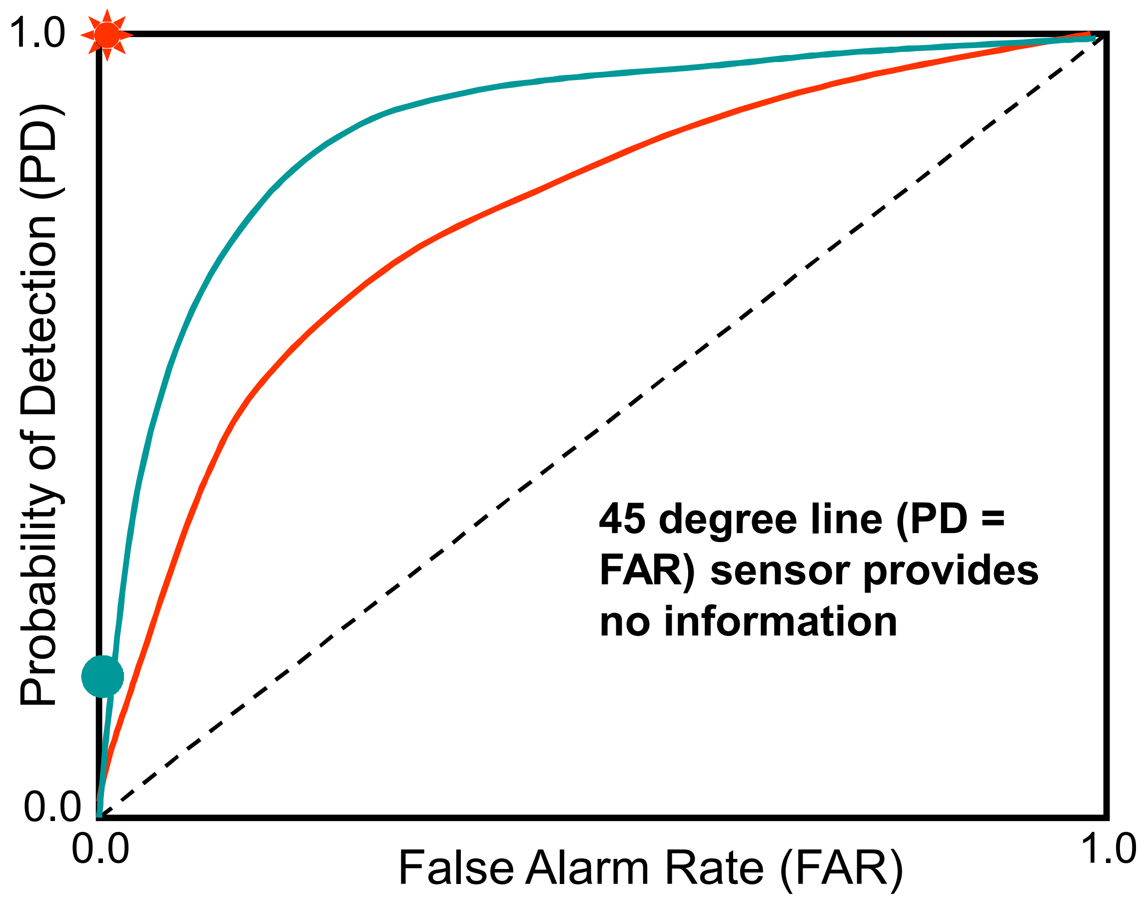
FP = False Positive

TP = True Positive

FN = False Negative

TN = True Negative

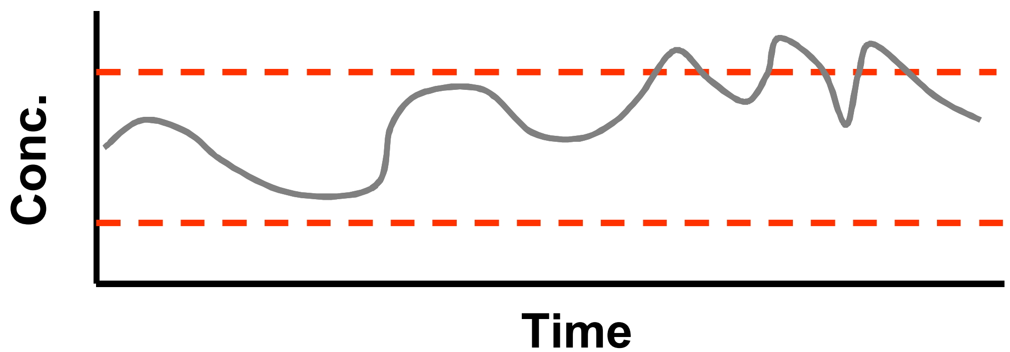
Receiver Operating Characteristic (ROC) Curve



Two ROC curves showing different sensor operating characteristics. Points along each curve are calculated by varying the threshold value above which events are identified and determining the number of correct and false positive characterizations

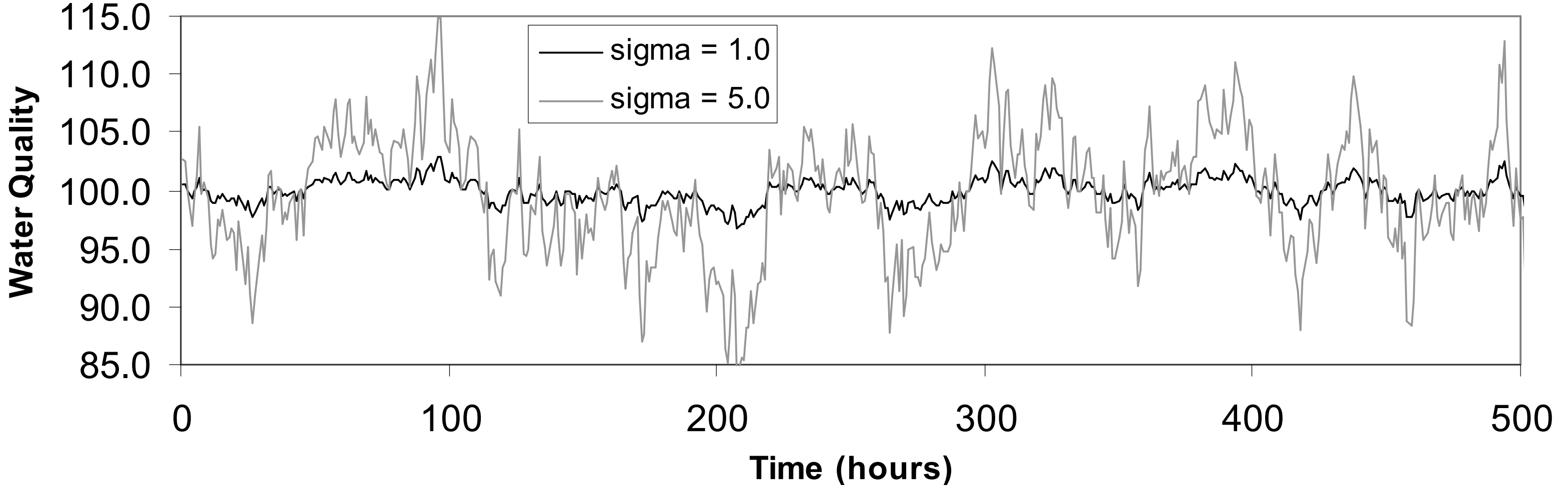
Simple Example

- Predictive approaches use recent history of the of the time series to identify changes
  - Change is relative to predicted value
- Set points use absolute limits on what is acceptable water quality to identify excursions



Simple Example (Cont.)

- Simulated water quality value (mean = 100.0, σ = 1.0, 2.0, . . ., 5.0



Case	Lower Set Point	Upper Set Point	Deviation Threshold
1	99.0	101.0	1.0
2	98.0	102.0	2.0
3	97.0	103.0	3.0

Simple Example: Results

- Total number of false positives for each case and each level of variability are tabulated
- Change in variance has no impact on change detection algorithms
- Large number of false positives renders set points ineffective

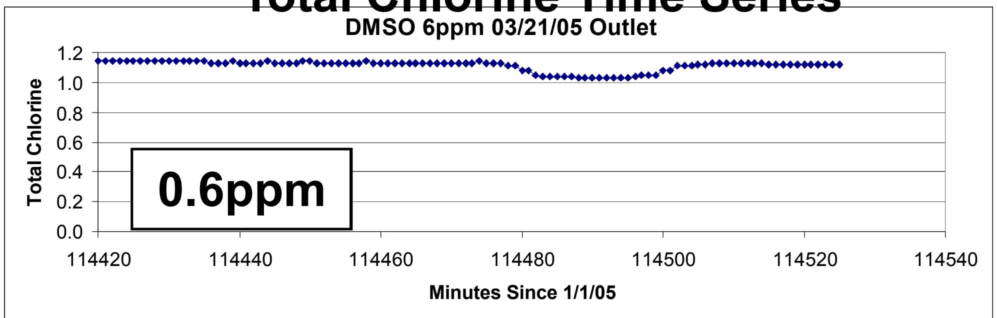
Threshold / Set Point Values	Algorithms (all σ)			Set Points				
	INC	LF	MV-NN	$\sigma = 1$	$\sigma = 2$	$\sigma = 3$	$\sigma = 4$	$\sigma = 5$
± 1.0	245	4	0	845	1576	1882	2014	2113
± 2.0	5	0	0	163	845	1312	1576	1783
± 3.0	0	0	0	14	419	845	1177	1416

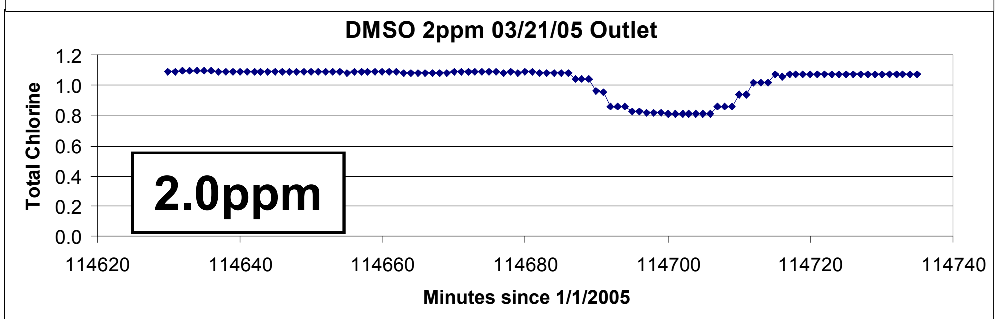
EPA Test Loop Data

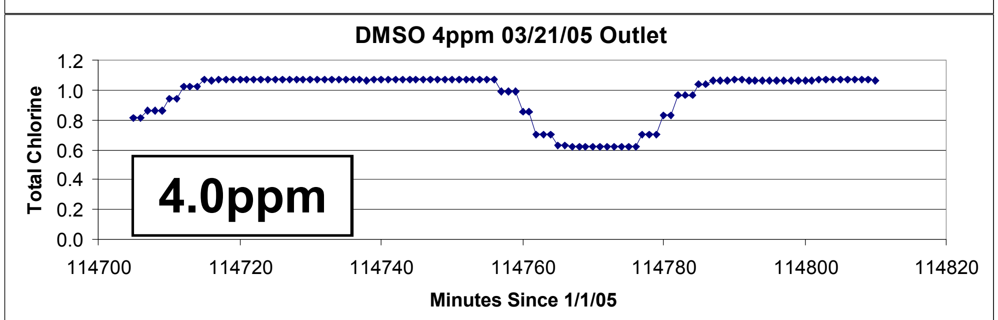
- Experiments conducted in March 2005 at EPA Test and Evaluation Facility
- Use DMSO “single pass” test with water quality sensors 1100 feet from contaminant source
- Three different source strengths of DMSO
  – 0.6 ppm, 2.0 ppm, 4.0 ppm,
  – Water quality collected on 1 minute intervals
  – Only examine changes in total Chlorine here

DMSO: 3 Source Strengths

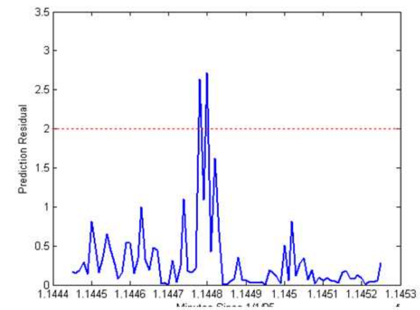
Total Chlorine Time Series

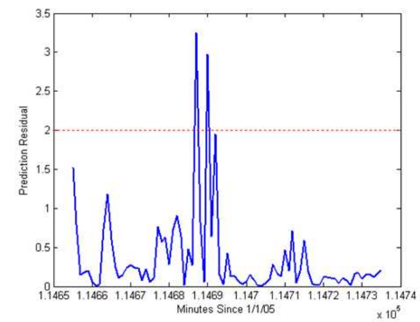


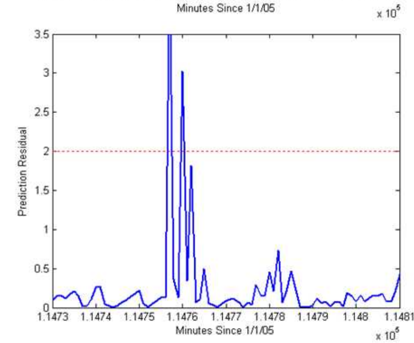




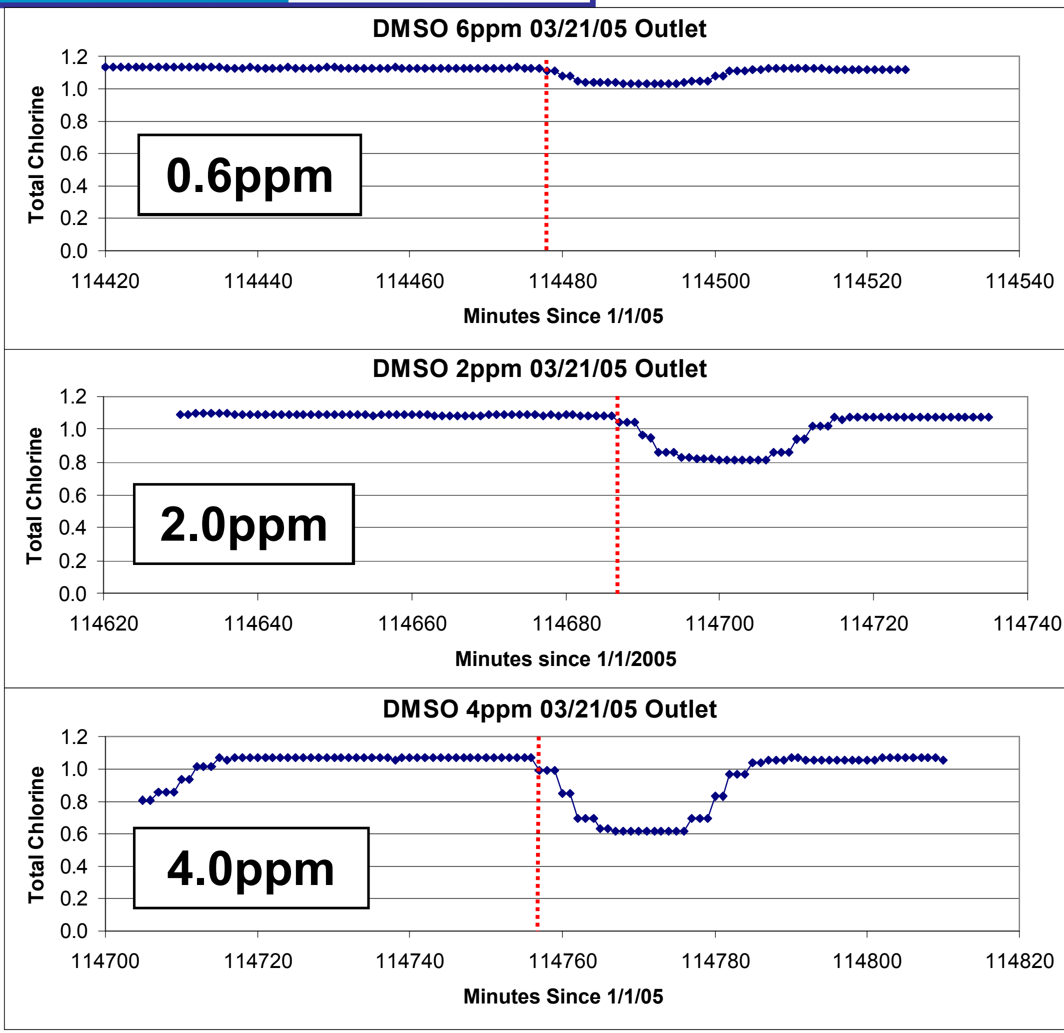
Prediction Errors







DMSO Change Detect Results



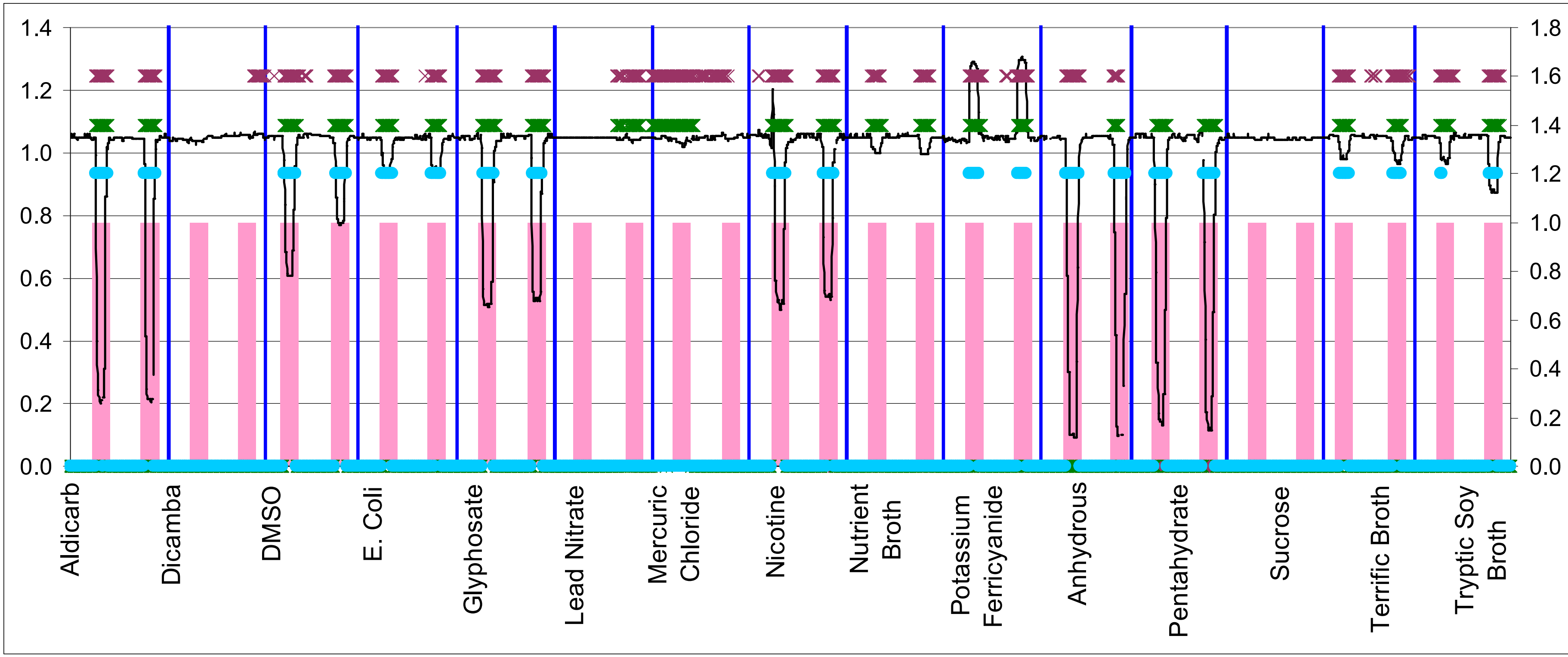
Linear filter with order of 25 minutes is used

Prediction residual threshold of 2.0 defines beginning of water quality change

Algorithm correctly identifies beginning of change for all three strengths

Algorithm Testing on EPA T&E Data

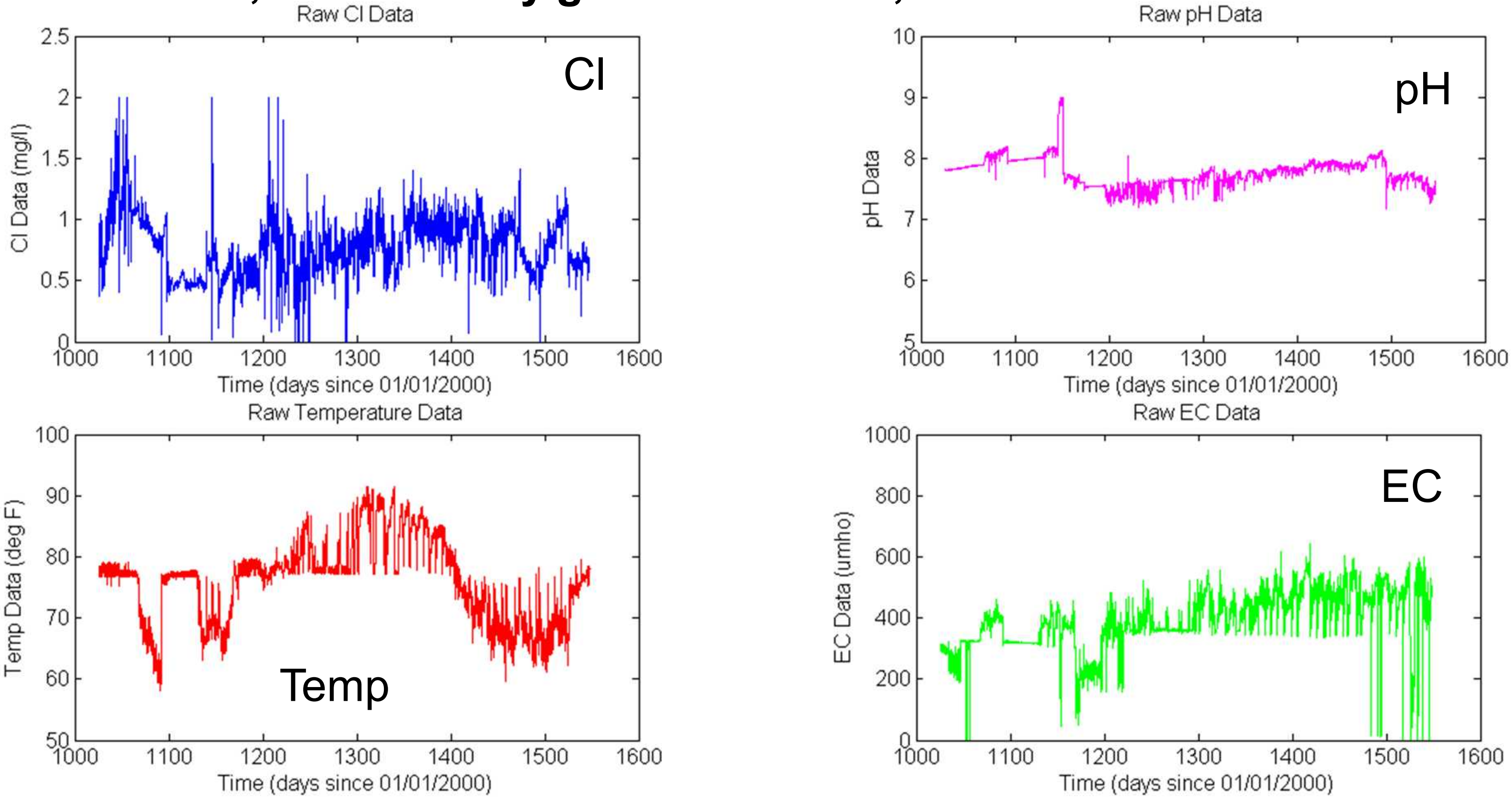
ATI Chlorine response to high source concentrations at 1100 feet



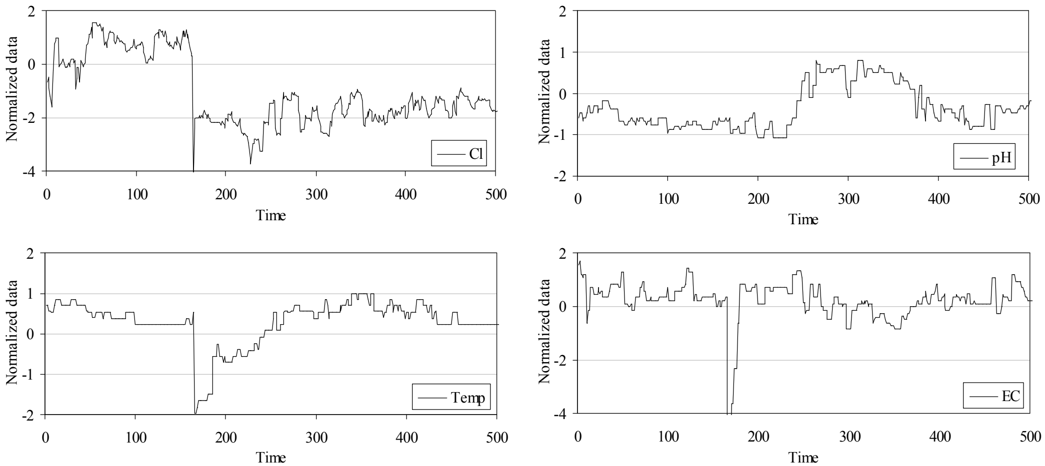
X = INC   X = LF   ● = MV

Water Quality Data

Real world water quality data are noisy, have daily and seasonal cycles, contain drift, occasionally get recalibrated, etc.

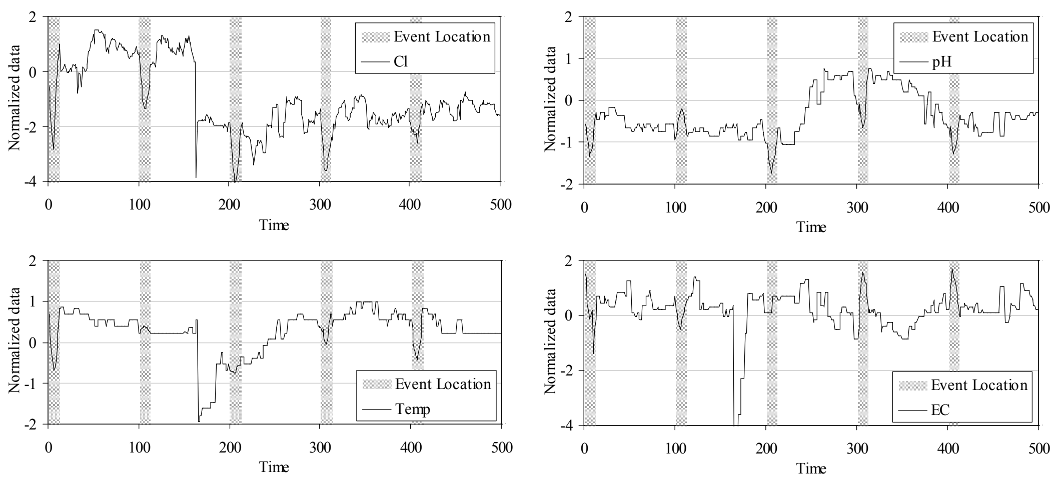


Simulated Events



Observed water quality signals: Cl, pH, Temp and EC

Water quality for a single location (P2). Three other locations also considered



Gaussian events simulated on top of observed water quality signals.

Strength = 2.5

Duration = 11 hours (events start every 100 hours)

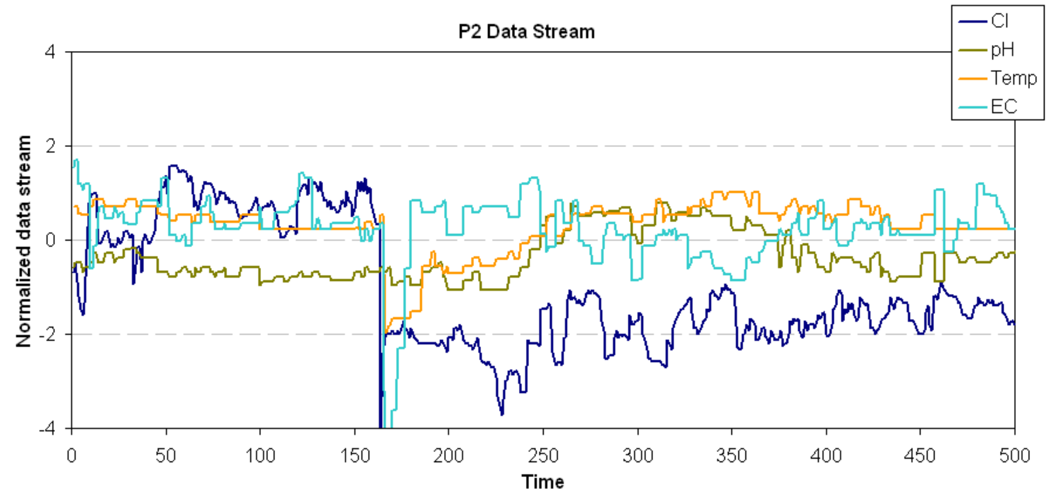
Event Strength

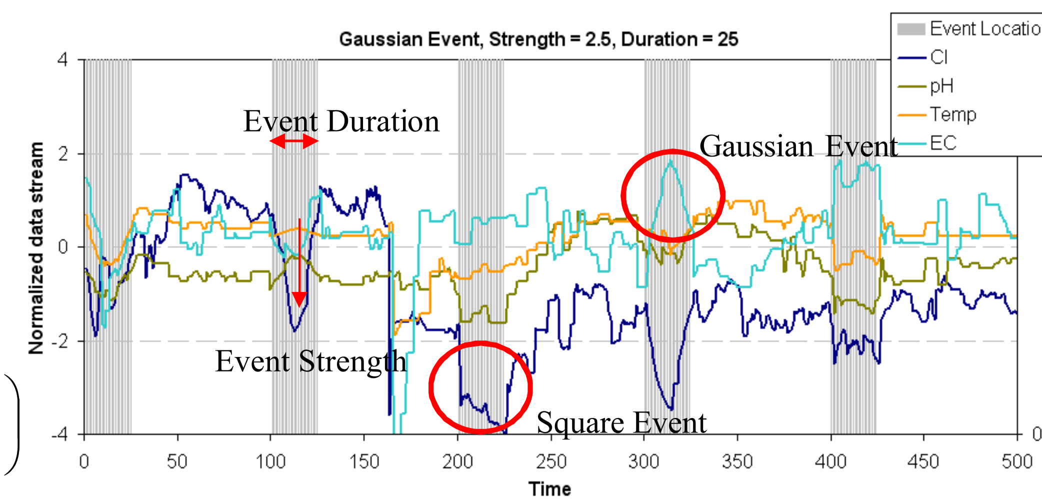
Simulated Event

Event Type
  Gaussian or Square

Event Duration
  Time span for event

Event Strength
  Deviation from measured data
  Deviations selected at random for each signal



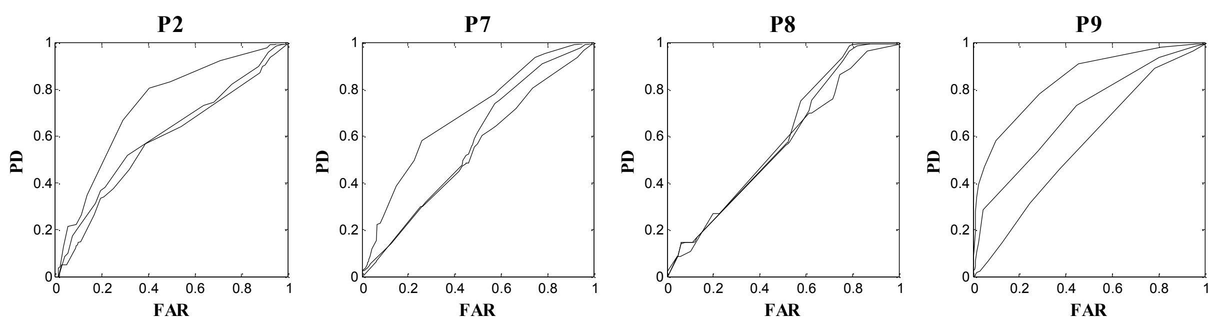


For N water quality signals:

$$\text{Strength} = \text{SQRT}\left(\sum_{i=1}^{N} (\text{deviation from measured})_i^{\,2}\right)$$

Results

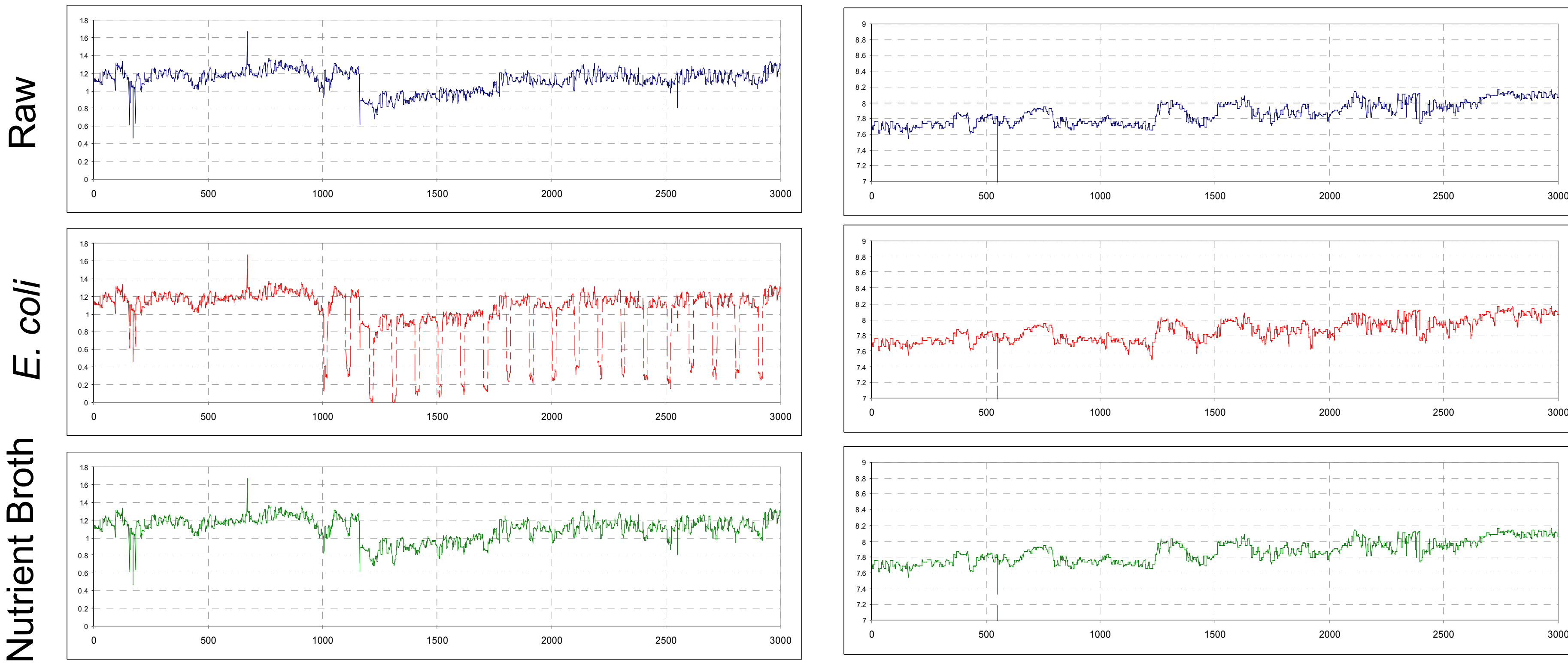
		Square Event				Gaussian Event			
Strength	**Duration**	**P2**	**P7**	**P8**	**P9**	**P2**	**P7**	**P8**	**P9**
0.5	**3**	0.586	0.529	0.586	0.702	0.557	0.503	0.545	0.588
	11	0.617	0.524	0.574	0.646	0.568	0.497	0.52	0.565
	25	0.599	0.553	0.565	0.657	0.576	0.538	0.559	0.568
1.5	**3**	0.839	0.767	0.844	0.943	0.728	0.586	0.712	0.93
	11	0.814	0.76	0.849	0.93	0.641	0.552	0.581	0.772
	25	0.797	0.731	0.892	0.914	0.606	0.575	0.576	0.698
2.5	**3**	0.94	0.784	0.915	0.995	0.891	0.801	0.841	0.956
	11	0.923	0.867	0.918	0.994	0.75	0.64	0.593	0.898
	25	0.939	0.863	0.923	0.991	0.73	0.694	0.59	0.838



ROC curves for all 3 strengths for each location. Gaussian events with duration of 25 hours

EPA T&E Events Superimposed on Measured Water Quality

Cl (left) and pH (right) after introduction of “high” dose of *E. coli* and nutrient broth superimposed on water quality measured in distribution system

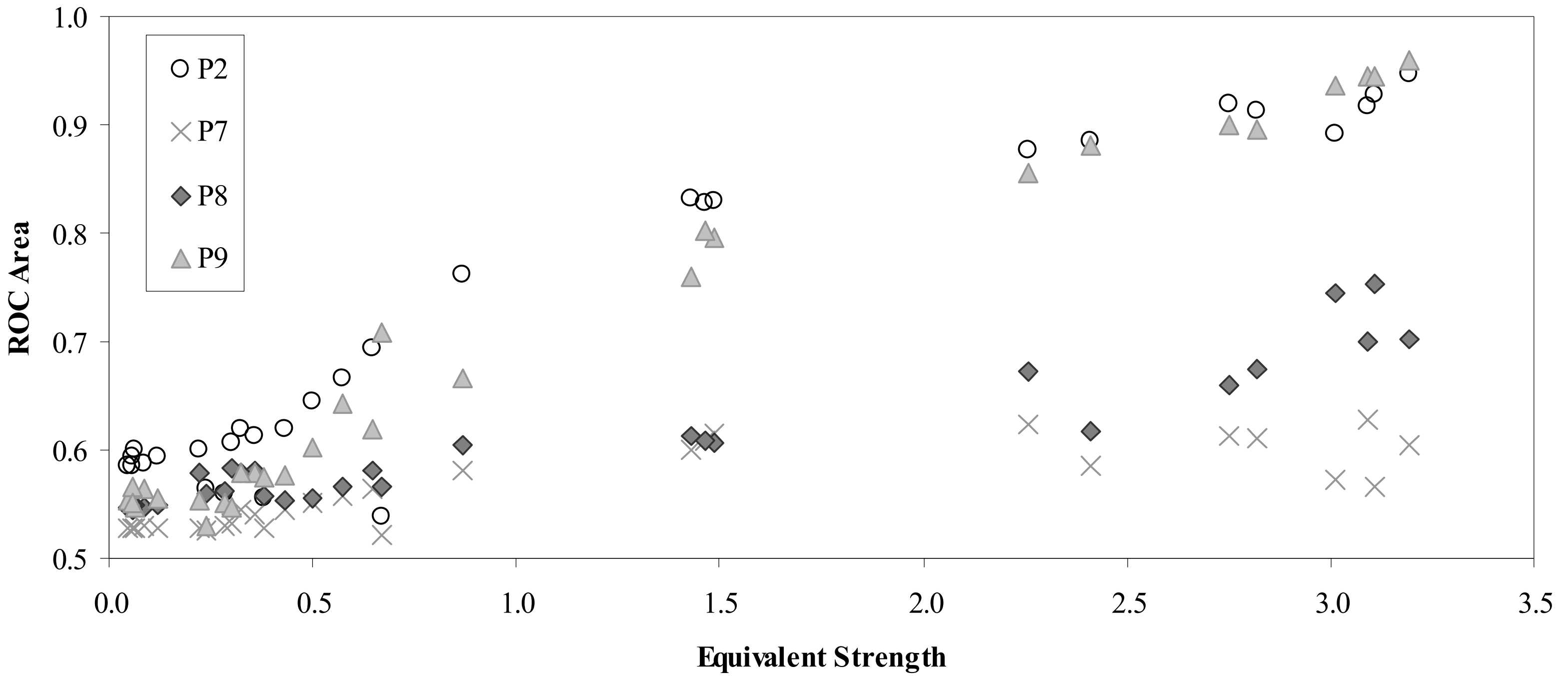


EPA T&E Events Results

	Level 1 Concentration					Level 2 Concentration				
Contaminant	**ES**	**P2**	**P7**	**P8**	**P9**	**ES**	**P2**	**P7**	**P8**	**P9**
Aldicarb	2.82	0.912	0.610	0.675	0.897	1.47	0.827	0.608	0.609	0.803
Dicamba	0.29	0.560	0.529	0.561	0.550	0.24	0.563	0.526	0.561	0.529
DMSO	0.57	0.667	0.558	0.567	0.643	0.36	0.612	0.540	0.580	0.578
E. coli	3.01	0.891	0.572	0.746	0.936	2.41	0.886	0.585	0.617	0.882
Glyphosate	3.20	0.947	0.605	0.703	0.961	3.11	0.928	0.566	0.753	0.945
Lead nitrate	0.05	0.586	0.527	0.547	0.554	0.06	0.594	0.529	0.545	0.567
Mercuric chloride	0.67	0.539	0.522	0.566	0.708	0.38	0.555	0.528	0.558	0.574
Nicotine	1.48	0.829	0.615	0.607	0.796	0.87	0.762	0.580	0.604	0.667
Nutrient broth	0.43	0.619	0.545	0.554	0.577	0.22	0.599	0.528	0.579	0.552
Potassium ferricyanide	0.08	0.587	0.531	0.547	0.564	0.12	0.593	0.528	0.550	0.555
Sodium hyposulfite	3.09	0.919	0.614	0.659	0.900	2.26	0.832	0.600	0.612	0.759
Sodium thiosulfate	2.75	0.916	0.628	0.699	0.944	1.43	0.876	0.624	0.672	0.854
Sucrose	0.06	0.601	0.528	0.549	0.547	0.06	0.586	0.528	0.549	0.552
Terrific broth	0.64	0.694	0.564	0.580	0.618	0.32	0.619	0.544	0.578	0.579
Tryptic soy broth	0.50	0.645	0.552	0.555	0.601	0.30	0.607	0.532	0.584	0.546

EPA T&E Events Results

Ability to accurately identify events is a strong function of the event strength (both strengths shown) and also the variability of the background water quality at the sensor location



Summary

- Algorithms developed to predict water quality at next time step
  - Large deviation between expected and measured water quality indicates a change
- Proportion of events identified cannot be considered independently of proportion of false positives
- Different contaminants produce different water quality responses

Next Steps

- Work to date has focused on change/event detection at a single location
  - Move to multiple locations providing a “network-wide” determination of an event
- Always looking for ways to test these approaches in real systems
  - Complaint data have proven problematic
  - Tracer tests? Known maintenance issues?
- Interested in feedback on “dual-use” capability of these tools.

Additional Reading

- Klise, K.A. and S.A. McKenna, 2006, Water quality change detection: multivariate algorithms, in Proceedings of SPIE (International Society for Optical Engineering), Defense and Security Symposium 2006, April 18-20, Orlando. Florida, 9pp.
- McKenna, S.A., K.A. Klise and M.P. Wilson, 2006, Testing Water Quality Change Detection Algorithms, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium WDSA’06, Cincinnati, OH, August 27-30, 2006.
- Klise, K.A. and S.A. McKenna, 2006, Multivariate Applications for Detecting Anomalous Water Quality, in Proceedings of the 8th Annual Water Distribution System Analysis Symposium, Cincinnati (WDSA ’06), OH, August 27-30, 2006.