

# Continuous monitoring of river water quality using in situ monitoring stations: Efficient data quality assessment

M3 Workshop

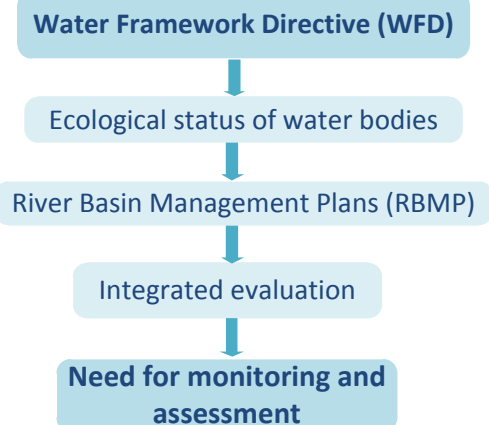
Cologne, Germany

14-15 06 2012

Janelcy Alferes, Pascal Poirier and Peter Vanrolleghem



## Problem definition



2



## Problem definition

- In situ monitoring stations
  - WFD investigative monitoring.
  - High sampling frequency, huge data sets
- Reliability of sensors insufficient
- Real data are mostly noisy



Data quality assessment crucial for practical use: ex. “modelling”

## In situ monitoring stations

- Urban river (QC, Canada)
- Water quality variables:
  - pH, °C,  $\mu\text{S}/\text{cm}$ ,  $\text{O}_2$ ,  $\text{NH}_4$ ,  $\text{NO}_3$ ,  $\text{K}^+$ , TSS, COD...
- Sample time: 5-60 sec  
More than 1 million measurements per variable!



Representative data??

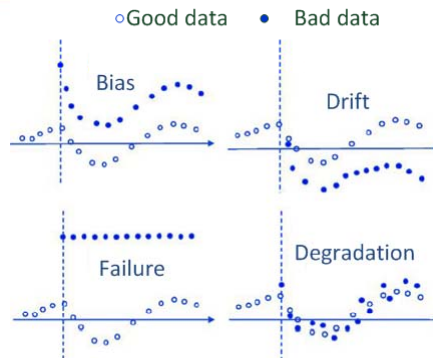
## In situ monitoring stations

- Some observations:

- Faulty probes
- Practical issues
- Maintenance
- Low water level



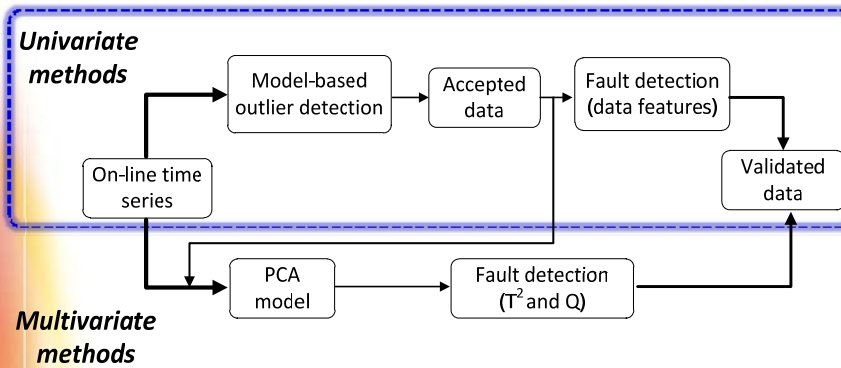
Data reliability can decrease



## Data quality assessment tools

- Software tools for automatic data quality evaluation
- Using time series information

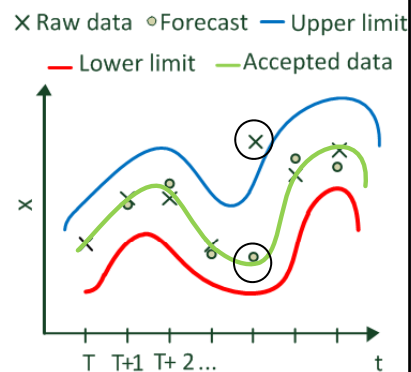
## Data quality assessment tools



## Data quality assessment tools

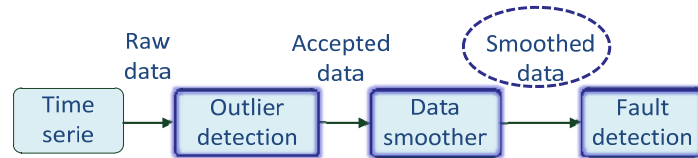
- Univariate methods
  - Outlier detection
  - Forecasting of variable  $\hat{x}$  and std of error  $\hat{\sigma}_e$
  - Prediction interval:

$$x_{lim} = \hat{x} \pm K \cdot \hat{\sigma}_e$$



## Data quality assessment tools

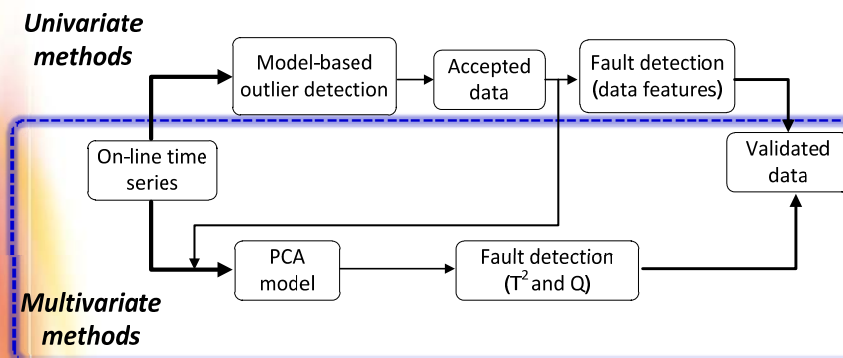
### Univariate methods



### Fault detection --> data features and their limits:

- % replaced data
- Slope
- Residual standard deviation (RSD)

## Data quality assessment tools



## Data quality assessment tools

- Multivariate methods
  - Large multidimensional dataset  $X$ , redundant and correlated.
  - Allow to reduce the dimension of  $X$  by identifying “key” variables.

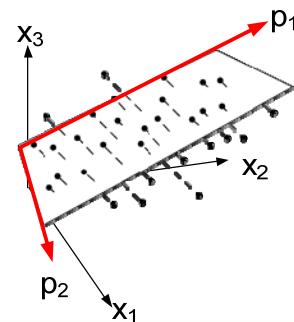
➡ New coordinated system

## Data quality assessment tools

- Multivariate methods
  - Let be  $n$  variables ( $x_1 \dots x_n$ ) and  $m$  samples
  - New variables ( $p_1 \dots p_a$ ) as linear combinations:

$$\begin{aligned} p_1 &= c_{11}x_1 + c_{12}x_2 + \dots + c_{1n}x_n \\ p_2 &= c_{21}x_1 + c_{22}x_2 + \dots + c_{2n}x_n \\ p_a &= c_{a1}x_1 + c_{a2}x_2 + \dots + c_{an}x_n \end{aligned}$$

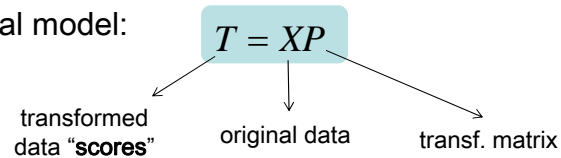
- Axes of a new coordinate system
- Directions of max. variability



## Data quality assessment tools

- Multivariate methods

- Final model:

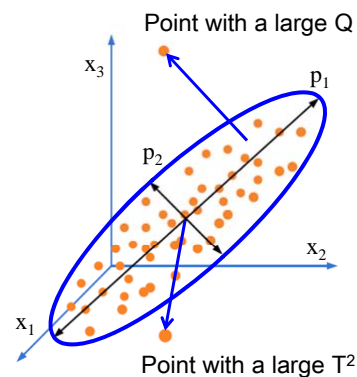


- Fault detection using the model by monitoring two statistics ( $T^2$ ,  $Q$ )

## Data quality assessment tools

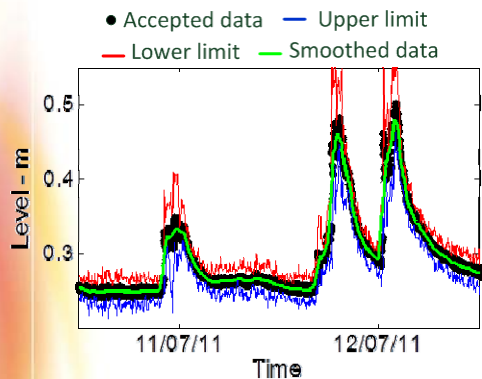
- Multivariate methods

- $T^2$ : normalized sum of scores: variations within the model
  - $Q$ : sum of squared residuals: goodness of fit of samples to the model
  - Detection limits are defined.



## Some preliminary results

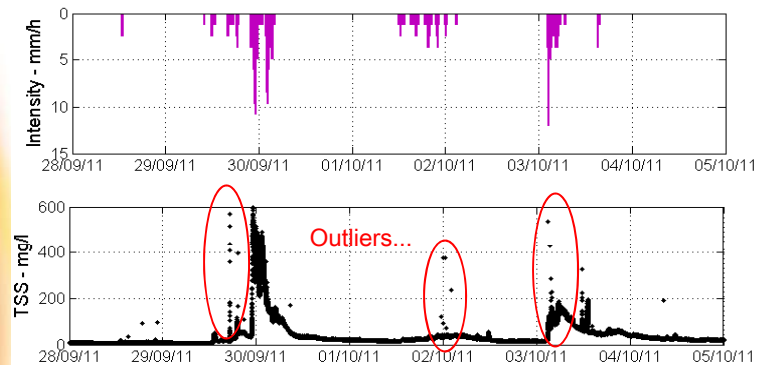
### ■ Univariate methods (Level)



- Variable prediction interval
- Adaptation to time-varying hydraulics

## Some preliminary results

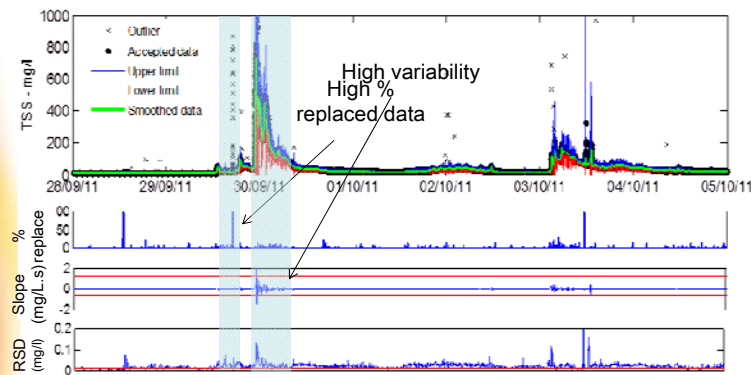
### ■ Univariate methods (TSS)





## Some preliminary results

### ■ Univariate methods (TSS)



## Some preliminary results

### ■ Multivariate methods

- Dataset with 8 variables
  - NTU, NO<sub>3</sub>, TOC, DOC, pH, K<sup>+</sup>, NH<sub>4</sub> and °C
- Obtaining the model
  - Two new variables (p<sub>1</sub>, p<sub>2</sub>) capture the max variability.
  - New space and transformed data ("scores"):

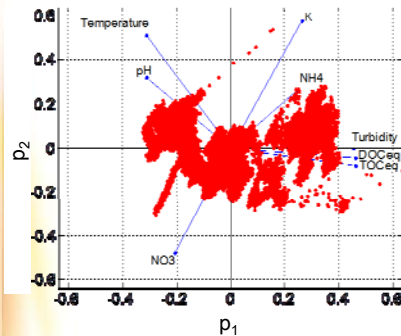
$$T = XP$$

- Monitoring of T<sup>2</sup> and Q
  - Define acceptability limits

## Some preliminary results

### ■ Multivariate methods

Representation of data in the new space

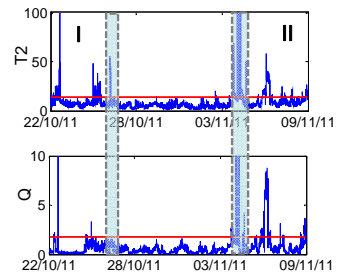
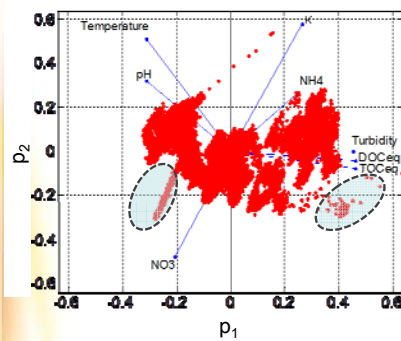


- Vectors represent the variables and their contribution to  $p_1$  and  $p_2$
- Each point corresponds to a sample in the new space (scores)

## Some preliminary results

### ■ Multivariate methods

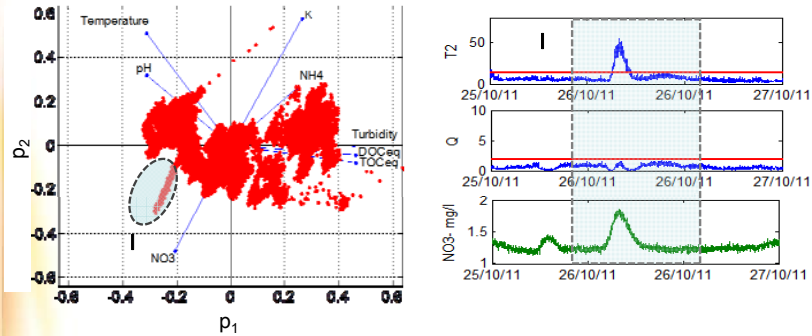
Representation of data in the new space



## Some preliminary results

- Multivariate methods

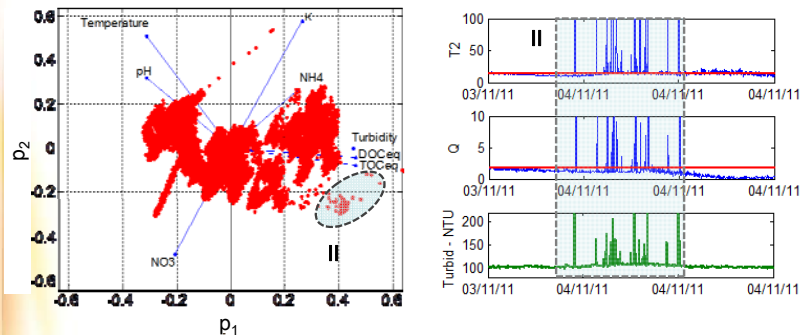
Representation of data in the new space



## Some preliminary results

- Multivariate methods

Representation of data in the new space



## Conclusions

- Data quality assessment tools satisfactory validated
- Methods allowed the detection of individual and multiple faults
- Validated data can be used within the WFD (river models, basin management...)
- Next steps: data coming from different stations

## Acknowledgement

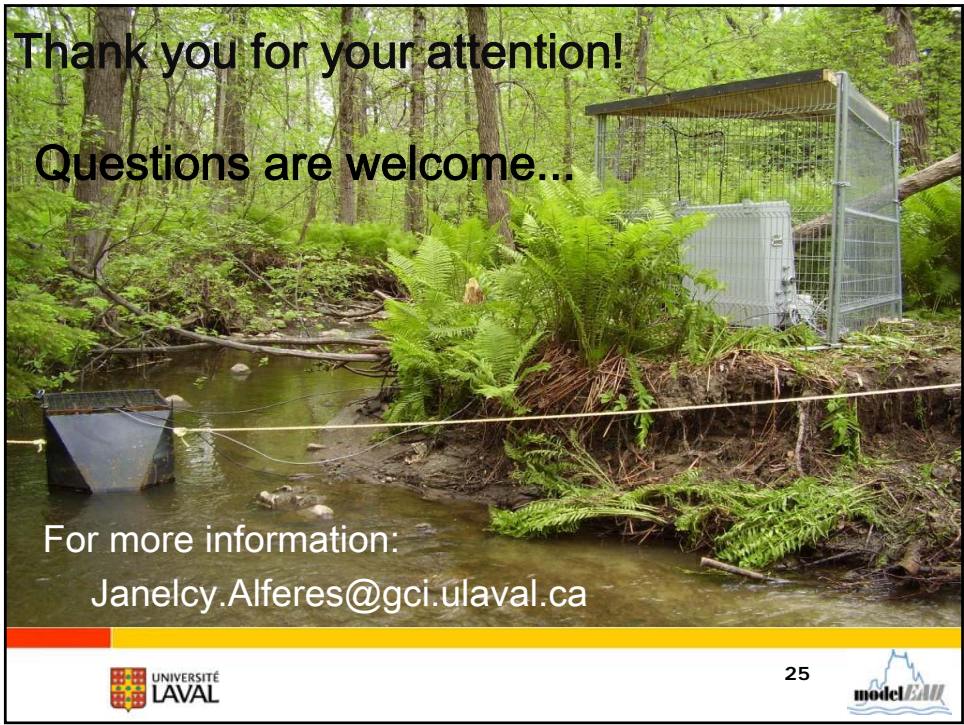



*Canada Research Chair  
in Water Quality Modeling*




**Thank you for your attention!**  
**Questions are welcome...**

For more information:  
[Janelcy.Alferes@gci.ulaval.ca](mailto:Janelcy.Alferes@gci.ulaval.ca)

A photograph of a stream in a forest. In the foreground, a metal cage is partially submerged in the water, with a net attached to it. In the background, a larger metal cage is situated on the bank, surrounded by lush green ferns and trees. The scene is a natural, wooded environment.

 UNIVERSITÉ  
LAVAL

25

 model