Statistical Validation of Water Quality Data

Lieven Clement, Olivier Thas, Peter Vanrolleghem and Jean-Pierre Ottoy

Department of Applied Mathematics, Biometrics and Process Control, Ghent University, Belgium

Outline

- Background
- Problem definition
- Methodology
- Validation in a simulation study
- Evaluation of deviating measurements
- Case study
- Implementation
- Conclusions

Background:
European Water Framework Directive

- Goal: maintain and improve aquatic environment

- Essential step: registration of the water quality

  - In Flanders, task performed by Flemish Environmental Protection Agency (VMM)

  - VMM established several monitoring networks

Problem Definition:
Data validation of the physico-chemical monitoring network

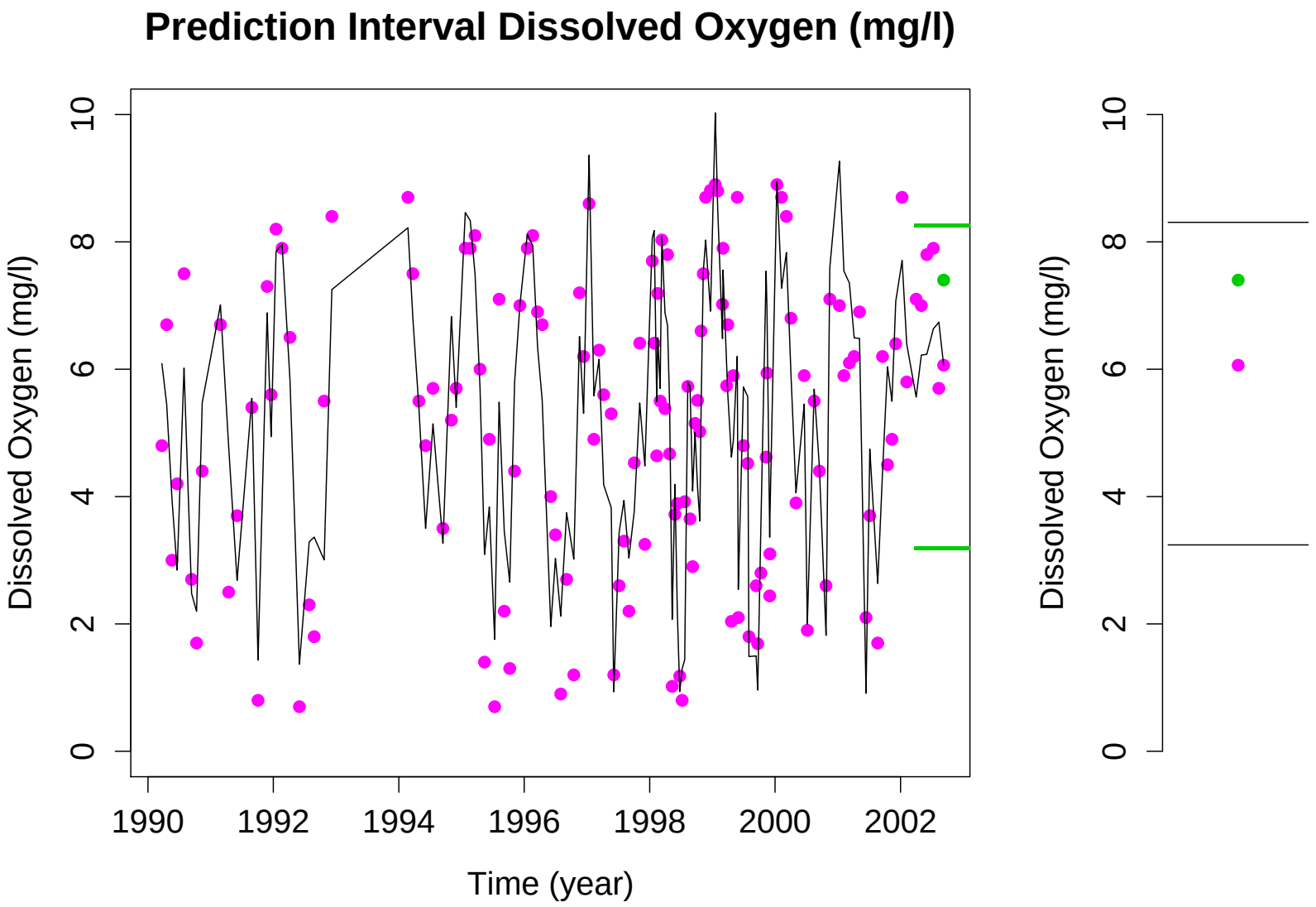
- 1317 measuring points located in different catchment area's of Flanders
- Every location is evaluated 12-26 times a year on a basic spectrum of physico-chemical variables
- Hugh amount of data to validate and to interpret
- It is important to detect possible anomalies, which can indicate water quality losses
- Data validation is carried out by human experts

Need for semi-automatic data validation procedure

- Automatic validation: data accepted $\Rightarrow$ addition to Dbase
- Deviation data is passed on to expert for further evaluation

Data validation: the statistical translation

- Fit model with historical data t=1..n
- Construct prediction and prediction interval at t=n+1
- Evaluate if new measurement is covered by prediction interval


Prediction Interval Dissolved Oxygen (mg/l)
Dissolved Oxygen (mg/l)
0
2
4
6
8
10
1990
1992
1994
1996
1998
2000
2002
Time (year)
Dissolved Oxygen (mg/l)
0
2
4
6
8
10

Methodology:

- The GAM-model Family
- Model selection
- Bootstrap prediction intervals

Methodology:
GAM-Family

- Flexible model needed:
  - Adaptation to changing environment
  - Valid for different rivers and different locations
  - Able to model non-linear relations
  - No intervention of the user needed

- A flexible model family was chosen: GAM-models

GAM-Family: Model Structure

- Model structure

$$Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \delta$$

- f_j: Local linear regression smoother

- No analytical solution: back-fitting algorithm is used

GAM-Family: Model Structure Characterization

- Model structure: Long term trend, Seasonal trend, several physico-chemical variables are selected

- Model selection is performed at every location and at each time-step

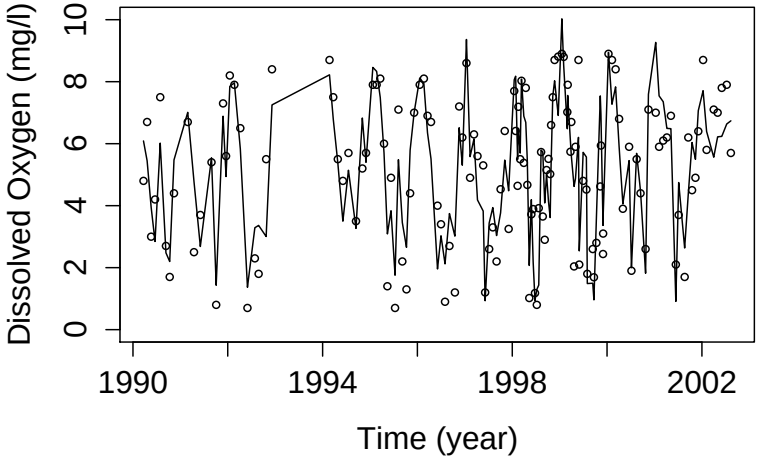
- Selection Statistic:

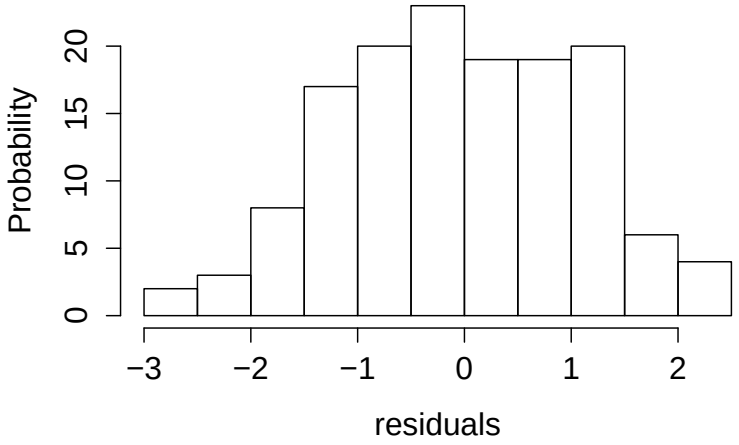
$$AIC = \frac{1}{n} \sum_{i=1}^{n} D(y_i, \widehat{y_i}) + 2\frac{df\phi}{n}$$

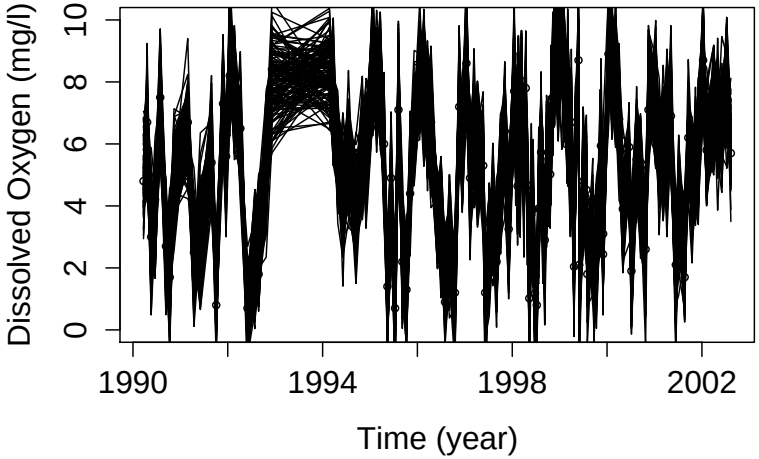
Methodology:
Bootstrap Prediction Intervals

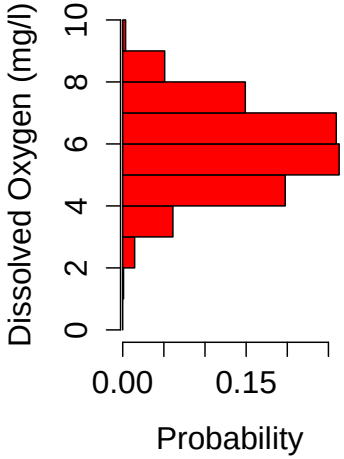
- GAM-model: no analytical solution of confidence intervals

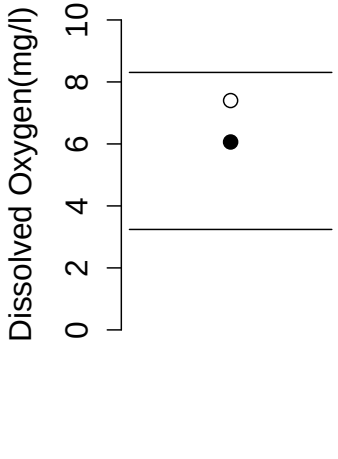
- Bootstrap methodology is used
  - New dataset: sample multivariate data with replacement
  - Fit GAM-model
  - Perform prediction
  - add random residual to the prediction
  - repeat this procedure


Dissolved Oxygen (mg/l)
0
2
4
6
8
10
1990
1994
1998
2002
Time (year)


Probability
0
5
10
15
20
−3
−2
−1
0
1
2
residuals


Dissolved Oxygen (mg/l)
0
2
4
6
8
10
1990
1994
1998
2002
Time (year)

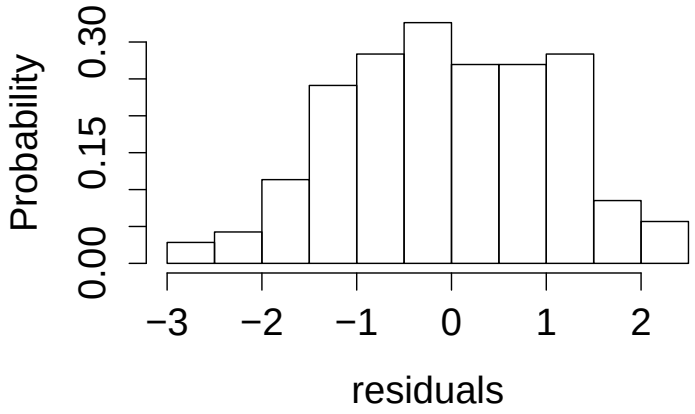

Dissolved Oxygen (mg/l)
0
2
4
6
8
10
0.00
0.15
Probability

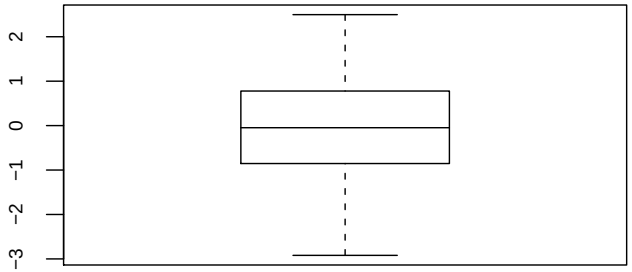

Dissolved Oxygen(mg/l)
0
2
4
6
8
10

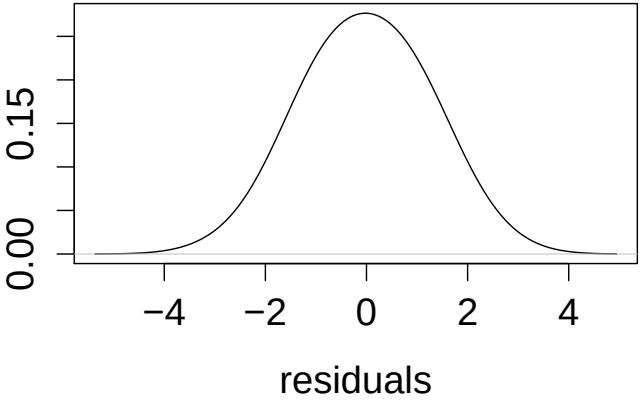
Validation:
Simulation study

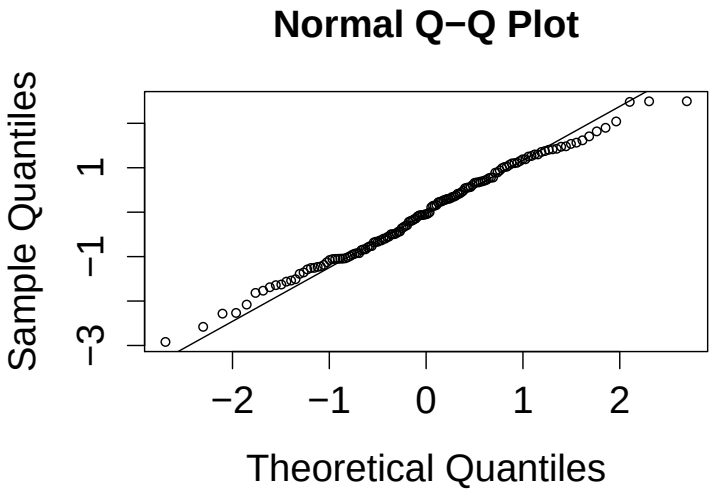
- Data generation
  - Model Fit: obtain $\widehat{Y}$
  - Estimate variance of errors: s^2
  - 10000 new datasets: predictor variables and new response $Y = \widehat{Y} + N(0, s^2)$

- Observed coverage of 95% bootstrap prediction intervals = 95.7%.

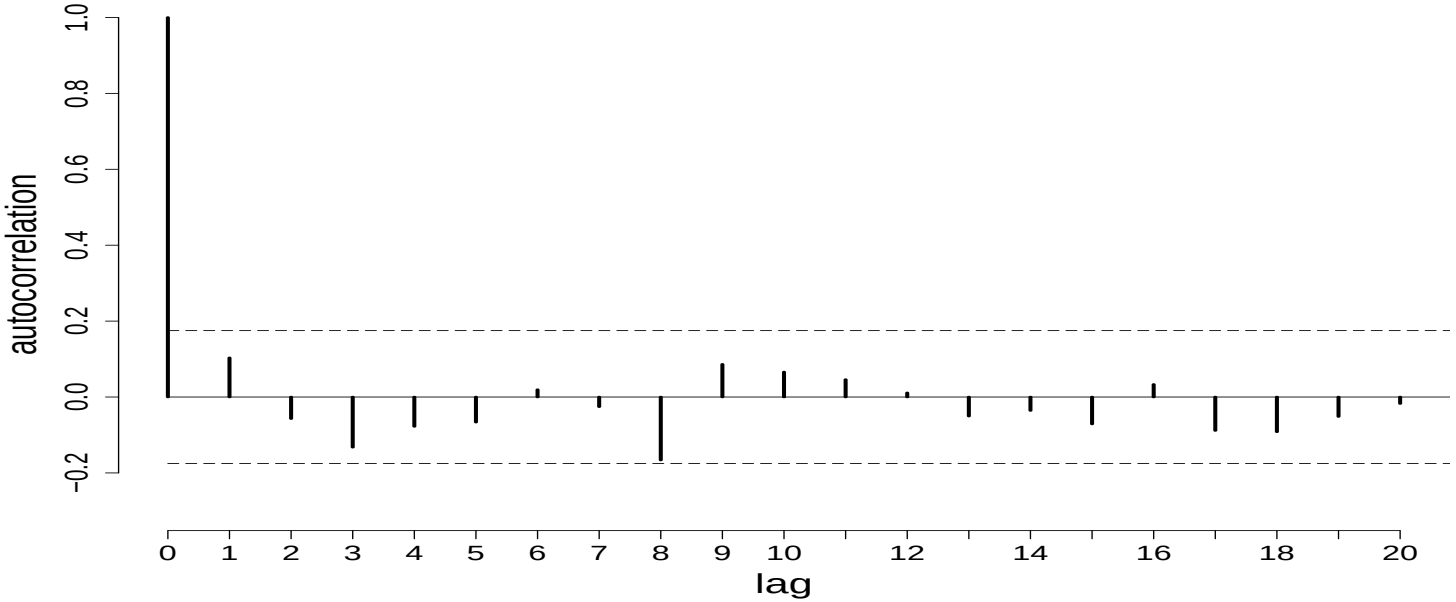

Probability
0.00
0.15
0.30
−3
−2
−1
0
1
2
residuals


−3
−2
−1
0
1
2


0.00
0.15
−4
−2
0
2
4
residuals


Normal Q−Q Plot
Sample Quantiles
−3
−1
1
−2
−1
0
1
2
Theoretical Quantiles

- No problems with serial correlation occurred
  no moving block bootstrap was needed



- Possible explanations: Serial correlation could be partially covered in predictor variables
  - Seasonal trend and long term trend in the model
  - Other physico-chemical variables

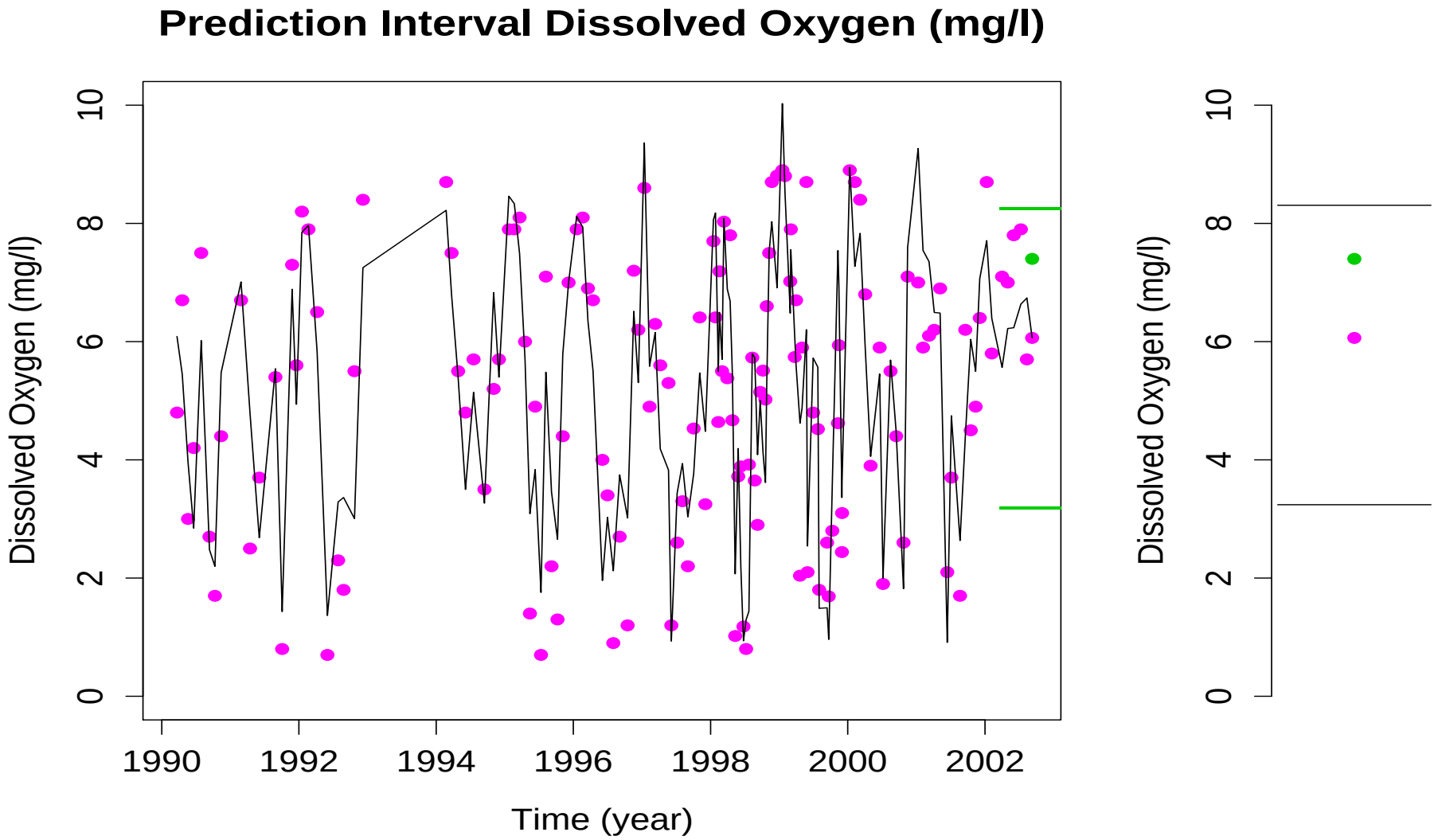
Evaluation of deviating measurements

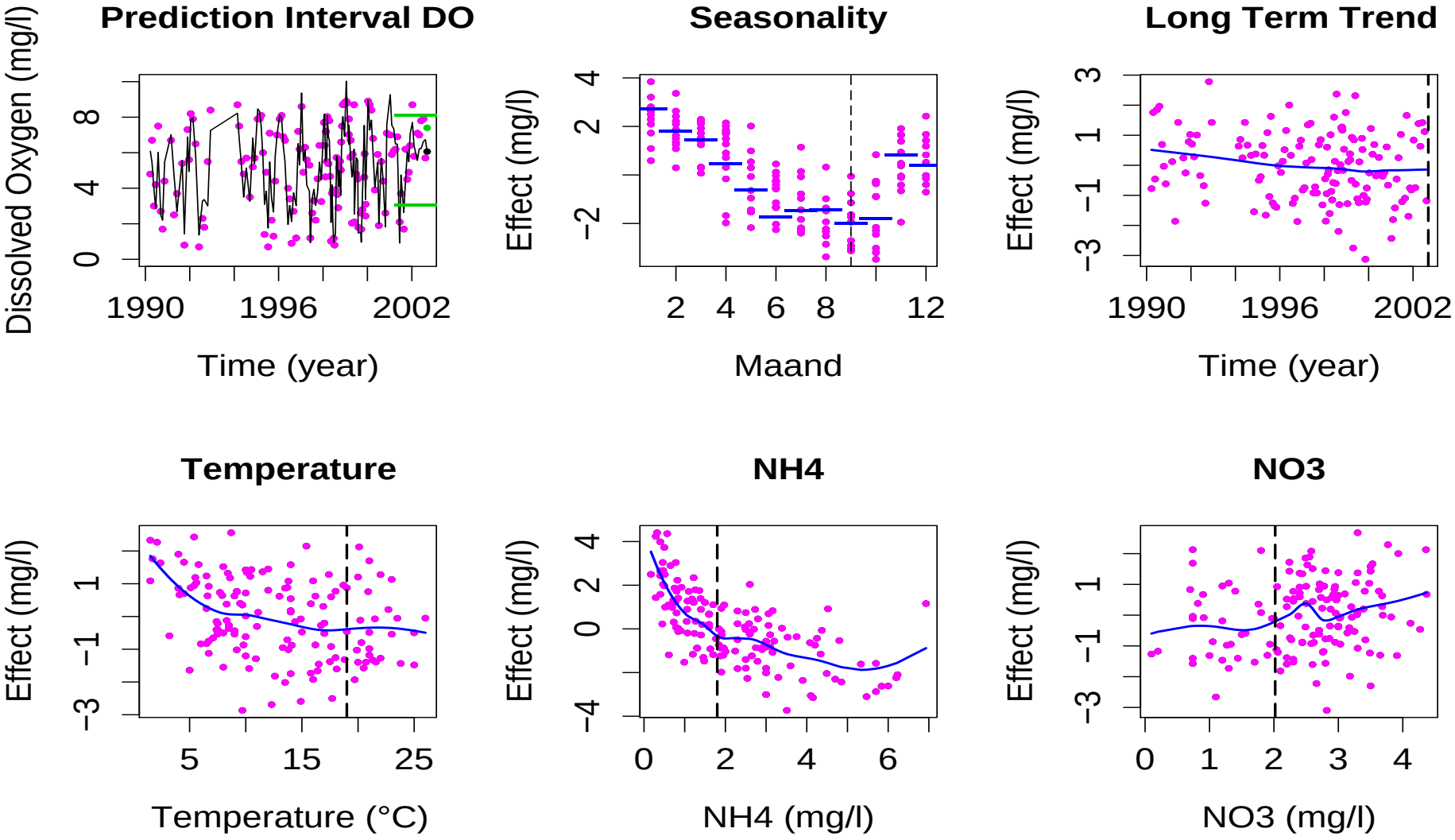
- Possible causes of deviating measurements
  - Due to the measurement
  - Due to a change in the system
  - Due to deviations in the predictor variables

- Detection of deviations in the predictor variables
  - Leave predictor variables one by one out the model
  - Construct prediction interval
  - Evaluate if measurement is covered by the new interval

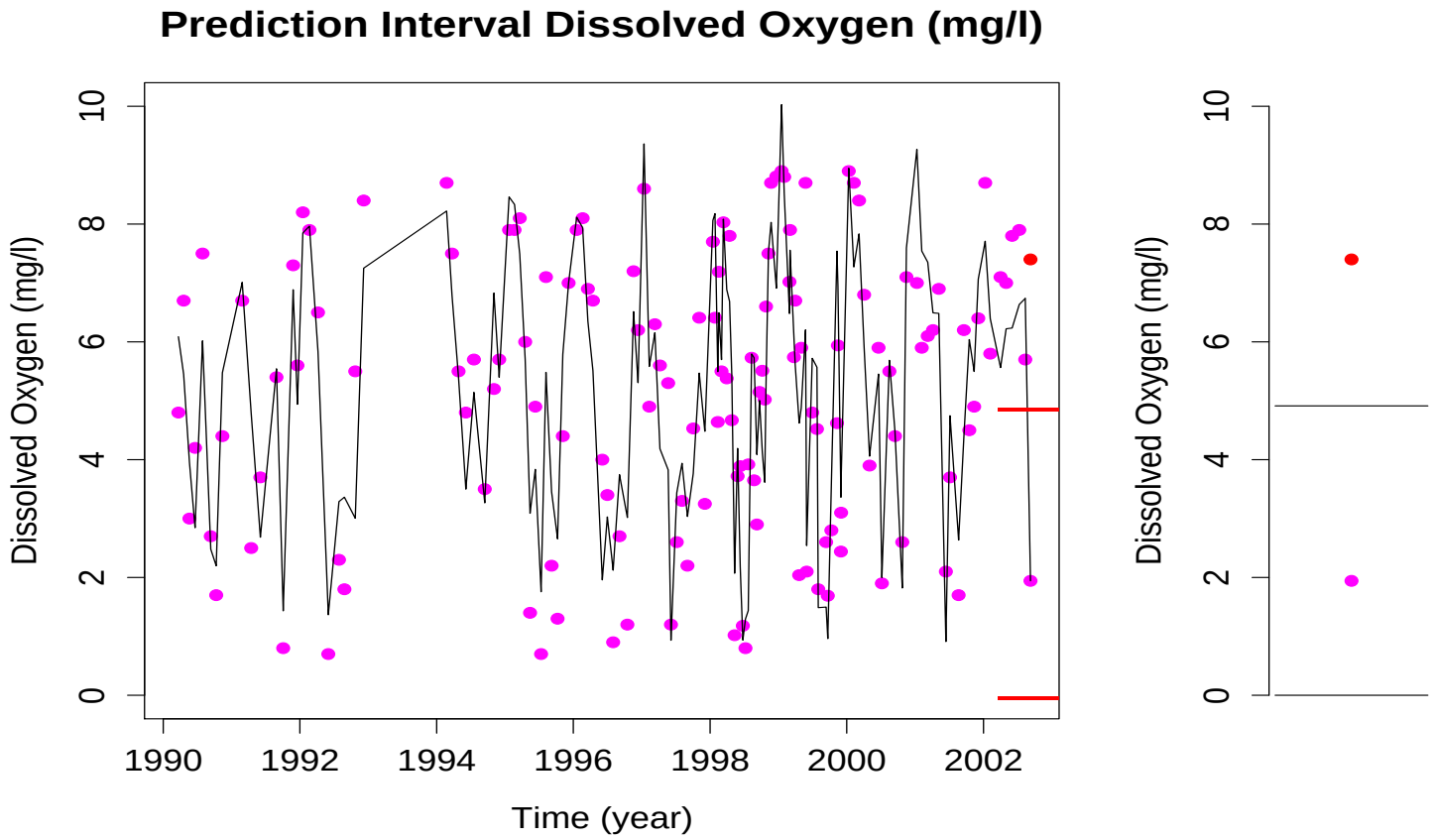
Case study

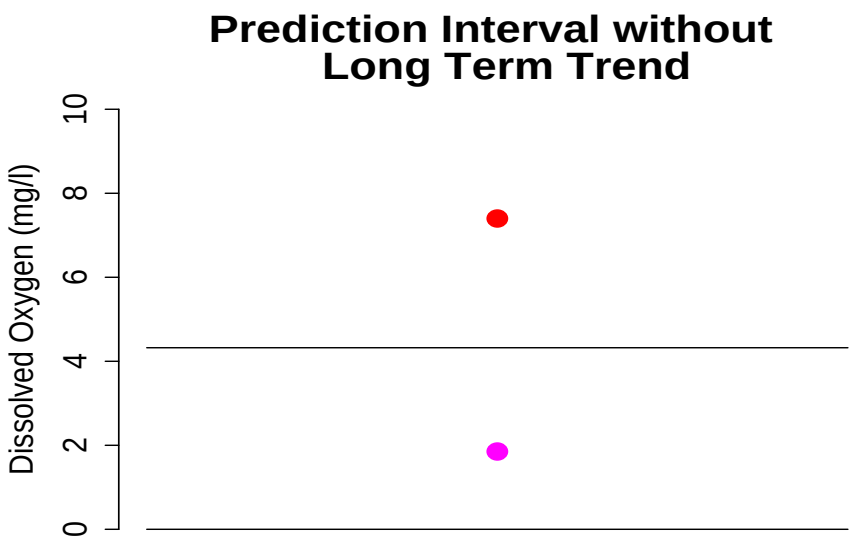
- Data validation of Dissolved Oxygen
- Introduction of error in prediction variable

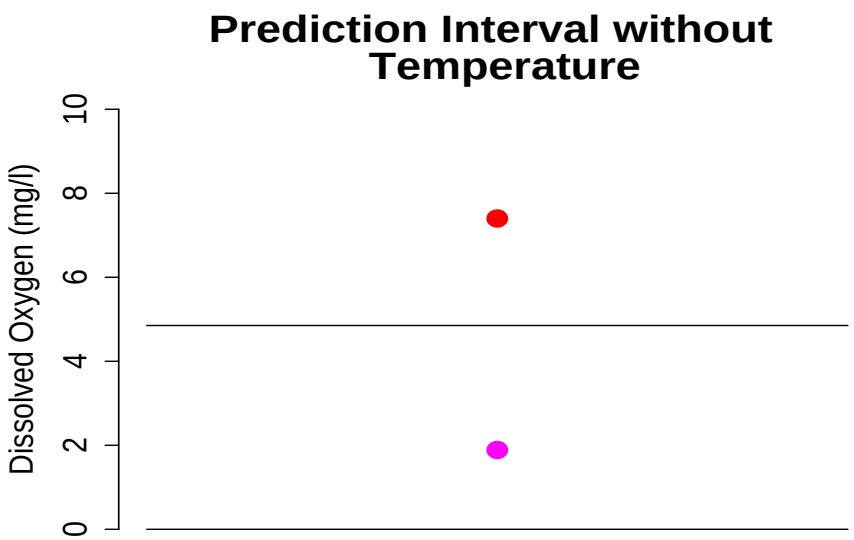

Prediction Interval Dissolved Oxygen (mg/l)
Dissolved Oxygen (mg/l)
0
2
4
6
8
10
1990
1992
1994
1996
1998
2000
2002
Time (year)
Dissolved Oxygen (mg/l)
0
2
4
6
8
10

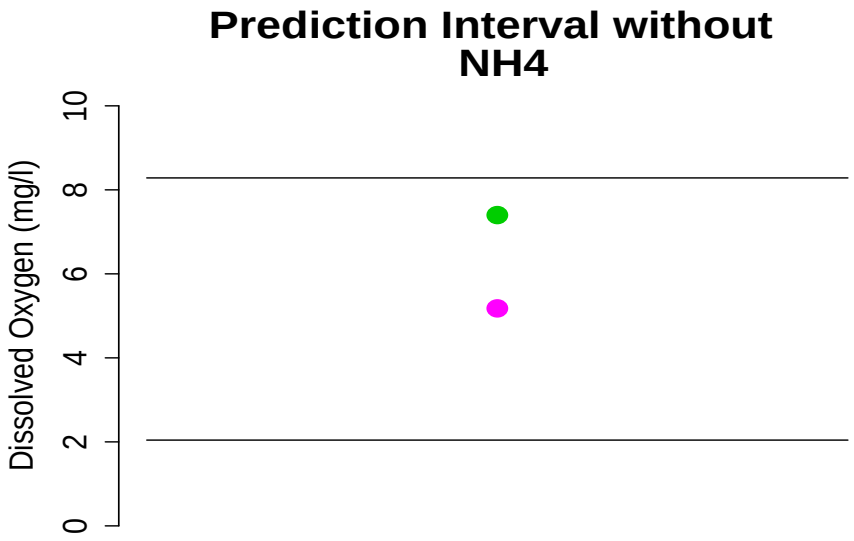

Prediction Interval DO
Dissolved Oxygen (mg/l)
0
4
8
1990
1996
2002
Time (year)
Seasonality
Effect (mg/l)
−2
2
4
2
4
6
8
12
Maand
Long Term Trend
Effect (mg/l)
−3
−1
1
3
1990
1996
2002
Time (year)
Temperature
Effect (mg/l)
−3
−1
1
5
15
25
Temperature (°C)
NH4
Effect (mg/l)
−4
0
2
4
0
2
4
6
NH4 (mg/l)
NO3
Effect (mg/l)
−3
−1
1
0
1
2
3
4
NO3 (mg/l)

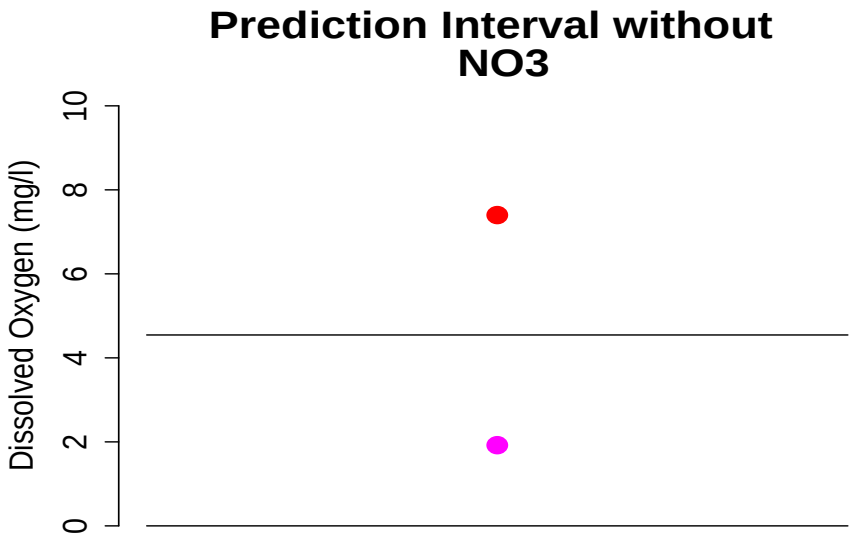
error introduced in NH4 concentration

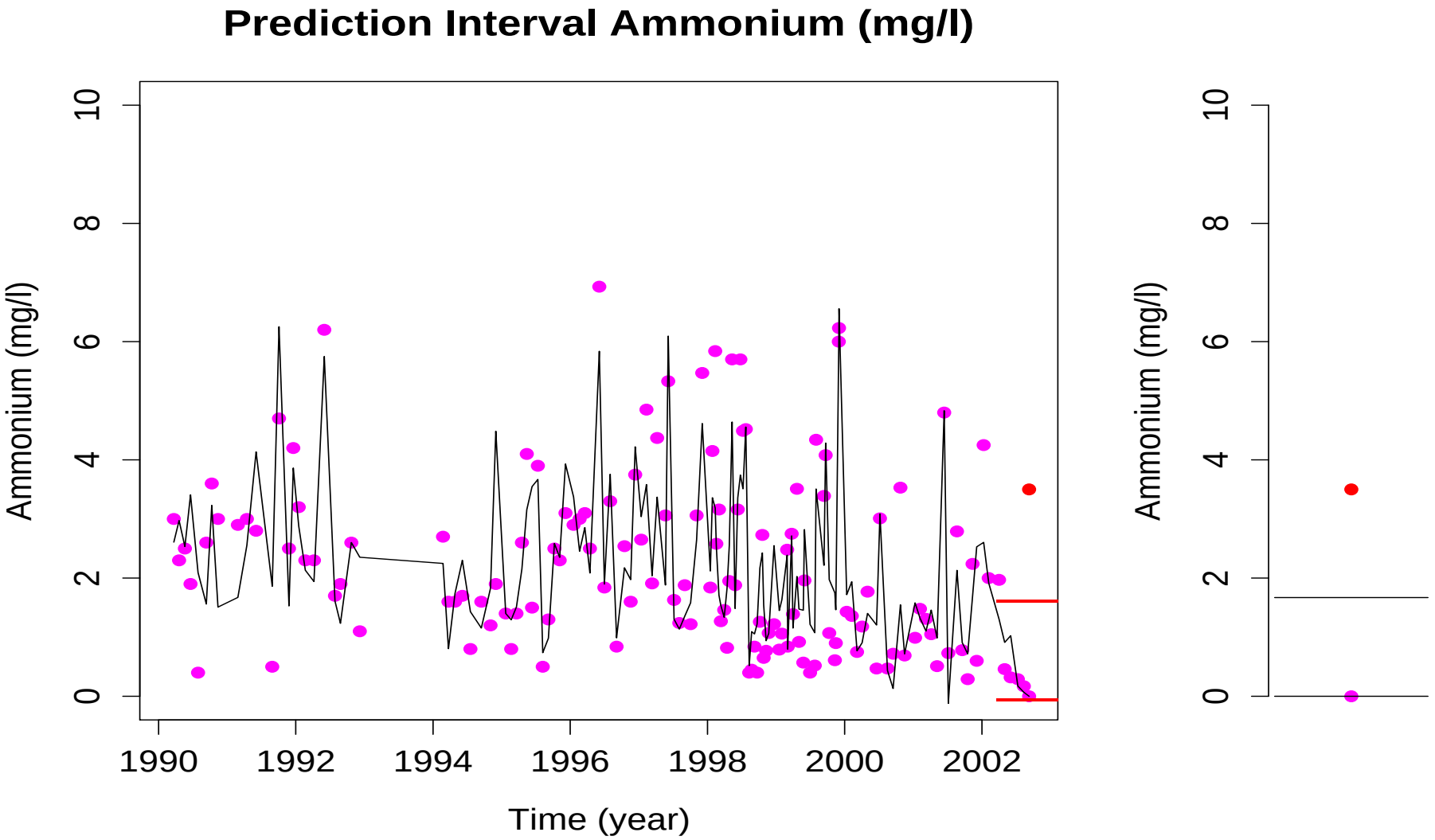


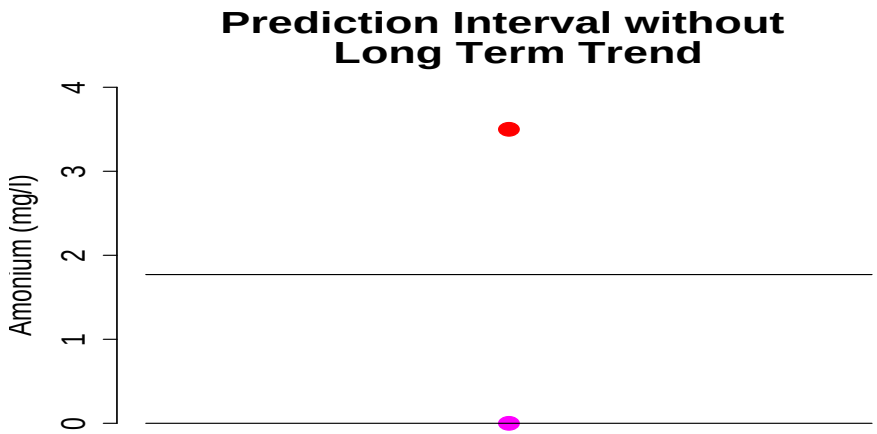

Prediction Interval without Long Term Trend
Dissolved Oxygen (mg/l)
0
2
4
6
8
10

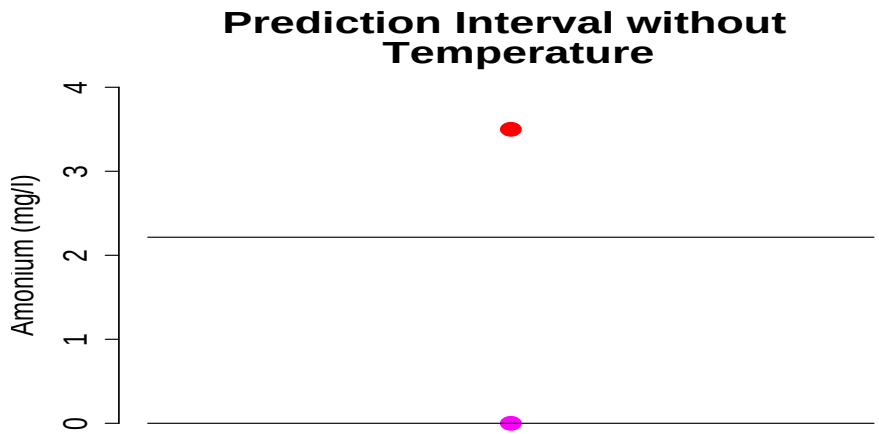

Prediction Interval without Temperature
Dissolved Oxygen (mg/l)
0
2
4
6
8
10

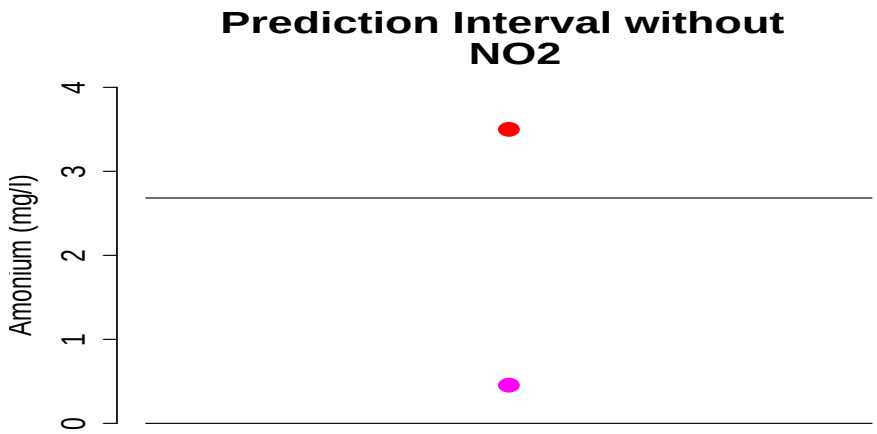

Prediction Interval without NH4
Dissolved Oxygen (mg/l)
0
2
4
6
8
10

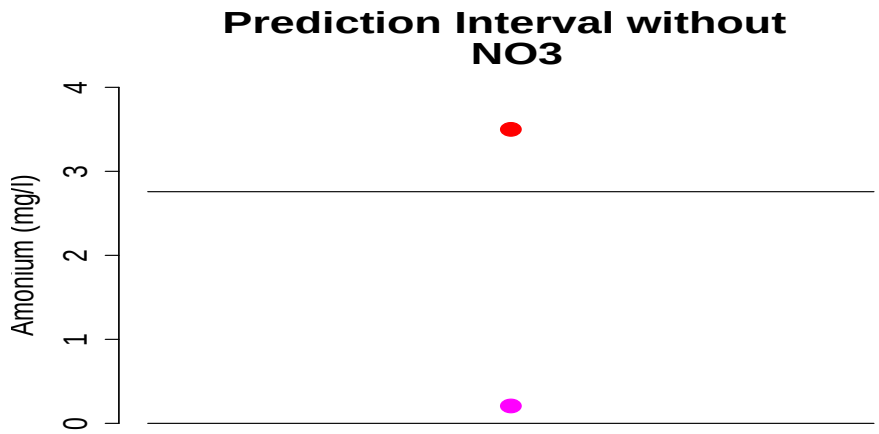

Prediction Interval without NO3
Dissolved Oxygen (mg/l)
0
2
4
6
8
10


Prediction Interval Ammonium (mg/l)
Ammonium (mg/l)
0
2
4
6
8
10
1990
1992
1994
1996
1998
2000
2002
Time (year)
Ammonium (mg/l)
0
2
4
6
8
10


Prediction Interval without Long Term Trend
Amonium (mg/l)
0
1
2
3
4


Prediction Interval without Temperature
Amonium (mg/l)
0
1
2
3
4


Prediction Interval without NO2
Amonium (mg/l)
0
1
2
3
4


Prediction Interval without NO3
Amonium (mg/l)
0
1
2
3
4

Practical Implementation

- The comming months the procedure will be implemented with the VMM

- Following loop will be implemented at each location:
  - Model structure characterization
  - Construction Prediction Intervals
  - Evaluation of new measurement

Conclusions:

- GAM Model Family:
  - Provide the flexibility needed for the semi-automatic procedure
  - Trends and relations between physico-chemical variables can be deduced
- Good coverage of the Prediction Intervals
- Methodology can give an indication why deviating measurements occur