

A framework for detecting anomalies in water-quality variables

Puwasala Gamakumara
ARCLP Workshop



Outline

- 1 Introduction
- 2 Framework
- 3 Data
- 4 Modeling
- 5 Outlier detection based on Extreme value theory
- 6 Evaluation
- 7 Conclusion

Outline

1 Introduction

2 Framework

3 Data

4 Modeling

5 Outlier detection based on Extreme value theory

6 Evaluation

7 Conclusion

Water-quality monitoring in river networks

- Low-cost in-situ sensors
- Produce high-frequency data
- Prone to errors due to miscalibration, biofouling, battery and technical errors

Objective

- Developing statistical tools to detect anomalies in water-quality variables measured by in-situ sensors
- Extend to utilising information from multiple sensors

Outline

1 Introduction

2 Framework

3 Data

4 Modeling

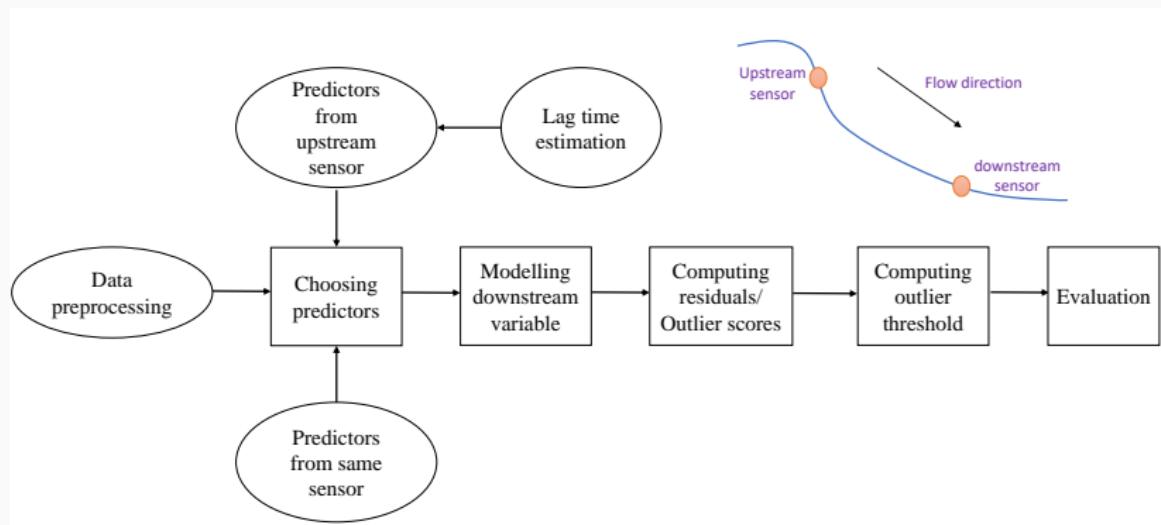
5 Outlier detection based on Extreme value theory

6 Evaluation

7 Conclusion

Anomaly detection framework

An anomaly is an observation that has an unexpectedly low conditional density



Outline

1 Introduction

2 Framework

3 Data

4 Modeling

5 Outlier detection based on Extreme value theory

6 Evaluation

7 Conclusion

Pringle Creek - Texas, USA

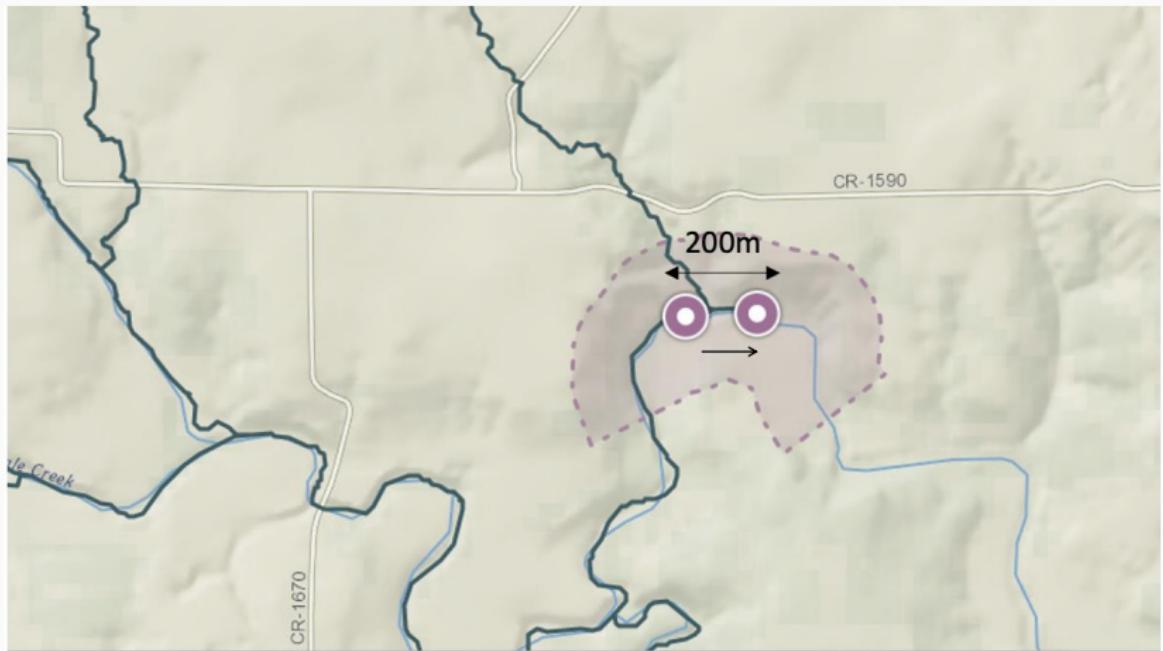
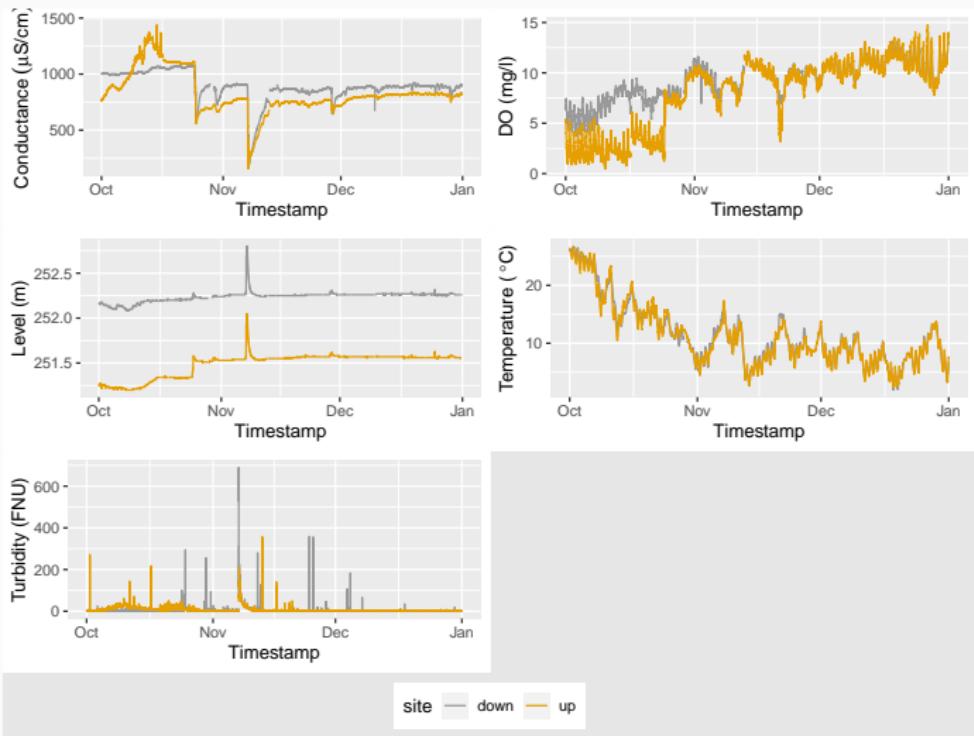


Figure 1: image courtesy neonscience.org

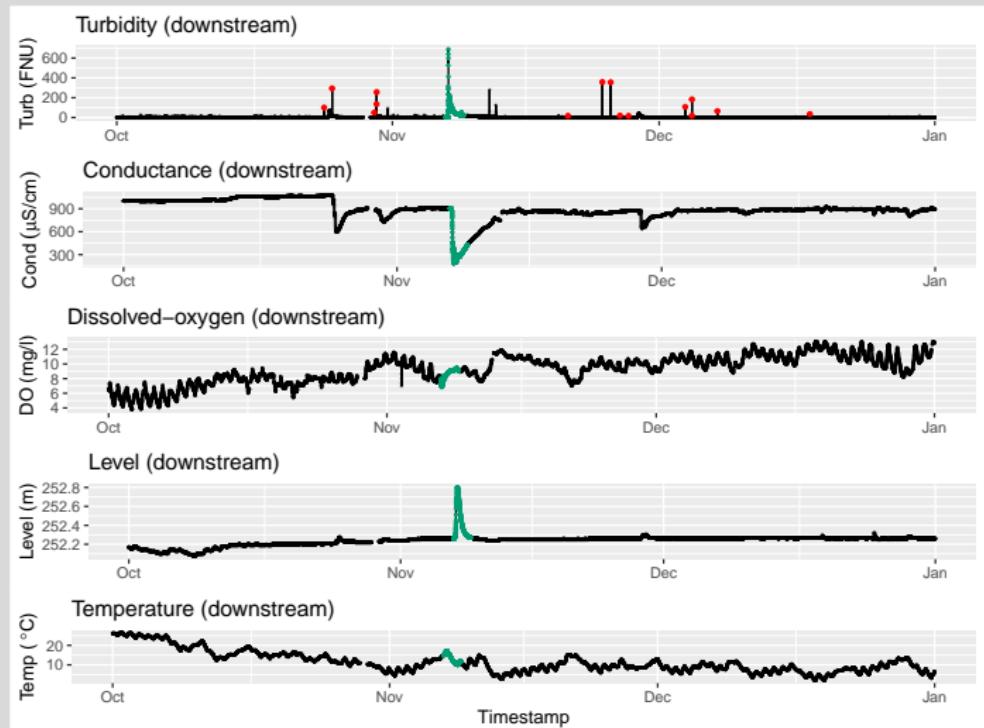
Data

- **Variables** - Turbidity, Conductivity, Dissolved oxygen, Level and Temperature
- **Time span** - 01-10-2019 to 31-12-2019
- **Frequency** - 5 minute intervals
- [https://data.neonscience.org/data-products]

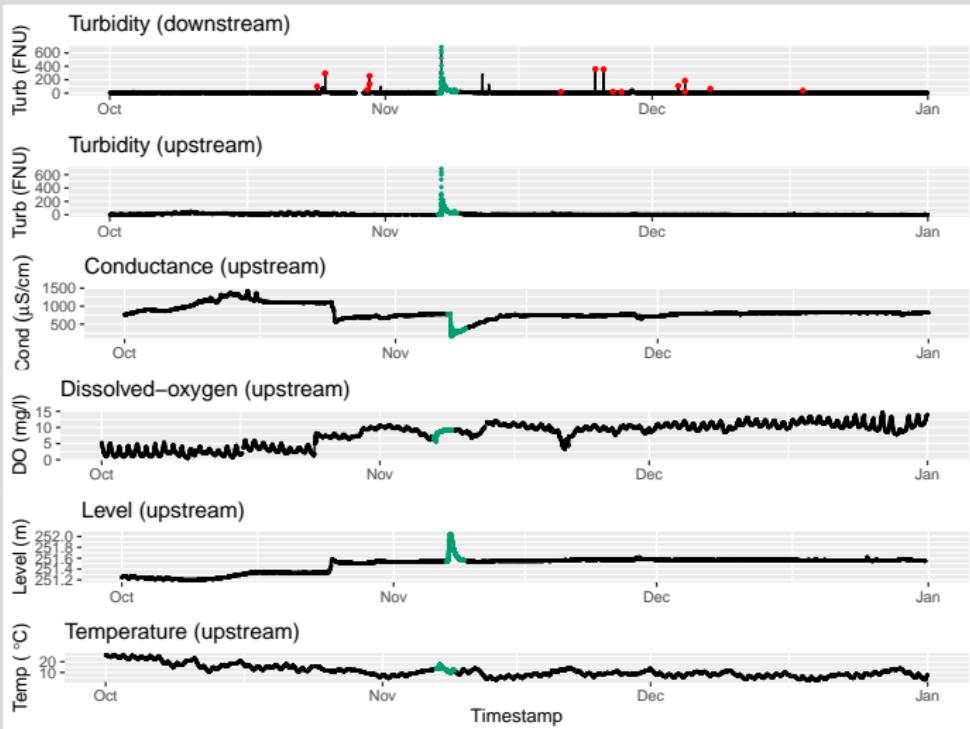
Time plots



Turbidity downstream vs other downstream variables



Turbidity downstream vs other upstream variables



Outline

1 Introduction

2 Framework

3 Data

4 Modeling

5 Outlier detection based on Extreme value theory

6 Evaluation

7 Conclusion

Modeling downstream turbidity

$$turbidity_down_t = \phi_0 + \sum_{i=1}^p g_i(z_{i,t_l}) + \sum_{j=1}^q h_j(turbidity_down_{t-j}) + \varepsilon_t$$

Choices for z_{i,t_l}

Model	Predictors
GAM-down	other downstream variables
GAM-down-AR	other downstream variables + lagged responses
GAM-up	upstream variables
GAM-up-AR	upstream variables + lagged responses
GAM-up-down	upstream variables + downstream variables

Lag time estimation

- Assume the lag time between two sensor locations depends on the upstream river behavior
- Use *conditional cross-correlations* to estimate the lag time
- let x_t : Turbidity upstream, y_t : Turbidity downstream and \mathbf{z}_t : {level upstream, temperature upstream}
- $x_t^* = \frac{x_t - E[x_t | \mathbf{z}_t]}{\sqrt{V[x_t | \mathbf{z}_t]}}$ and $y_t^* = \frac{y_t - E[y_t | \mathbf{z}_t]}{\sqrt{V[y_t | \mathbf{z}_t]}}$

Conditional cross-correlation

$$r_k(\mathbf{z}_t) = E[x_t^* y_{t+k}^* | \mathbf{z}_t] \quad \text{for } k = 1, 2, \dots$$

- To estimate $r_k(\mathbf{z}_t)$ we fit the following GAMs
- Let $x_t^* y_{t+k}^* | \mathbf{z}_t \sim N(r_k(\mathbf{z}_t), \sigma_r^2)$,

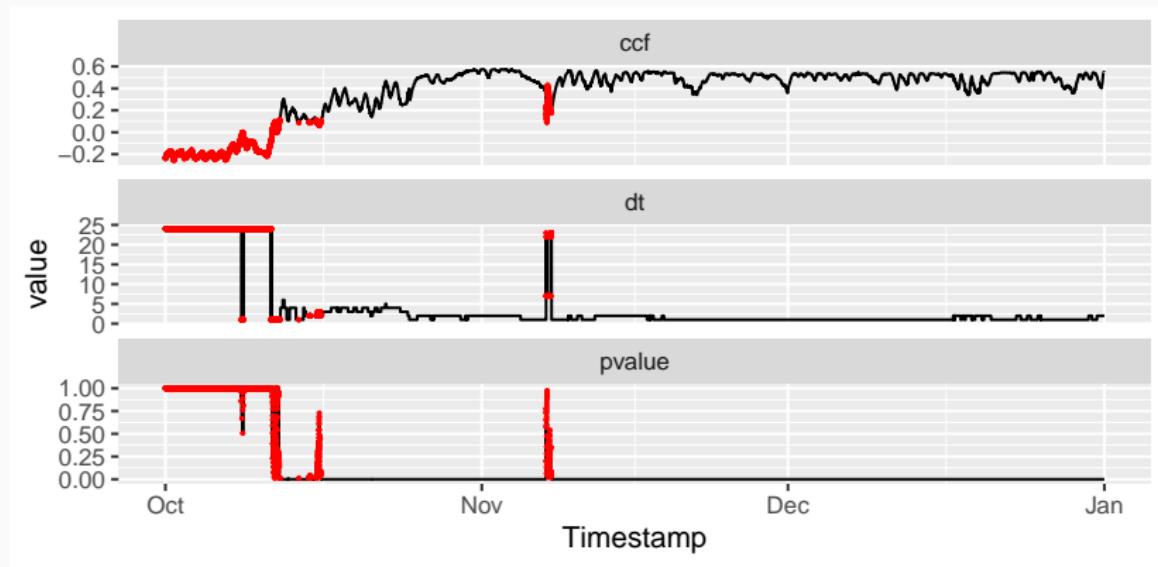
$$g(r_k(\mathbf{z}_t)) = \gamma_0 + \sum_{i=1}^p h_i(z_{i,t}) + \varepsilon_t$$

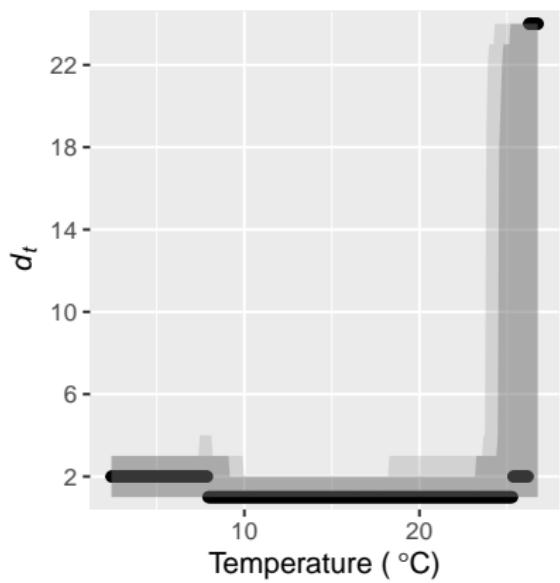
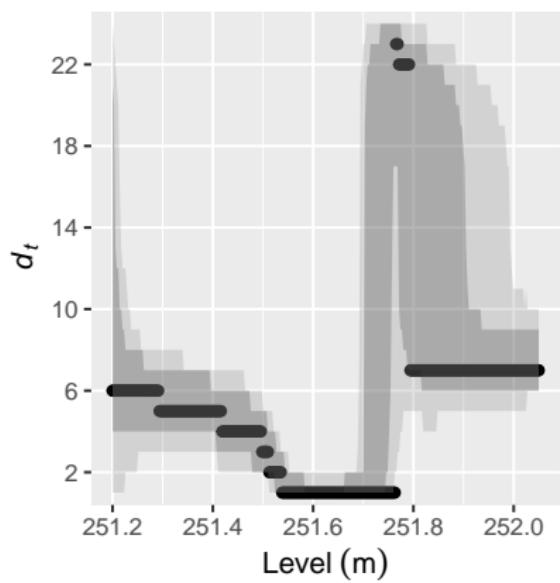
$$\hat{r}_k(\mathbf{z}_t) = g^{-1}(\hat{\gamma}_0 + \sum_{i=1}^p \hat{h}_i(z_{i,t}))$$

Estimating time delay

$$\hat{d}_t(\mathbf{z}_t) = \operatorname{argmax}_k \hat{r}_k(\mathbf{z}_t)$$

Conditional cross-correlation and lag time estimation

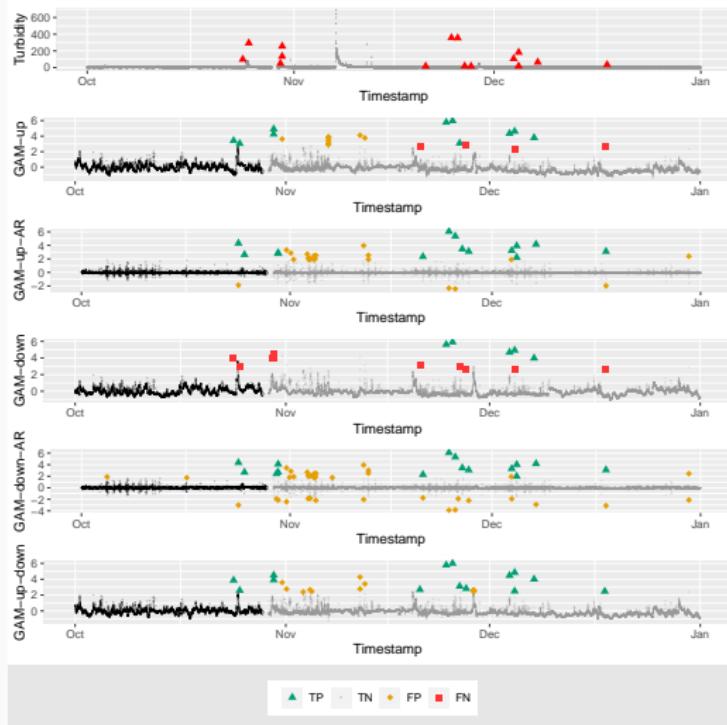




Outline

- 1 Introduction
- 2 Framework
- 3 Data
- 4 Modeling
- 5 Outlier detection based on Extreme value theory
- 6 Evaluation
- 7 Conclusion

Outlier detection using Peak over Threshold method



Outline

- 1 Introduction
- 2 Framework
- 3 Data
- 4 Modeling
- 5 Outlier detection based on Extreme value theory
- 6 Evaluation
- 7 Conclusion

Performance Evaluation

method	TP	TN	FP	FN	OP
GAM-down-AR	15	25948	39	0	0.9977
GAM-up-down	14	25916	14	1	0.9652
GAM-up-AR	14	25837	21	1	0.9651
GAM-up	10	25920	10	5	0.7996
stray($p=0.5$, $k=1$)	9	26007	5	6	0.7497
stray($p=0.5$, $k=5$)	6	26011	1	9	0.5711
stray($p=0.75$, $k=5$)	6	26011	1	9	0.5711
GAM-down	5	26012	0	10	0.4996
stray($p=0.75$, $k=1$)	15	9835	16177	0	-0.0728

Outline

1 Introduction

2 Framework

3 Data

4 Modeling

5 Outlier detection based on Extreme value theory

6 Evaluation

7 Conclusion

Conclusions

- A new framework to detect technical anomalies
- Uses temporal correlation between the water-quality variables to detect anomalies
- Utilising information from the nearby sensors improves the performance of the algorithm
- Place low-cost sensors close together

Thank You!

Slides: <https://github.com/PuwasalaG/ARCLP-workshop/tree/main/presentation>

 @PuwasalaG

 puwasala.gamakumara@gmail.com