

A framework for anomaly detection in river networks

Puwasala Gamakumara
ACEMS Retreat 2021



Rob Hyndman



Catherine Leigh



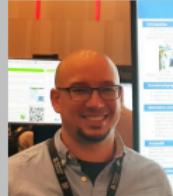
Dilini Talagala



Kerrie Mengersen



Erin Peterson



Edgar Santos-Fernandez



Katie Buchhorn

Water-quality monitoring in river networks

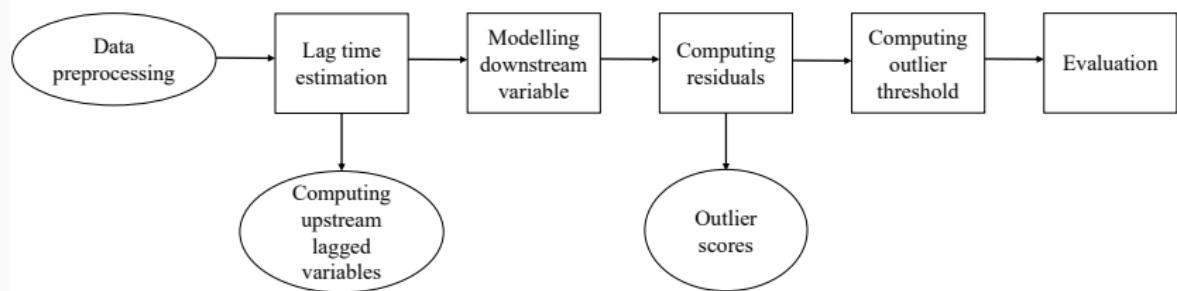
- Low-cost in-situ sensors
- Produce high-frequency data
- Prone to errors due to miscalibration, biofouling, battery and technical errors

Objective

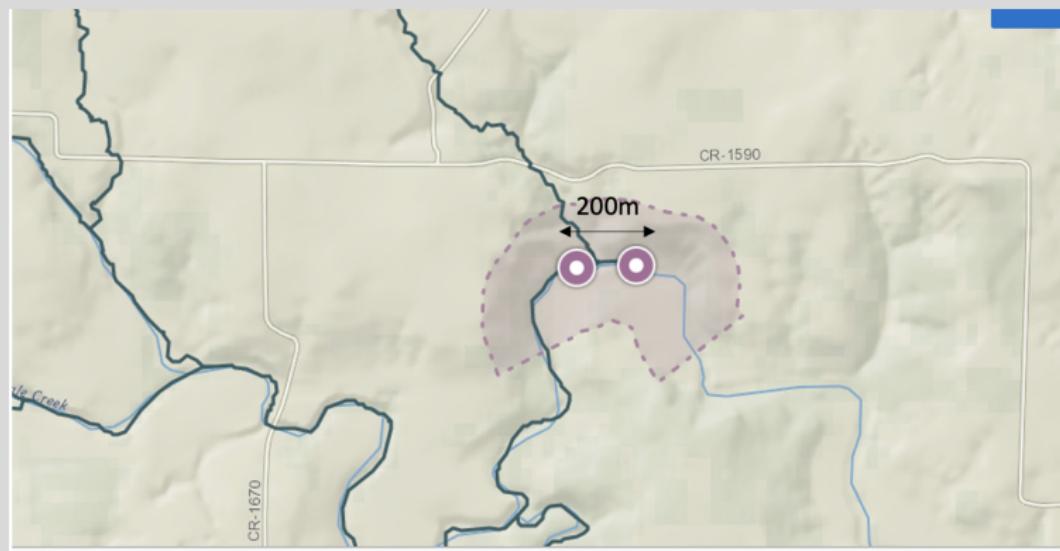
- Developing statistical tools to detect anomalies in water-quality variables measured by in-situ sensors

Proposed framework

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



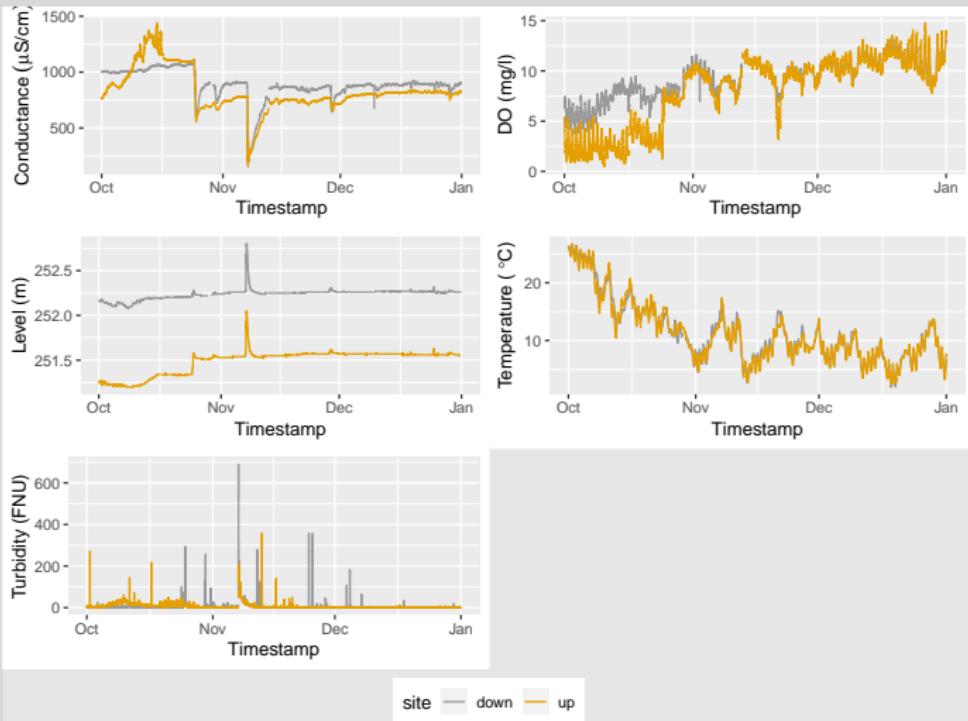
Pringle Creek - Texas, USA



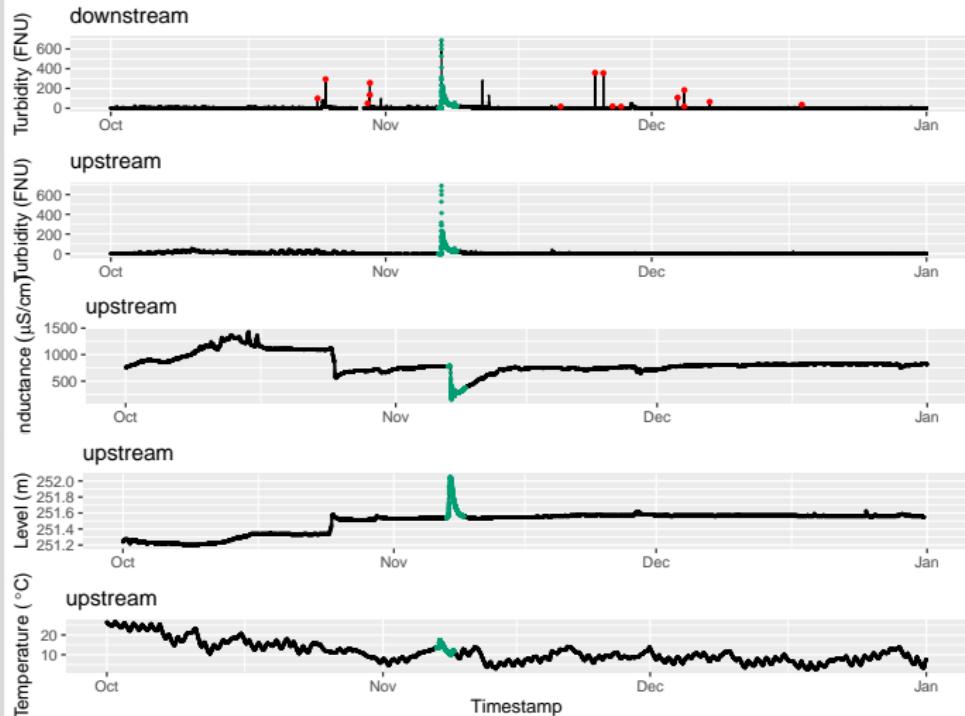
Data

- **Variables** - Turbidity, Conductivity, Dissolved oxygen, Level and Temperature
- **Time span** - 01-10-2019 to 31-12-2019
- **Frequency** - 5 minute intervals

Time plots



Turbidity downstream



Modeling turbidity with upstream variables

$$\log(\text{turbidity_downstream}_t) \sim \sum_{i=1}^p f_i(\log(\text{turbidity_downstream}_{t-i})) + s_1(\log(\text{turbidity_upstream}_{t-d_t})) + s_2(\text{conductance_upstream}_{t-d_t}) + s_3(\text{level_upstream}_{t-d_t}) + s_4(\text{temperature_upstream}_{t-d_t}) + \epsilon_t$$

- **Problem:** How to estimate d_t ?

Estimating lag time

- 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 $x_t^* y_{t+k}^* | \mathbf{z}_t \sim N(r_k(\mathbf{z}_t), \sigma_r^2)$,

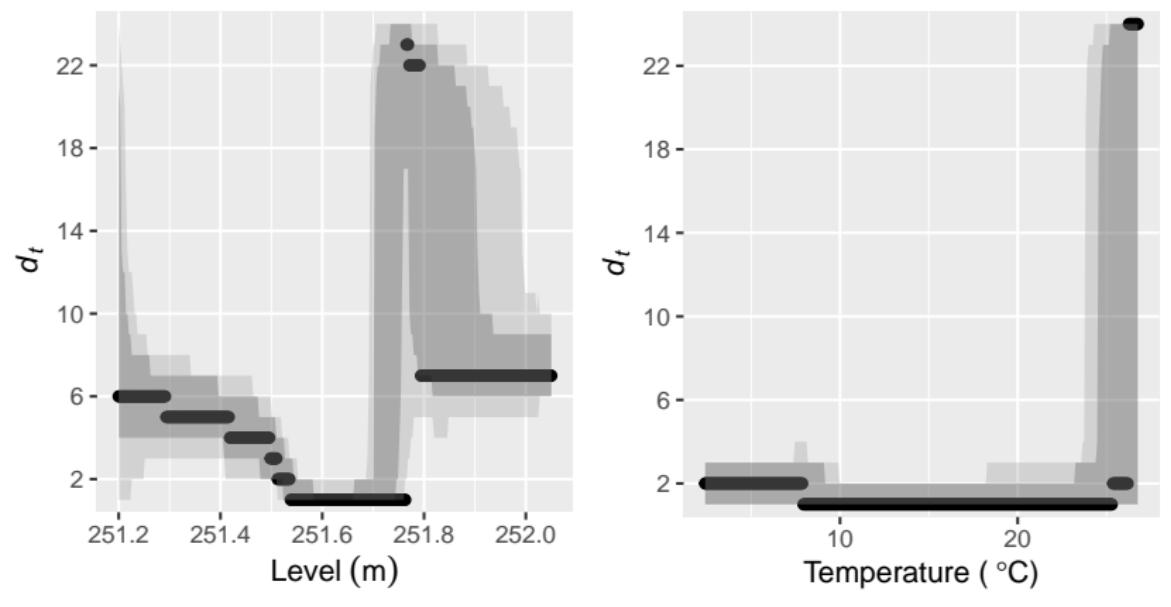
$$g(r_k(\mathbf{z}_t)) = \gamma_0 + \sum_{i=1}^p h_i(z_{i,t}) + \varepsilon_t$$

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

Estimating time delay

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

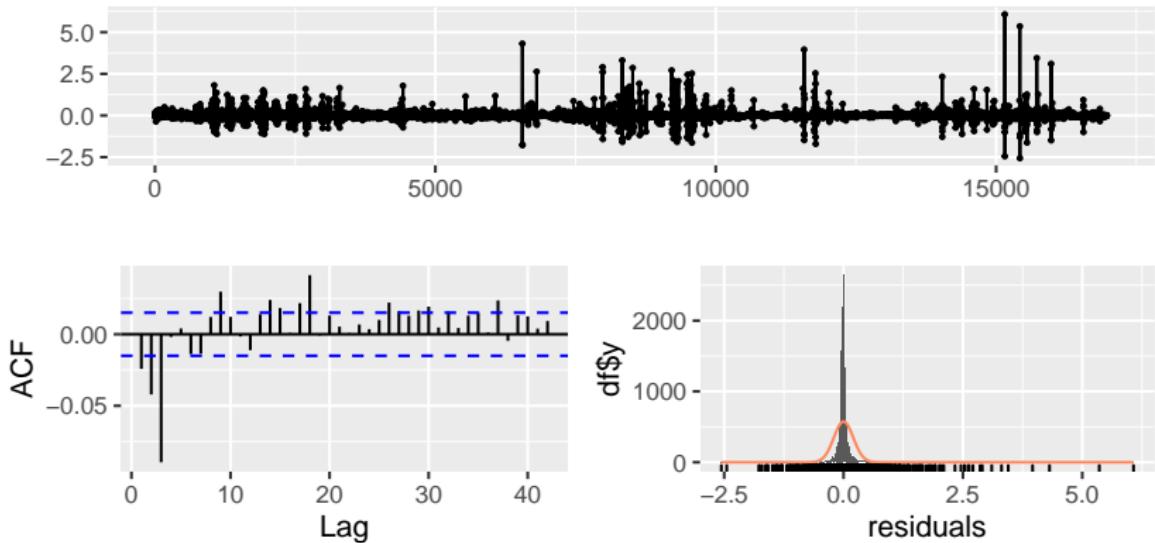
Visualising d_t vs predictors



Back to our GAM for turbidity downstream

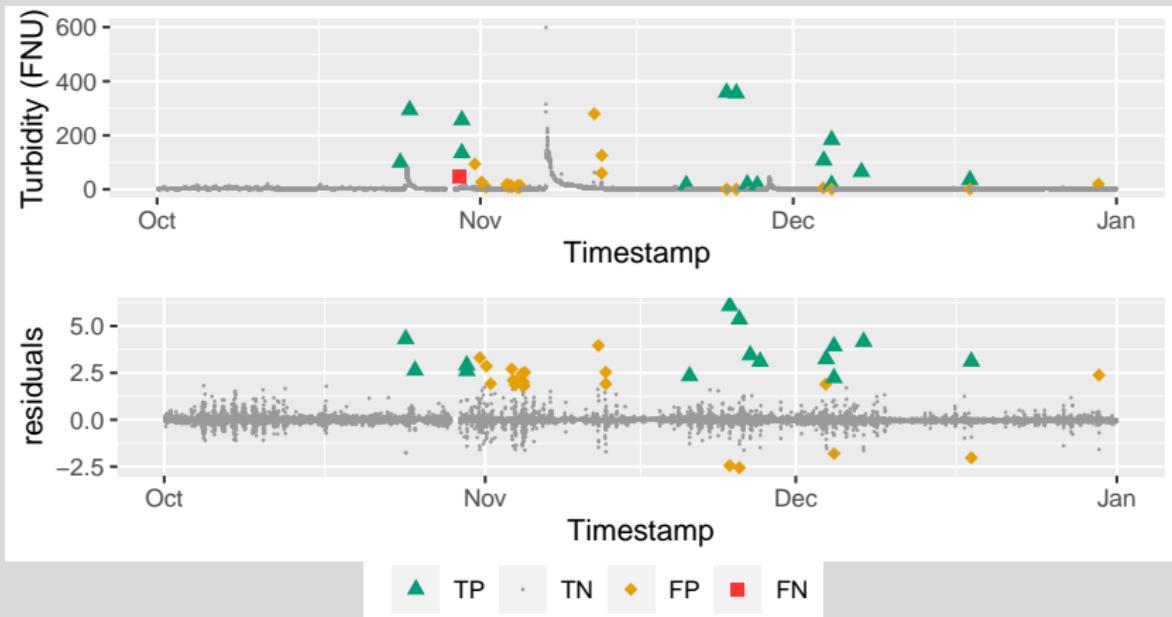
$$\log(\text{turbidity_downstream}_t) \sim \sum_{i=1}^3 f_i(\log(\text{turbidity_downstream}_{t-i})) + \\ s_1(\log(\text{turbidity_upstream}_{t-d_t})) + \\ s_2(\text{conductance_upstream}_{t-d_t}) + \\ s_3(\text{level_upstream}_{t-d_t}) + \\ s_4(\text{temperature_upstream}_{t-d_t}) + \epsilon_t$$

Residuals from the GAM



- We use Peak-Over-Threshold method to estimate the outlier threshold

Performance Evaluation



method	TP	TN	FP	FN
GAM-upstream-AR	14	25769	22	1

Comparison

GAM_up

$$\begin{aligned} \log(\text{turbidity_down}_t) \sim & s_1(\log(\text{turbidity_up}_{t-d_t})) \\ & + s_2(\text{conductance_up}_{t-d_t}) + s_3(\text{level_up}_{t-d_t}) \\ & + s_4(\text{temperature_up}_{t-d_t}) + s_5(\text{do_up}_{t-d_t}) + \epsilon_t \end{aligned}$$

GAM_down

$$\begin{aligned} \log(\text{turbidity_down}_t) \sim & g_1(\text{conductance_down}_{t-1}) \\ & + g_2(\text{level_down}_{t-1}) + g_3(\text{temperature_down}_{t-1}) \\ & + g_4(\text{do_down}_{t-1}) + \epsilon_t \end{aligned}$$

GAM_up_down

$$\begin{aligned} \log(\text{turbidity_down}_t) \sim & h_1(\log(\text{turbidity_up}_{t-d_t})) \\ & + h_2(\text{conductance_up}_{t-d_t}) + h_3(\text{do_up}_{t-d_t}) \\ & + h_4(\text{level_up}_{t-d_t}) + h_5(\text{temperature_up}_{t-d_t}) \\ & + h_6(\text{conductance_down}_{t-1}) + h_7(\text{temperature_down}_{t-1}) + \epsilon_t \end{aligned}$$

Recall models

- GAM-up-AR: GAM with upstream predictors and lagged response
- GAM-up: GAM with upstream predictors
- GAM-down: GAM with downstream predictors
- GAM-up-down: GAM with upstream and downstream predictors

Comparison

method	TP	TN	FP	FN	Accuracy	Error Rate	Optimised Precision
GAM-up-AR	14	25769	22	1	0.9991	0.0009	0.9651
GAM-up-down	13	25847	14	2	0.9994	0.0006	0.9282
GAM-up	11	25851	10	4	0.9995	0.0005	0.8458
GAM-down	5	26011	0	10	0.9996	0.0004	0.4996

Working papers and R packages

Working papers

- Conditional Normalisation in Time Series Analysis
- Anomaly Detection in River Networks

R packages

- **conduits** (Conditional UI for Time Series normalisation) - <https://github.com/PuwasaLaG/conduits>
- **dori** (Data for Outlier Detection in River networks)

Acknowledgement

- ARC Linkage project - “Revolutionising high resolution water-quality monitoring in the information age”.
- Department of Environment and Science, Queensland

