MONASH University

MONASH BUSINESS SCHOOL

A framework for anomaly detection in river networks

Puwasala Gamakumara

ACEMS Retreat 2021

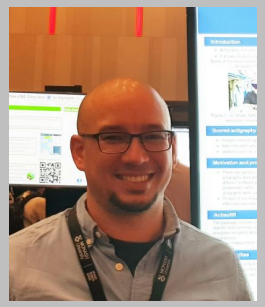
Rob Hyndman

Catherine Leigh

Dilini Talagala

Kerrie Mengersen

Erin Peterson



Edgar Santos-Fernandez

Katie Buchhorn

Water-quality monitoring in river networks

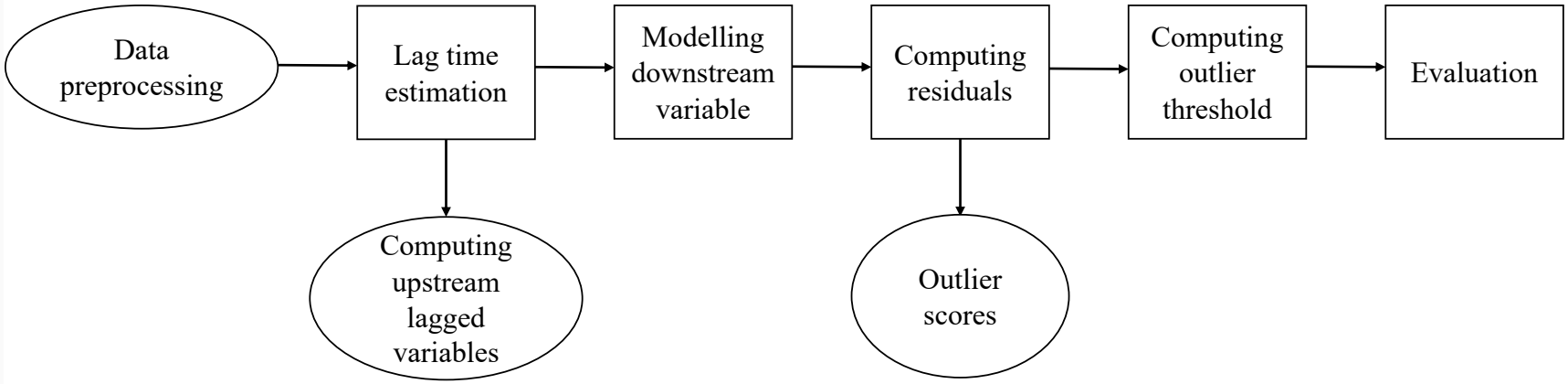
- Low-cost in-situ sensors
- Produce high-frequency data
- Prone to errors due to miscalibration, biofouling, battery and technical errors

Objective

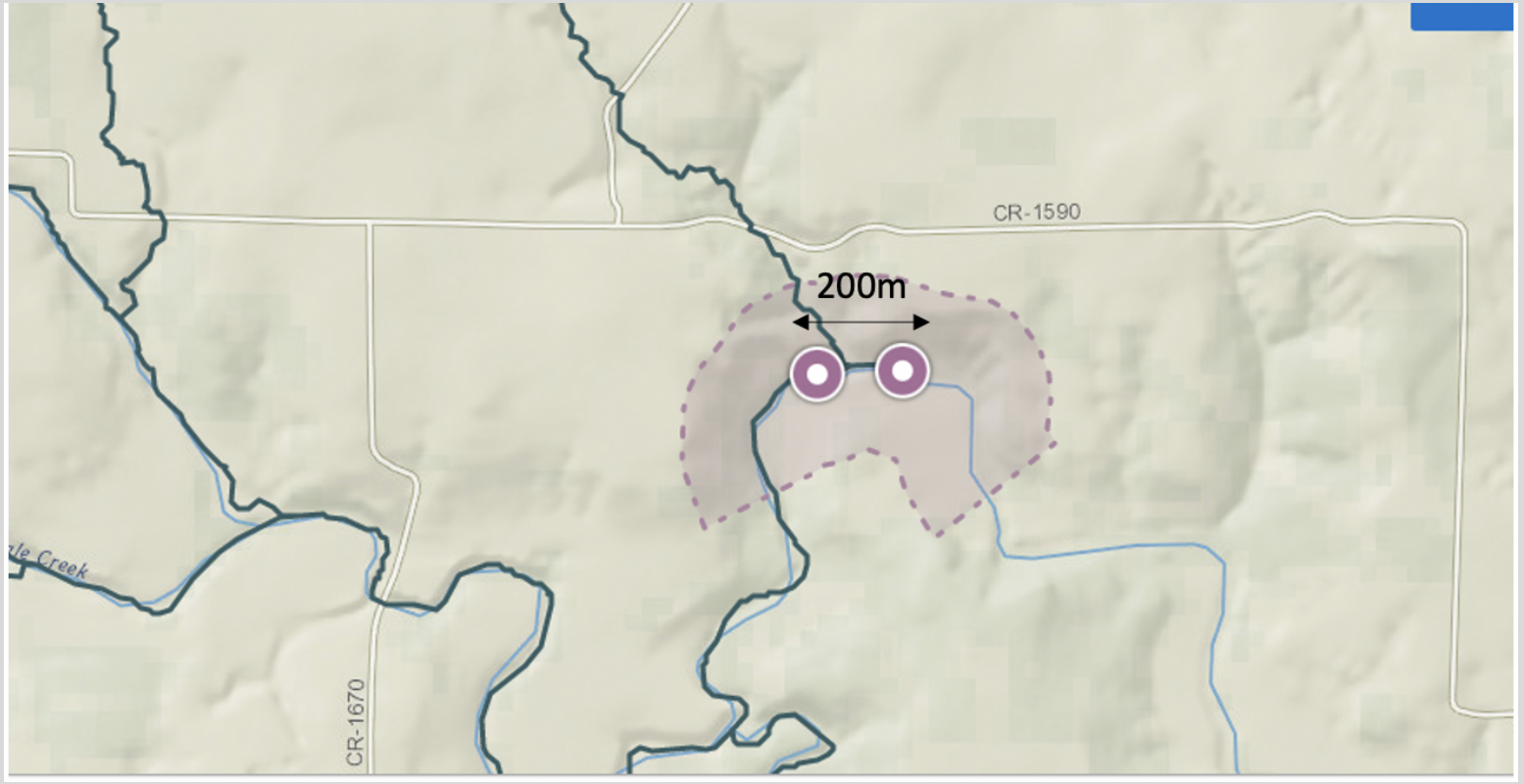
- Developing statistical tools to detect anomalies in water-quality variables measured by in-situ sensors

Proposed framework

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



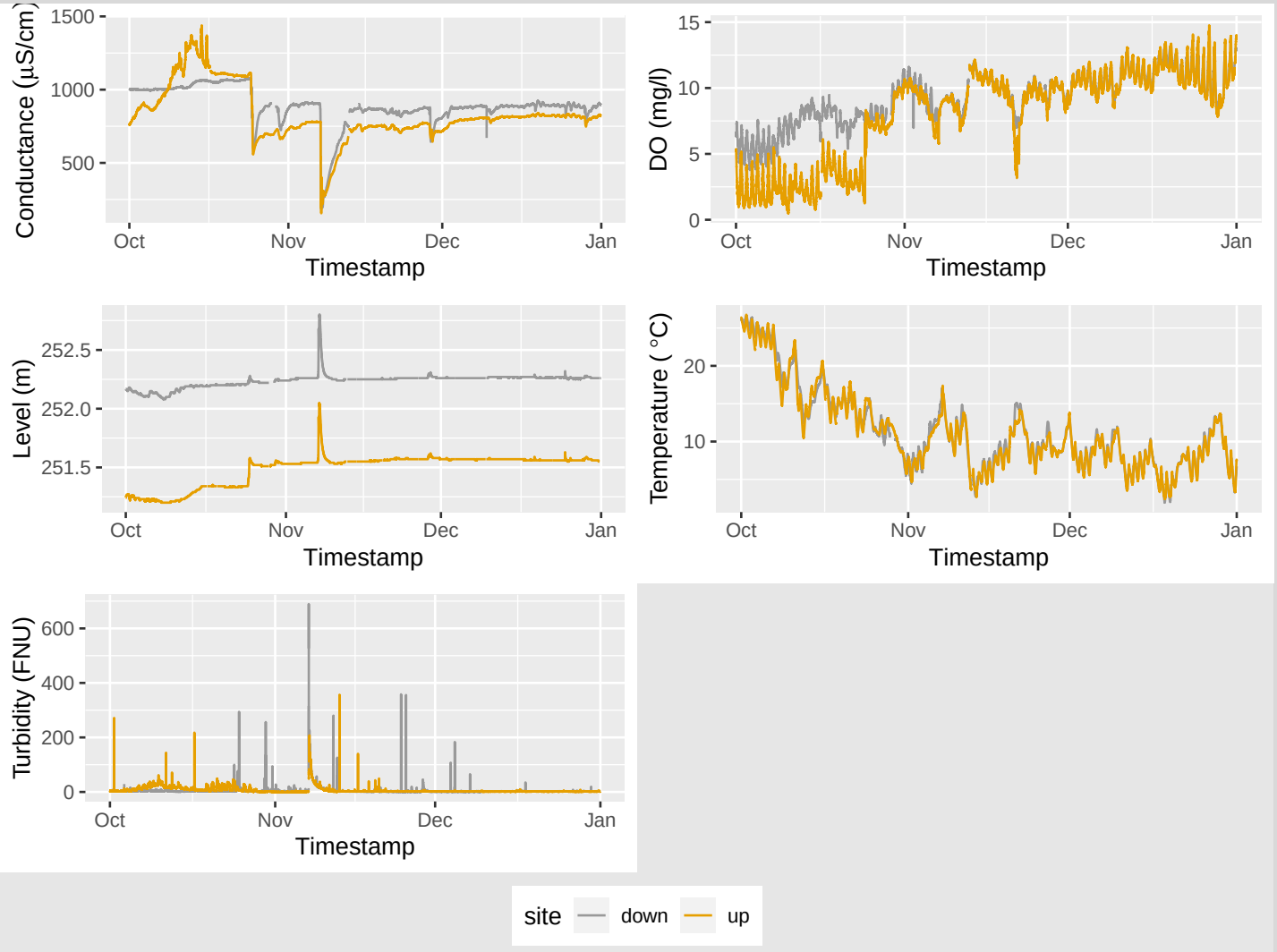
Pringle Creek - Texas, USA



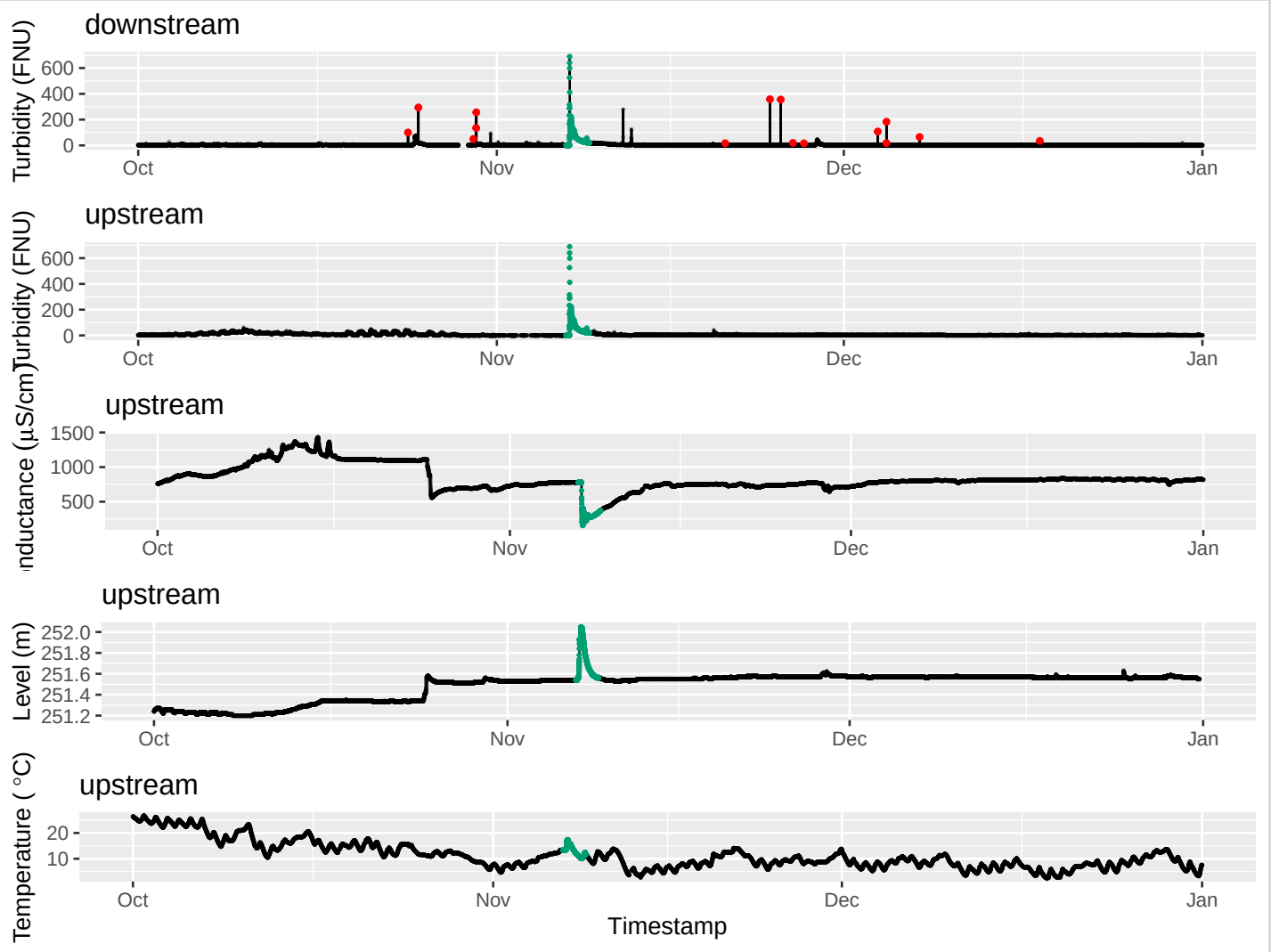
Data

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

Time plots


Conductance (µS/cm)
1500
1000
500
Oct
Nov
Dec
Jan
Timestamp
DO (mg/l)
15
10
5
0
Oct
Nov
Dec
Jan
Timestamp
Level (m)
252.5
252.0
251.5
Oct
Nov
Dec
Jan
Timestamp
Temperature (°C)
20
10
Oct
Nov
Dec
Jan
Timestamp
Turbidity (FNU)
600
400
200
0
Oct
Nov
Dec
Jan
Timestamp
site
down
up

Turbidity downstream



Modeling turbidity with upstream variables

$$log(turbidity\_downstream_t) \sim \sum_{i=1}^{p} f_i(log(turbidity\_downstream_{t-i})) + s_1(log(turbidity\_upstream_{t-d_t})) + s_2(conductance\_upstream_{t-d_t}) + s_3(level\_upstream_{t-d_t}) + s_4(temperature\_upstream_{t-d_t}) + \epsilon_t$$

- **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 $\boldsymbol{z}_t$: {level upstream, temperature upstream}
- $x_t^* = \frac{x_t - \mathrm{E}[x_t|\boldsymbol{z}_t]}{\sqrt{\mathrm{V}[x_t|\boldsymbol{z}_t]}}$ and $y_t^* = \frac{y_t - \mathrm{E}[y_t|\boldsymbol{z}_t]}{\sqrt{\mathrm{V}[y_t|\boldsymbol{z}_t]}}$

Conditional cross-correlation

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

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

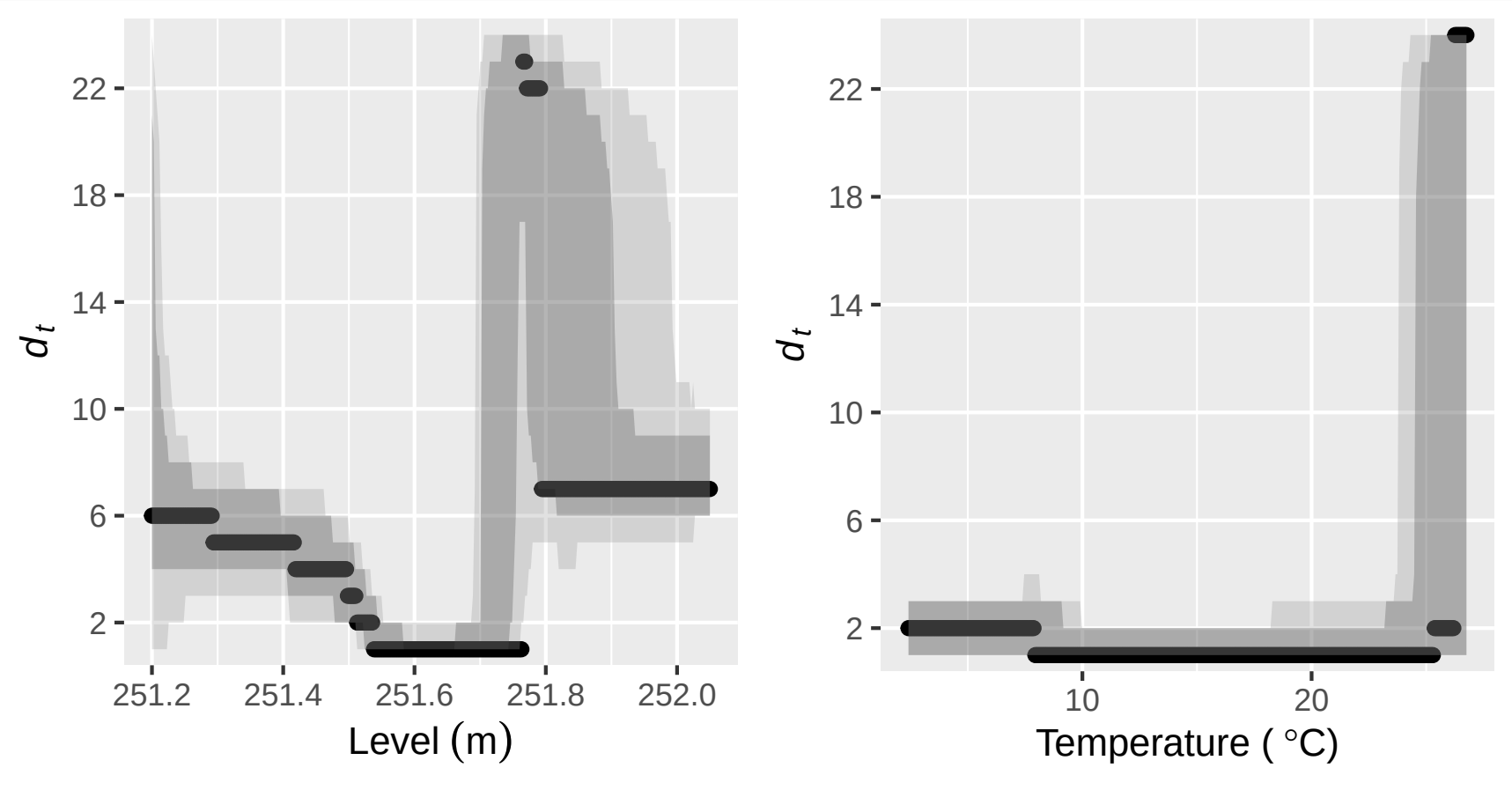
$$g(r_k(\boldsymbol{z}_t)) = \gamma_0 + \sum_{i=1}^{p} h_i(z_{i,t}) + \varepsilon_t$$

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

Estimating time delay

$$\hat{d}_t(\boldsymbol{z}_t) = \underset{k}{\operatorname{argmax}} \quad \hat{r}_k(\boldsymbol{z}_t)$$

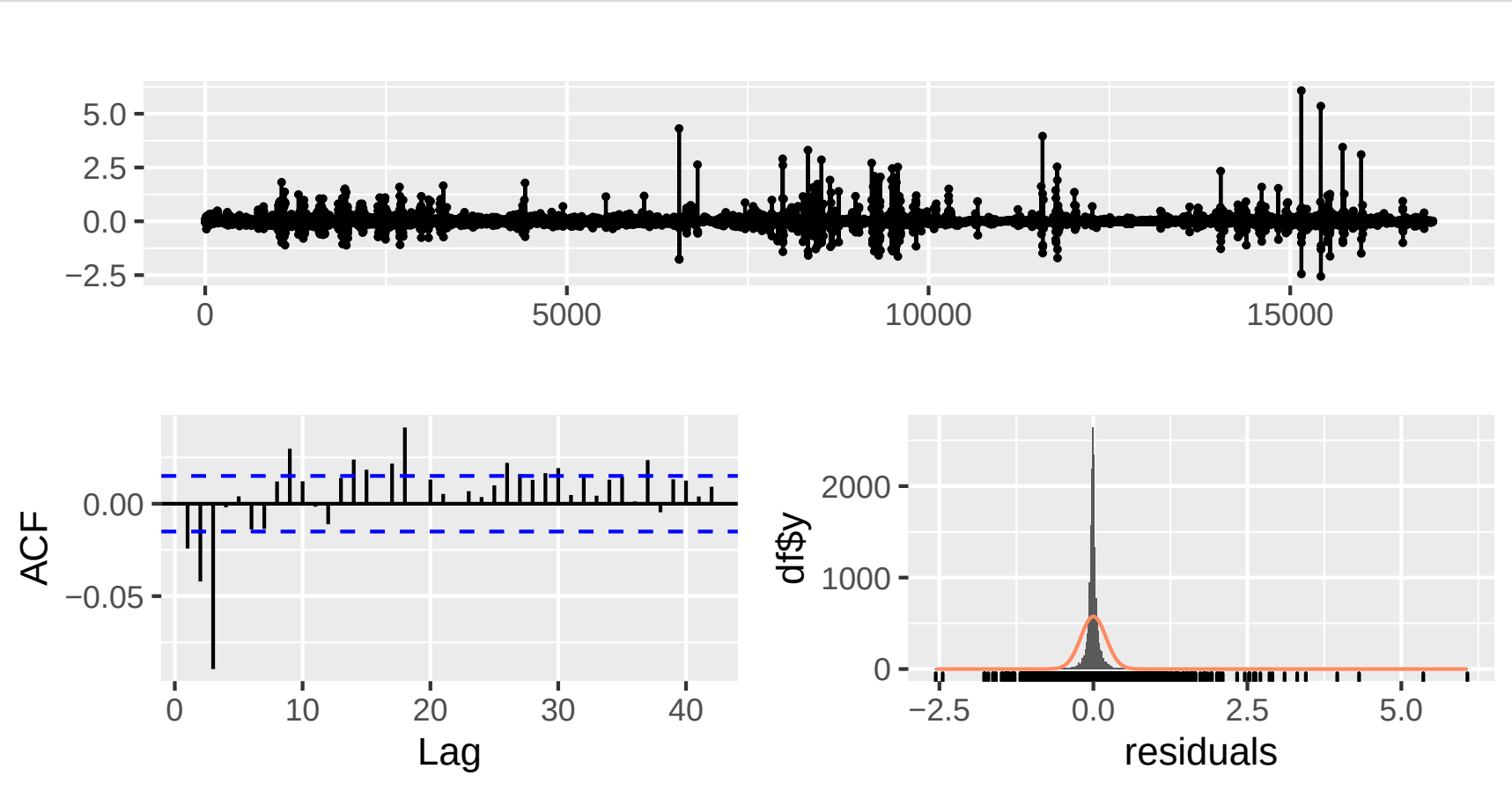
Visualising dt vs predictors



Back to our GAM for turbidity downstream

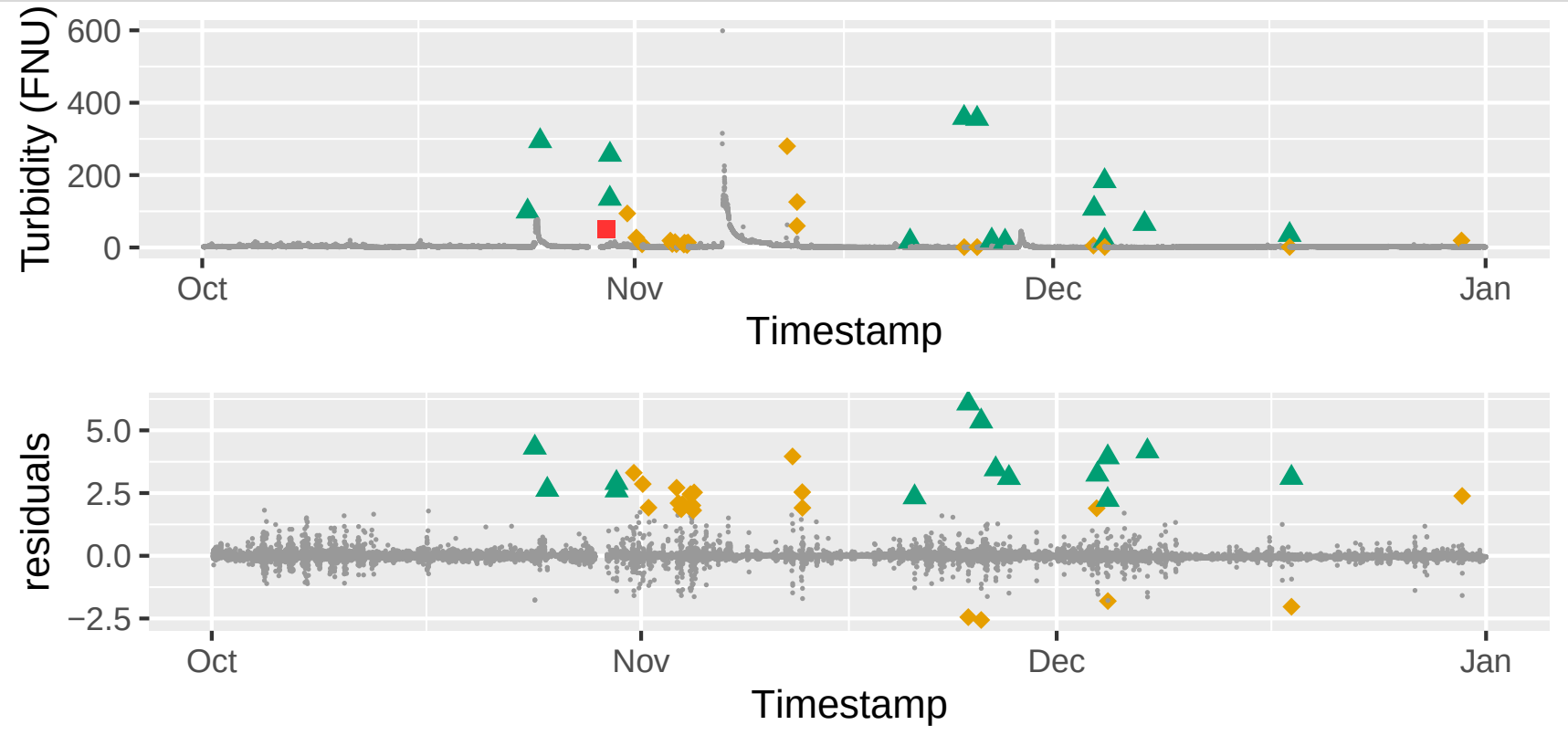
$$log(turbidity\_downstream_t) \sim \sum_{i=1}^{3} f_i(log(turbidity\_downstream_{t-i}))+$$
$$s_1(log(turbidity\_upstream_{t-d_t}))+$$
$$s_2(conductance\_upstream_{t-d_t})+$$
$$s_3(level\_upstream_{t-d_t})+$$
$$s_4(temperature\_upstream_{t-d_t}) + \epsilon_t$$

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

$$log(turbidity\_down_t) \sim s_1(log(turbidity\_up_{t-d_t})) + s_2(conductance\_up_{t-d_t}) + s_3(level\_up_{t-d_t}) + s_4(temperature\_up_{t-d_t}) + s_5(do\_up_{t-d_t}) + \epsilon_t$$

GAM_down

$$log(turbidity\_down_t) \sim g_1(conductance\_down_{t-1}) + g_2(level\_down_{t-1}) + g_3(temperature\_down_{t-1}) + g_4(do\_down_{t-1}) + \epsilon_t$$

GAM_up_down

$$log(turbidity\_down_t) \sim h_1(log(turbidity\_up_{t-d_t})) + h_2(conductance\_up_{t-d_t}) + h_3(do\_up_{t-d_t}) + h_4(level\_up_{t-d_t}) + h_5(temperature\_up_{t-d_t}) + h_6(conductance\_down_{t-1}) + h_7(temperature\_down_{t-1}) + \epsilon_t$$

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