MONASH University

MONASH BUSINESS SCHOOL

A framework for detecting anomalies in water-quality variables via conditional density estimation

Puwasala Gamakumara

November 18, 2021

Outline

Outline

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

Why multiple sensors?

Outlier detection

- Single sensor may not be able to detect certain type of anomalies
- If the sensors are malfunctioning, then the covariates or lagged values of them may be incorrect

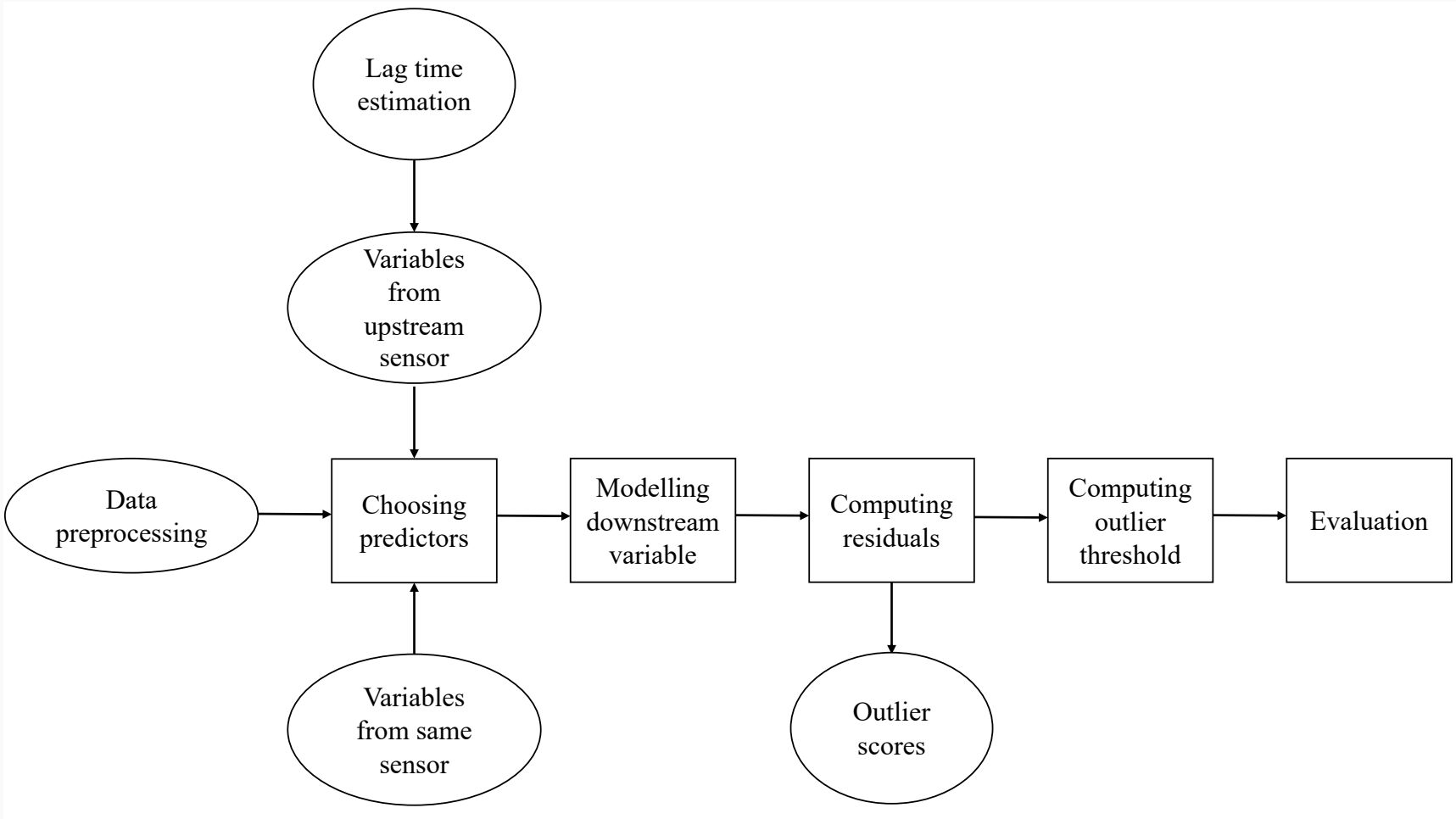
Importance of replication

- Uncertain measurements or unreliable equipment
- It’s important to have replications to measure uncertainty and bias

Outline

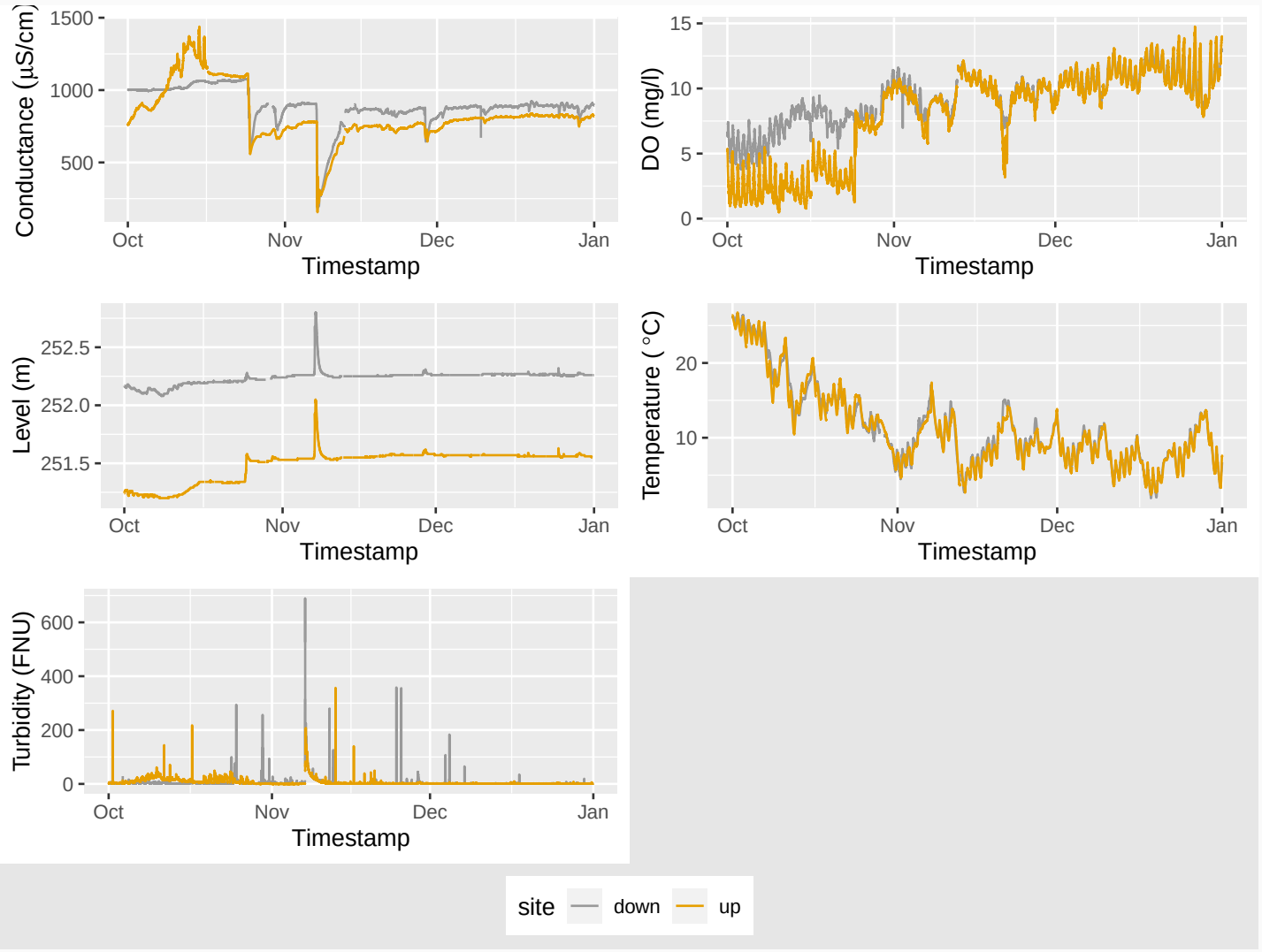
Anomaly detection framework

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

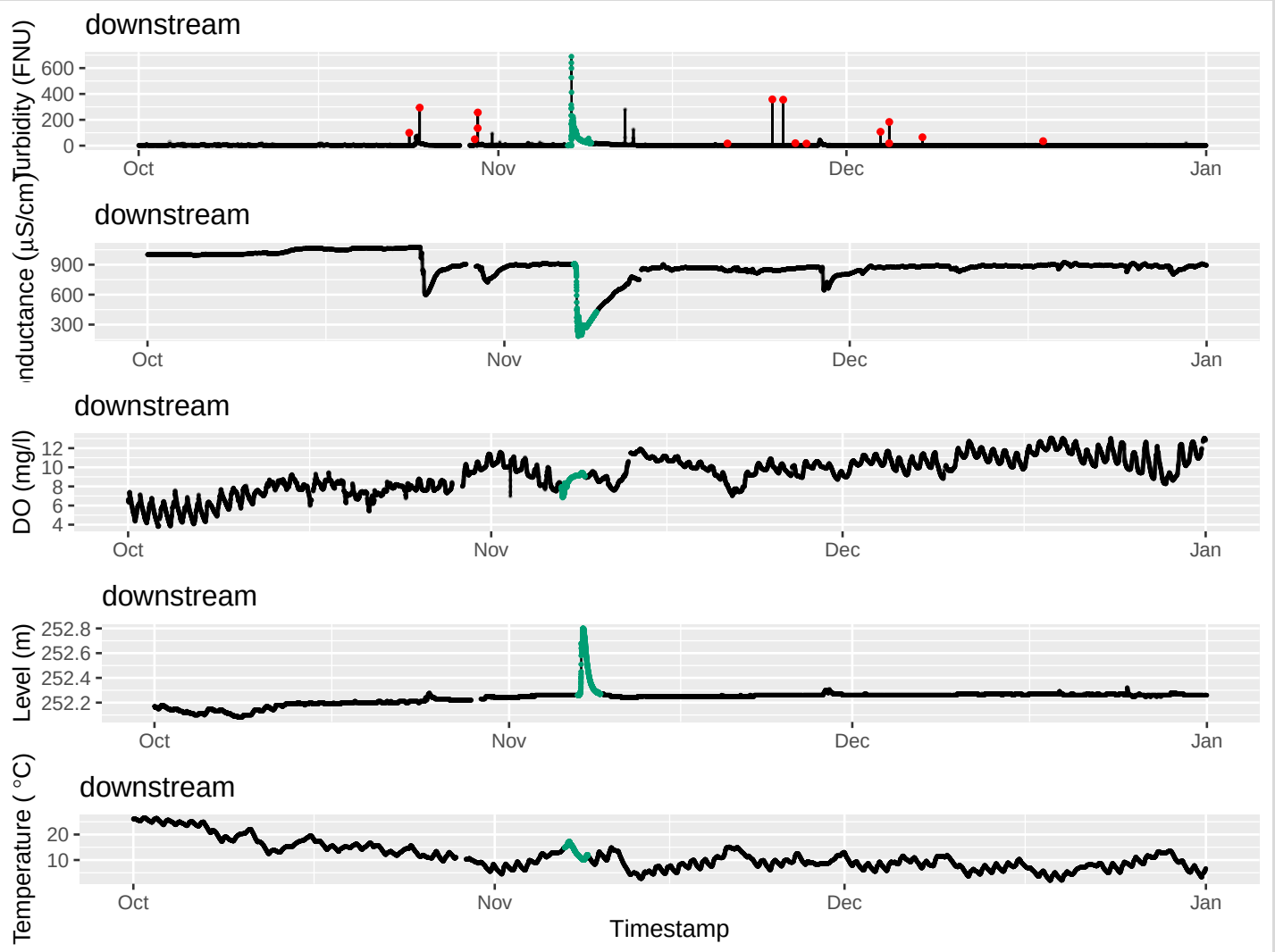


Outline

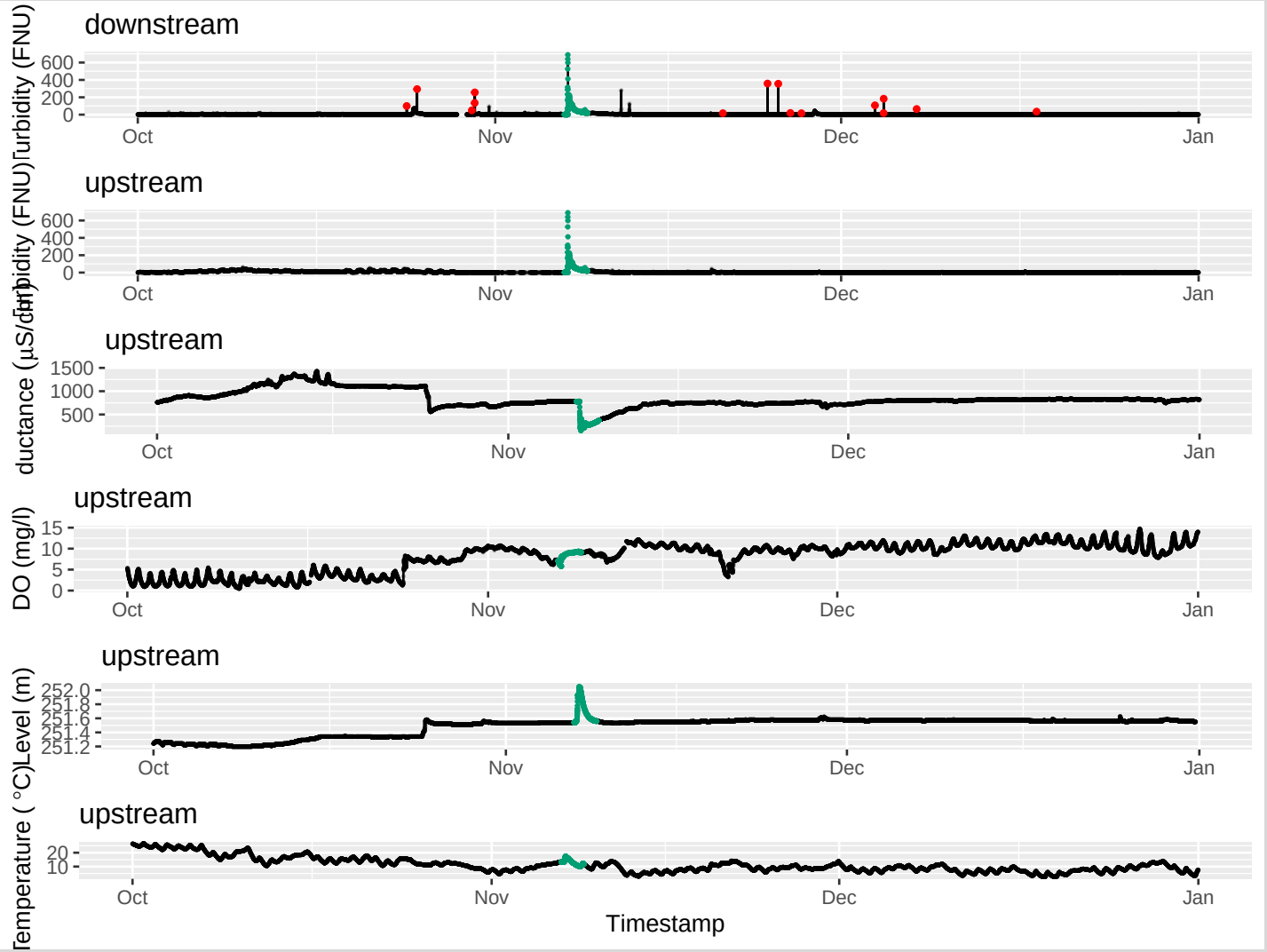
Time plots


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 vs other downstream variables



Turbidity downstream vs other upstream variables



Outline

Modeling downstream turbidity

$$y_t = \phi_0 + \sum_{i=1}^{p} g_i(z_{i,t_l}) + \sum_{j=1}^{q} h_j(y_{t-j}) + r_t$$

Choices for z_{i,t_l}

- **GAM-down:** Downstream contemporaneous variables
- **GAM-down-AR:** Downstream contemporaneous variables + AR terms
- **GAM-up:** Upstream lagged variables
- **GAM-up-AR:** Upstream lagged variables + AR terms
- **GAM-up-down:** Upstream lagged variables + Downstream contemporaneous variables

Outline

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 $\boldsymbol{z}_t$: {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

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

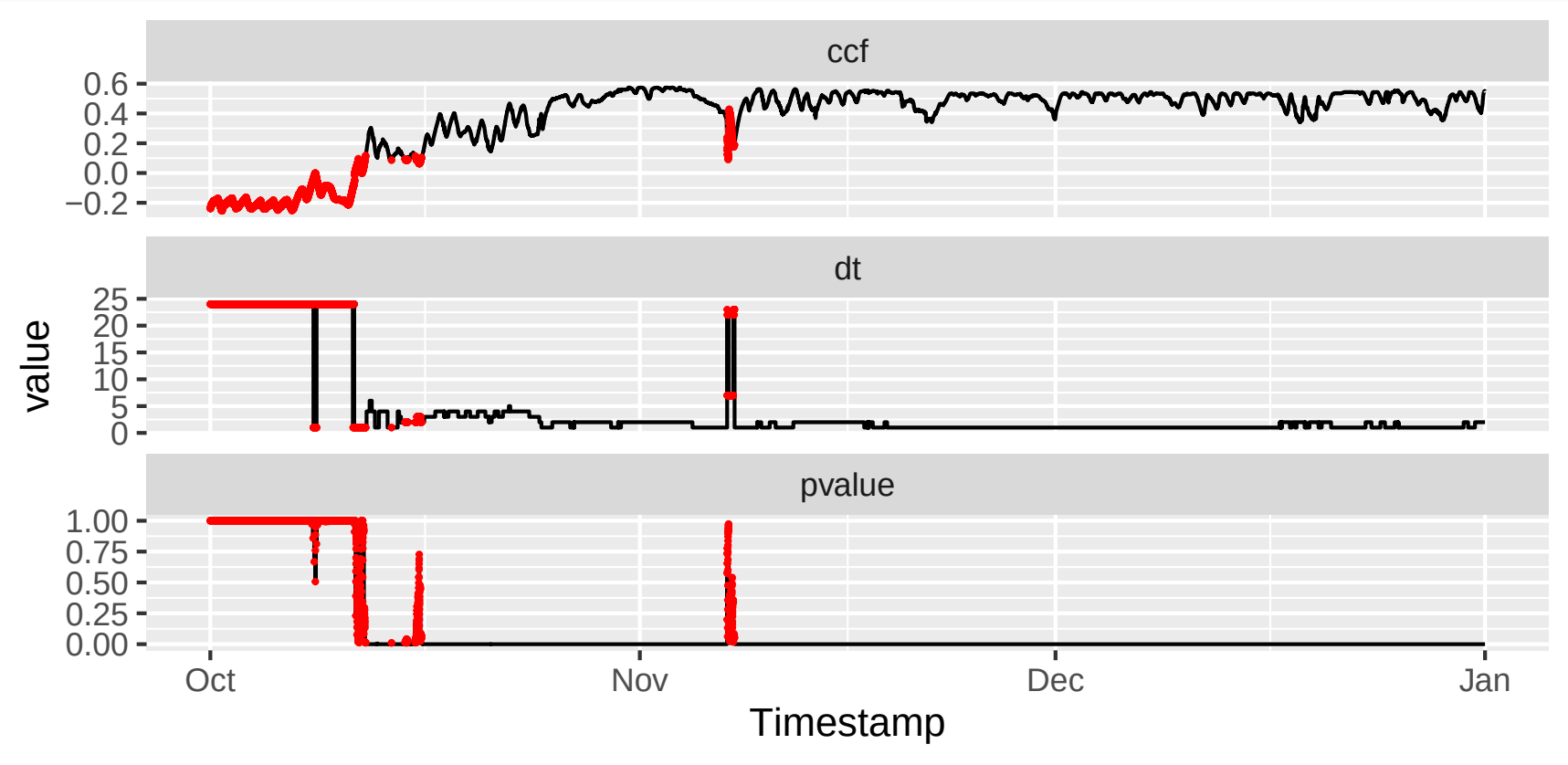
- To estimate $r_k(\boldsymbol{z}_t)$ we fit the following GAMs
- Let $x_t^* y_{t+k}^* | \boldsymbol{z}_t \sim \mathrm{N}(r_k(\boldsymbol{z}_t), \sigma_r^2)$,

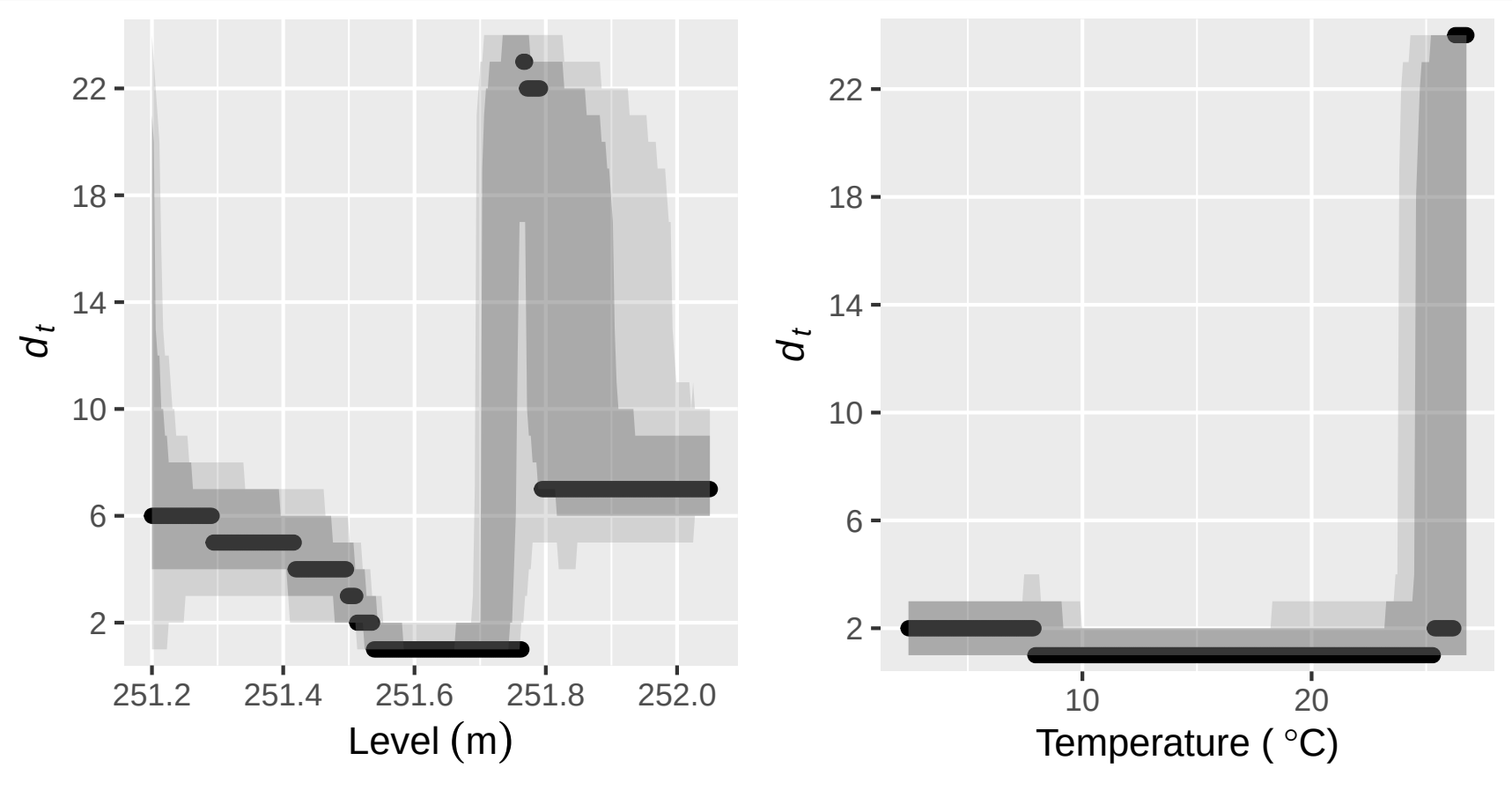
$$g(r_k(\boldsymbol{z}_t)) = \gamma_0 + \sum_{i=1}^{p} h_i(z_{i,t}) + \varepsilon_t$$

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

Estimating time delay

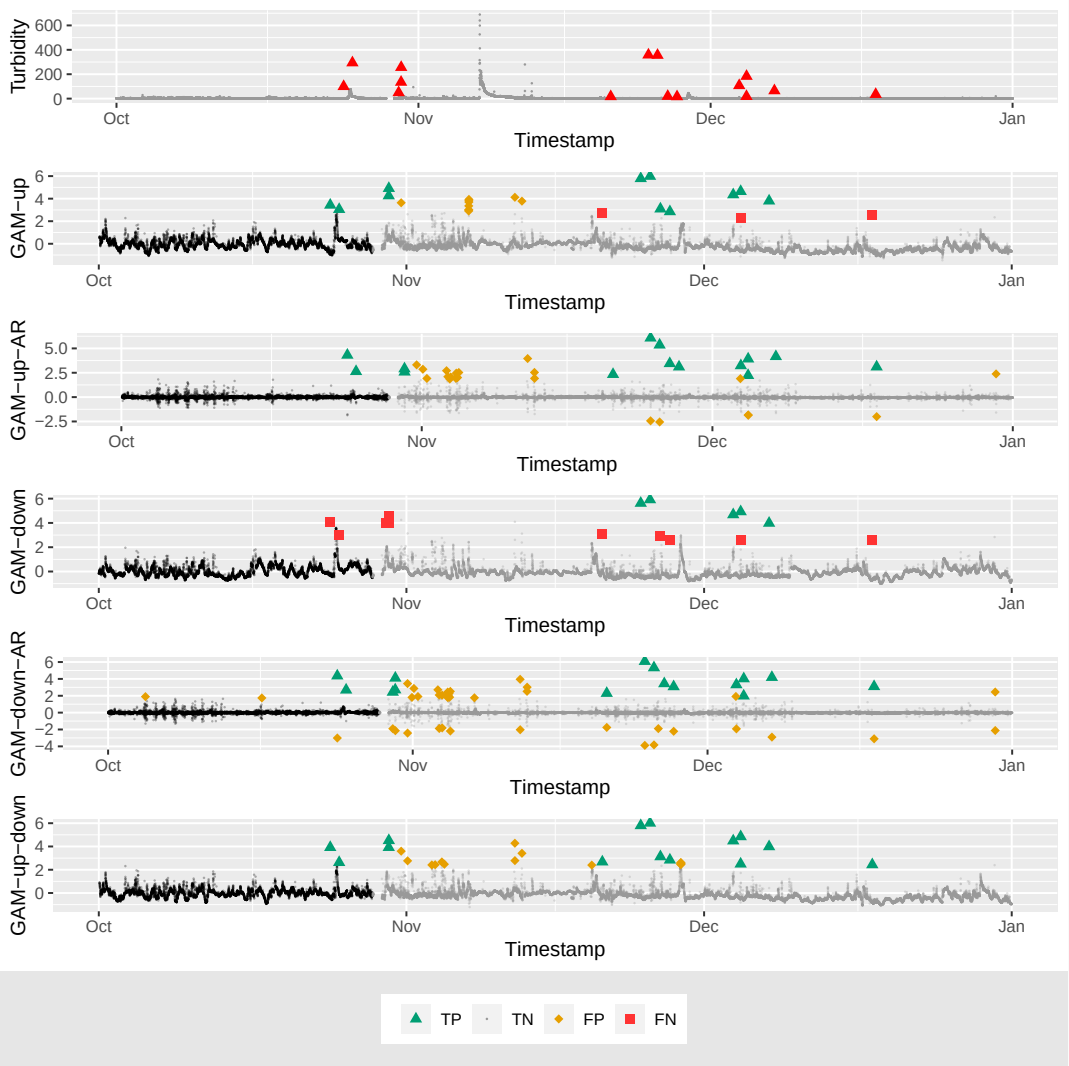
$$\hat{d}_t(\boldsymbol{z}_t) = \underset{k}{\operatorname{argmax}} \quad \hat{r}_k(\boldsymbol{z}_t)$$


ccf
dt
pvalue
value
0.6
0.4
0.2
0.0
−0.2
25
20
15
10
5
0
1.00
0.75
0.50
0.25
0.00
Oct
Nov
Dec
Jan
Timestamp


22
18
14
10
6
2
d_t
251.2
251.4
251.6
251.8
252.0
Level (m)
22
18
14
10
6
2
d_t
10
20
Temperature (°C)

Outline

Outlier detection using Peak over Threshold method



Comparison

method	TP	TN	FP	FN	OP
GAM-down-AR	15	25948	39	0	0.9977
GAM-up-down	14	25841	16	1	0.9652
GAM-up-AR	14	25765	22	1	0.9651
GAM-up	11	25847	10	4	0.8458
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