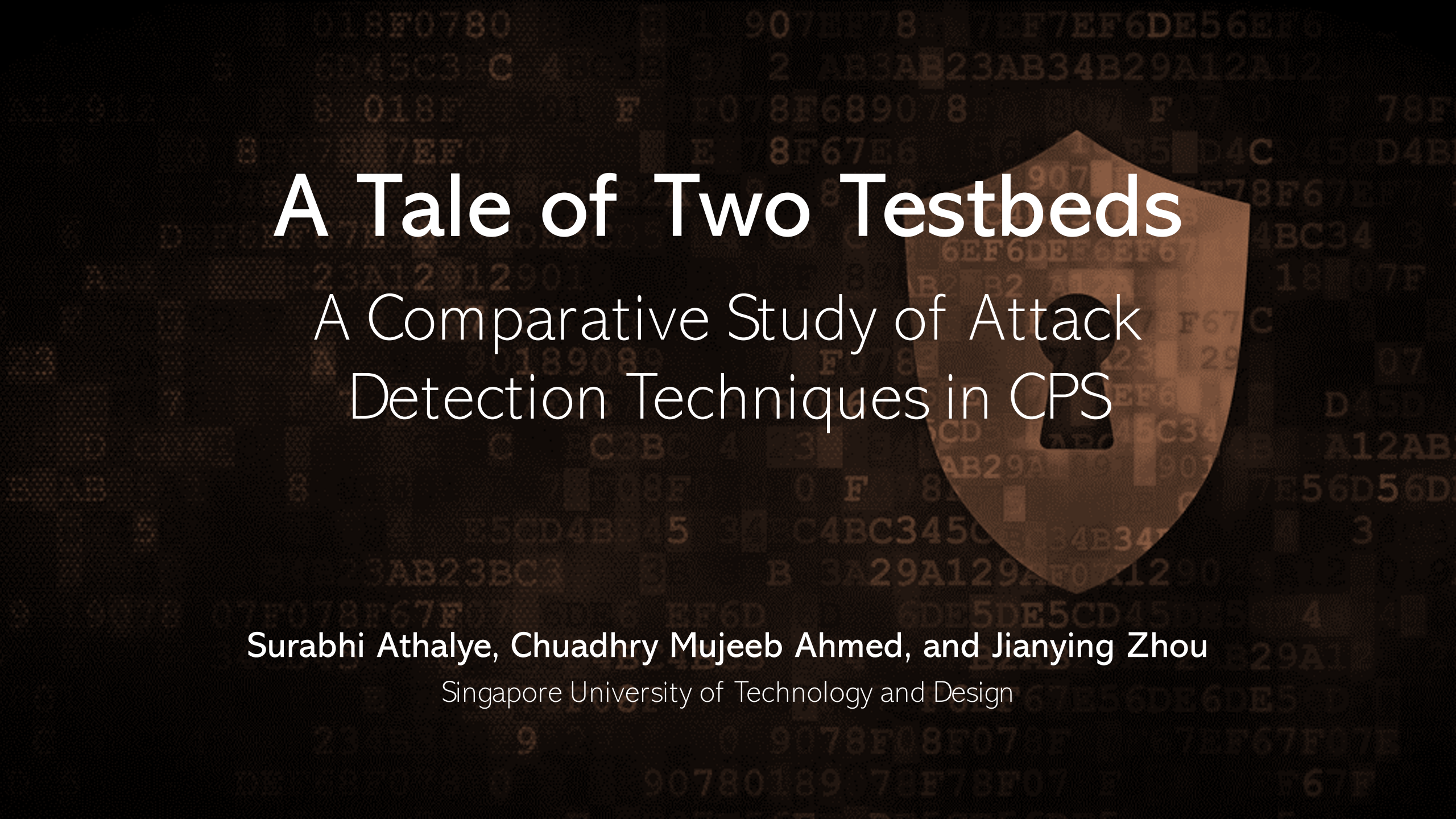
A Tale of Two Testbeds

A Comparative Study of Attack Detection Techniques in CPS



Surabhi Athalye, Chuadhry Mujeeb Ahmed, and Jianying Zhou

Singapore University of Technology and Design

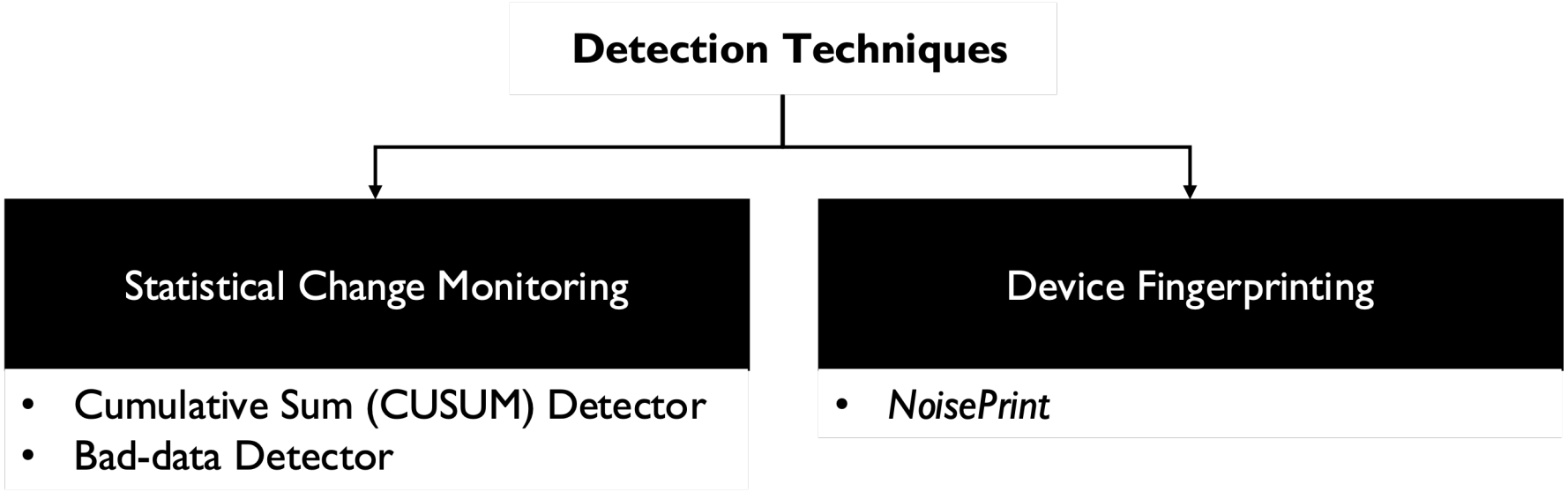
Contents

Introduction: Cyber-physical Systems

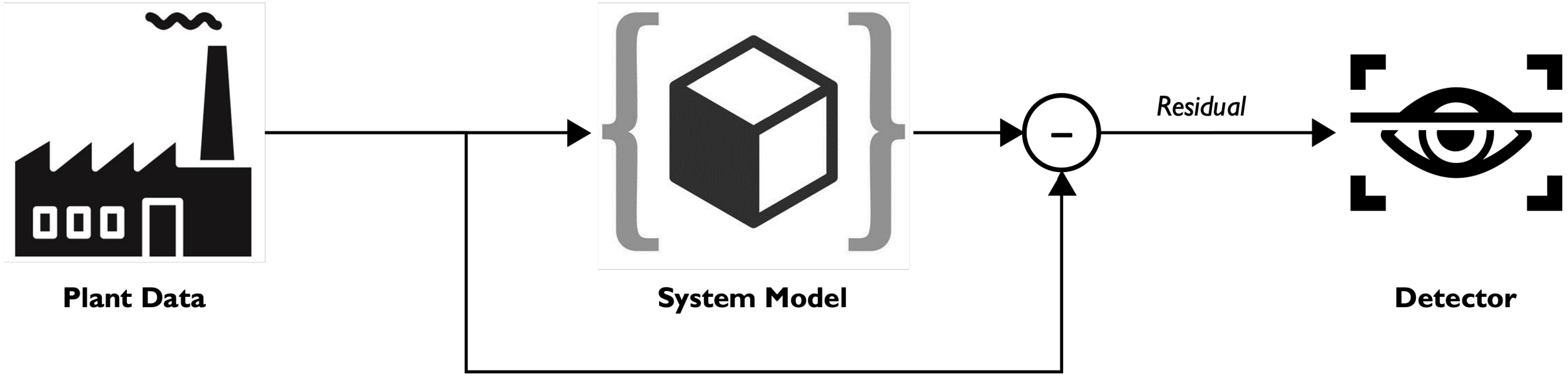
- Interconnected components:
  - Programmable Logic Controllers (PLCs)
  - Sensors, actuators
  - Supervisory Control and Data Acquisition (SCADA) workstation
  - Human Machine Interface (HMI)
  - Communication network
- Exposure to malicious entities.

Motivation

To exhaustively test and compare attack detection techniques for CPS on different testbeds.

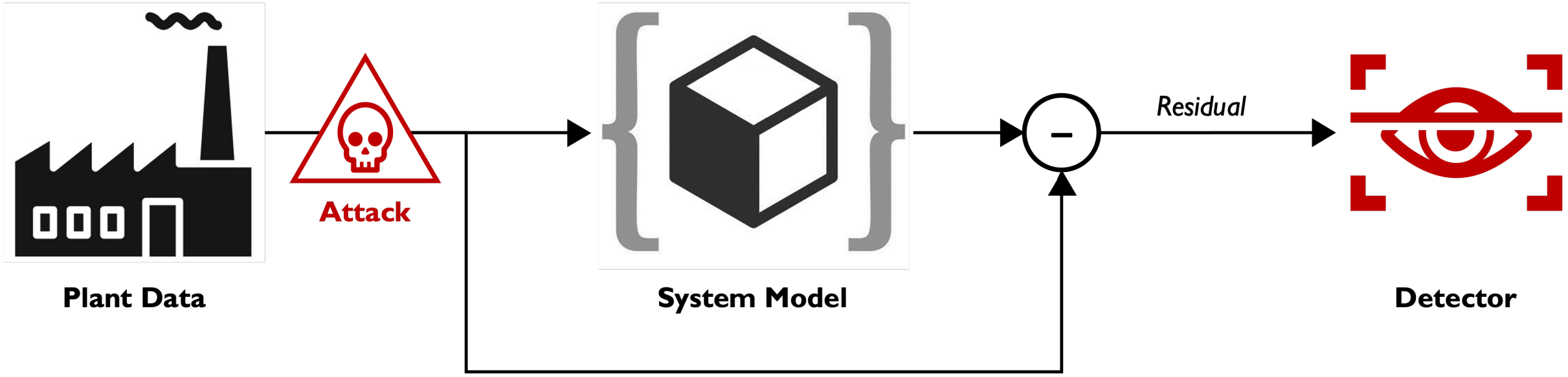


Methodology



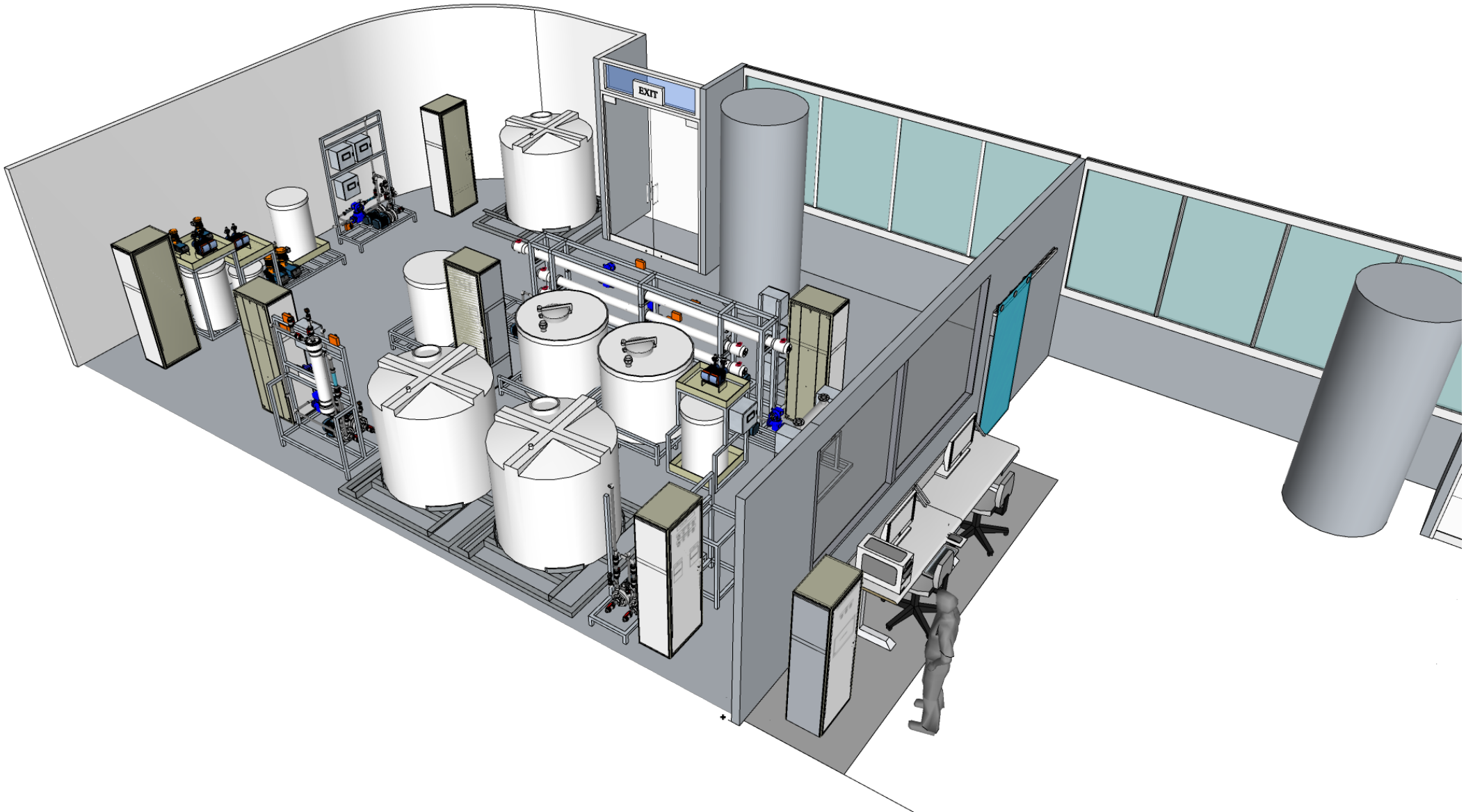
Model-based approach: normal operation

Methodology

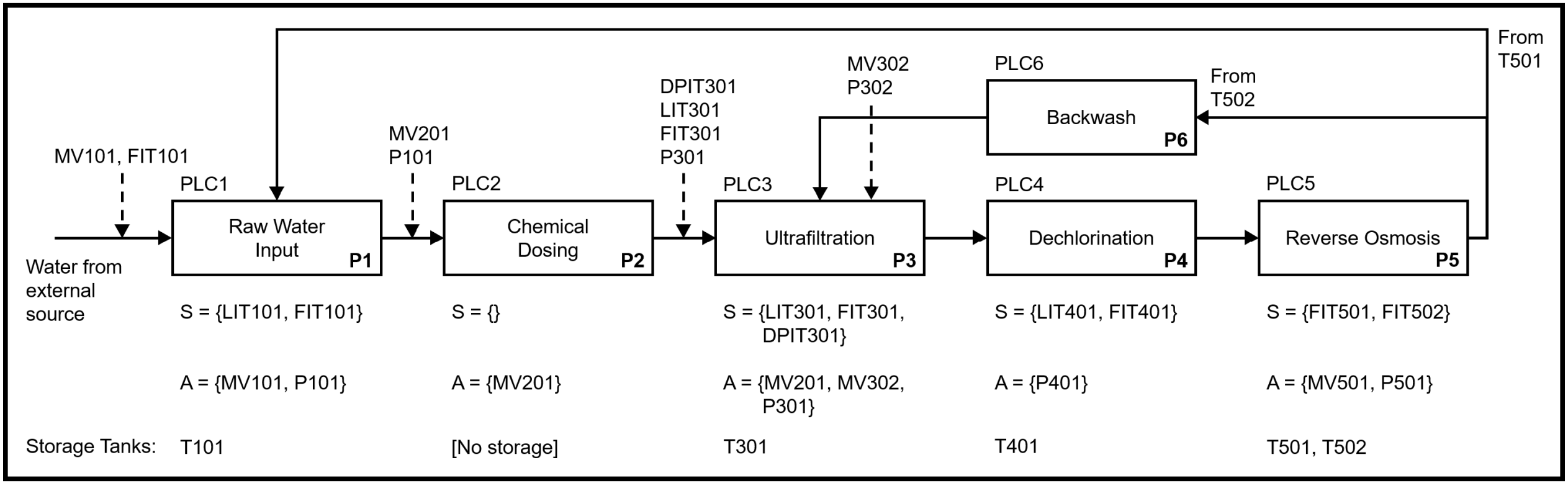


Model-based approach: under attack

Testbeds: SWaT

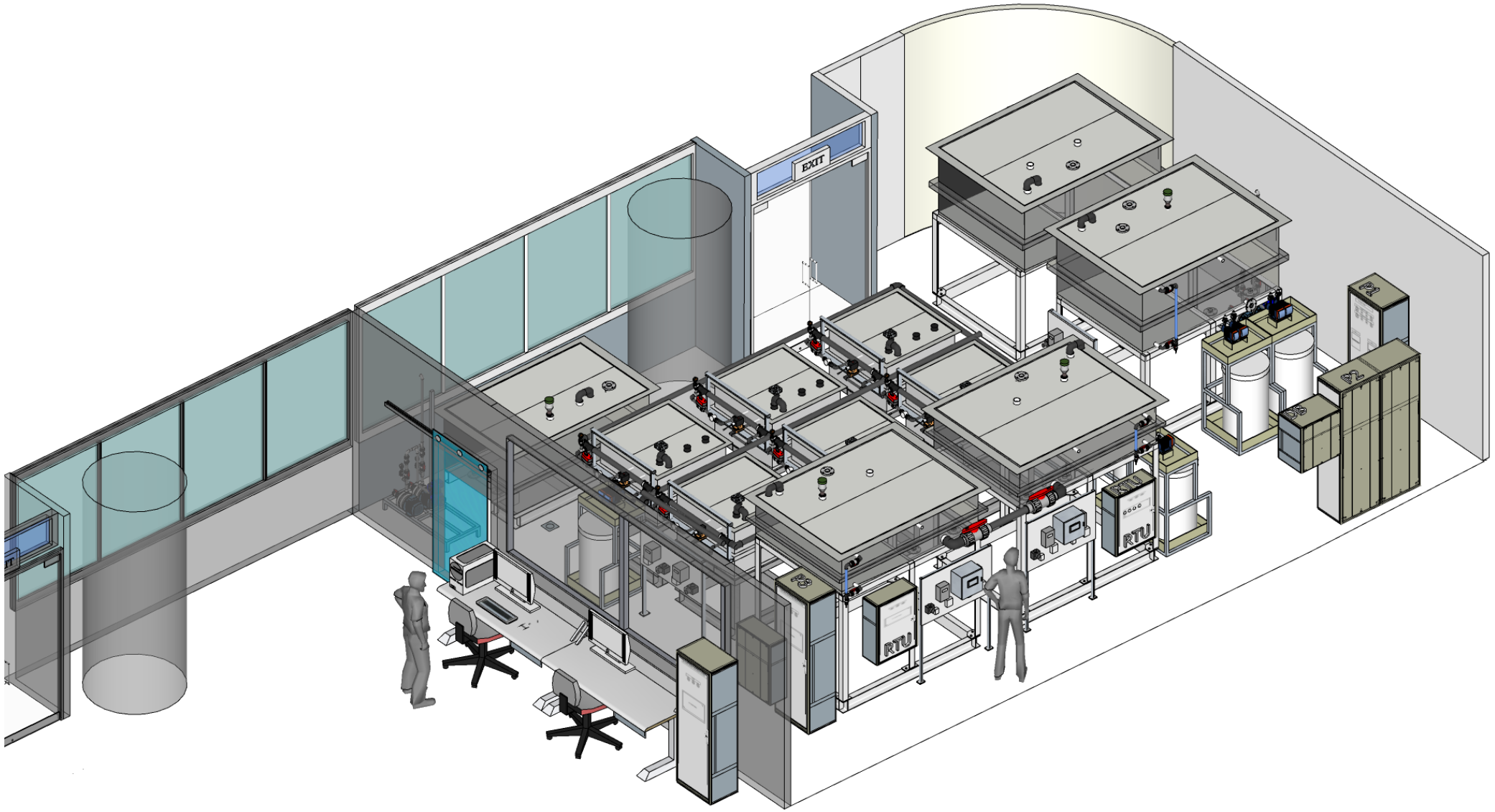


Testbeds: SWaT

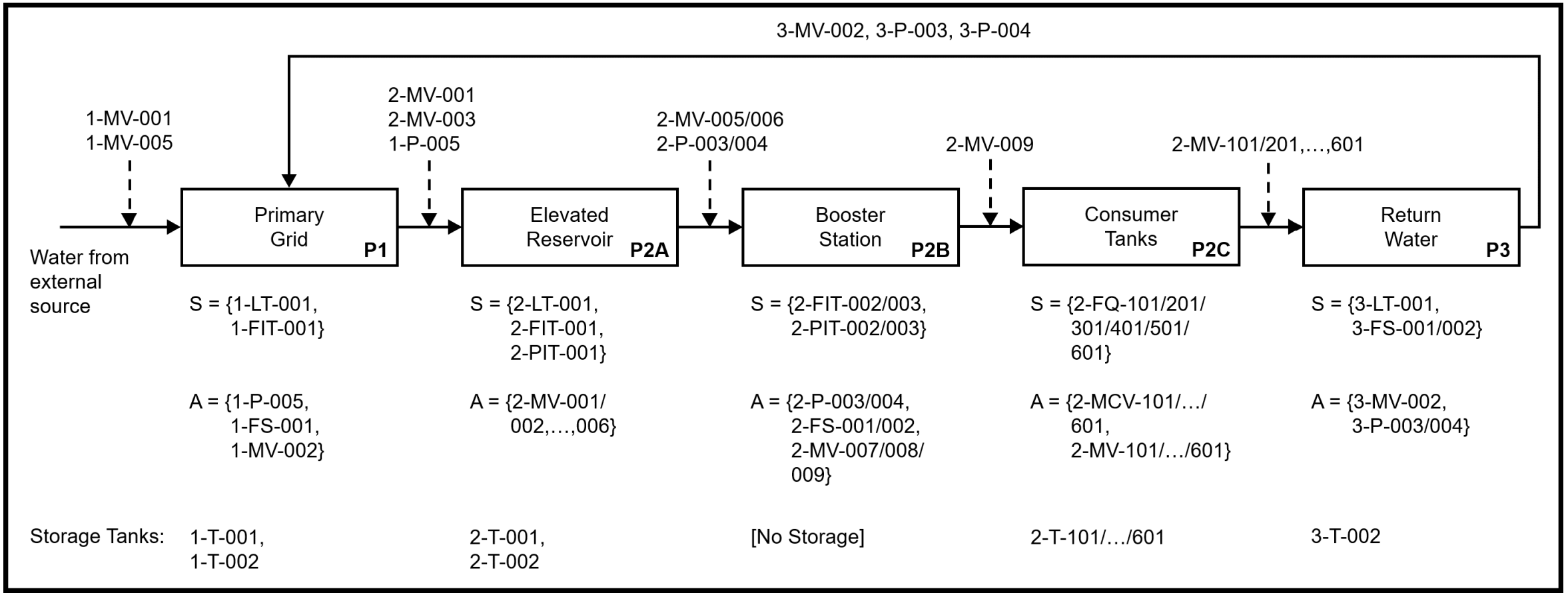


Architecture of the SWaT testbed

Testbeds: WADI



Testbeds: WADI



Architecture of the WADI testbed

System Modelling

- Actuators as control input, sensors as control output
- System model:

$$\begin{cases} x_{k+1} = Ax_k + Bu_k + v_k, \\ y_k = Cx_k + \eta_k \end{cases}$$

- The state-space matrices A, B and C capture the system dynamics and can be used to find a specific system state given an initial state.
- The sensor and process noise vectors are represented by η_k and v_k, respectively.

System Modelling

- Model validation: using Root Mean Square Error (RMSE)

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{N}}$$

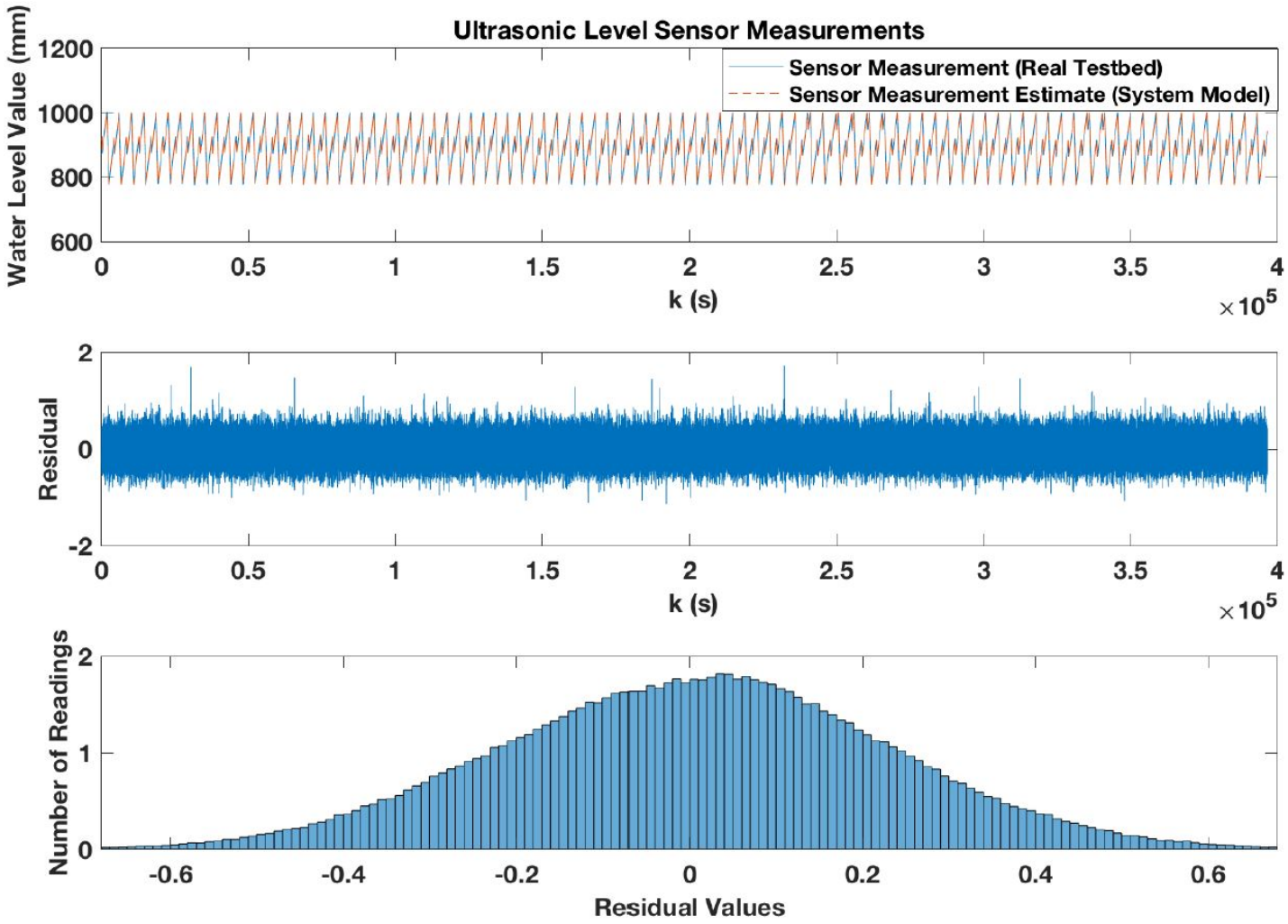
Sensor	FIT101	LIT101	LIT301	FIT301	LIT401	FIT401
RMSE	0.0363	0.2867	0.2561	0.0200	0.2267	0.0014
(1-RMSE)*100%	96.3670	71.3273	74.3869	98.0032	77.3296	99.8593

Table 1: Validating SWaT model obtained from sub-space system identification

- Accuracy as high as 70% is considered sufficiently precise*.

* Sensor fault detection and isolation for wind turbines based on subspace identification and kalman filter techniques. International Journal of Adaptive Control and Signal Processing, 2010
* Model-based attack detection scheme for smart water distribution networks. ASIACCS'17

Performance under Normal Operation

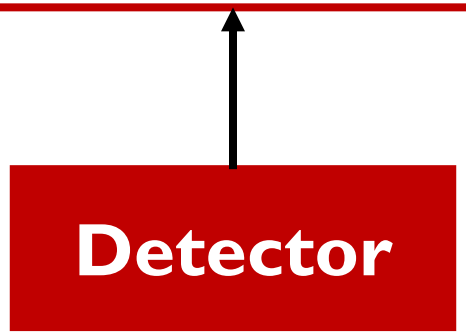


Validating system model obtained using sub-space system identification method for a level sensor in SWaT testbed

Attack Detection Framework

1) Estimation of the sensor output using the system model

2) Examination of the residual between the actual and estimated values and verifying the source of the sensor readings.



Attack Detection Framework

- Residual at time instance k:

$$r_k = y_k - \hat{y}_k$$

- Under *normal mode:* $E[r_k] = 0$
- Under *attack:* $E[r_k] \neq 0$

Attack Detectors: CUSUM

CUSUM: $S^-_{0,i} = 0, \quad S^+_{0,i} = 0, \quad \tilde{k}^+_i = 0, \quad \tilde{k}^-_i = 0,$

$$\begin{cases} S^+_{k,i} = \max(0, S^+_{k-1,i} + r_{k,i} - \bar{T}_i - \kappa_i), & \text{if } S^+_{k-1,i} \le \tau^+_i, \\ S^+_{k,i} = 0 \text{ and } \tilde{k}^+_i = \tilde{k}^+_i + 1, & \text{if } S^+_{k-1,i} > \tau^+_i. \end{cases}$$

$$\begin{cases} S^-_{k,i} = \min(0, S^-_{k-1,i} + r_{k,i} - \bar{T}_i + \kappa_i), & \text{if } S^-_{k-1,i} \ge \tau^-_i, \\ S^-_{k,i} = 0 \text{ and } \tilde{k}^-_i = \tilde{k}^-_i + 1, & \text{if } S^-_{k-1,i} < \tau^-_i. \end{cases}$$

Design parameters: bias $\kappa_i > 0$ and threshold $\tau_i > 0$.
Output: $alarm(s) = \tilde{k}^+_i + \tilde{k}^-_i$.

- CUSUM values $S^+_{k,i}$ and $S^-_{k,i}$ accumulate the distance measure $r_{k,i}$ (residual of i^{th} sensor) over time to measure how far are the values of the residual from the target mean ($\bar{T}_i$).
- Alarm is raised when the accumulation at any time instance k becomes greater or lesser than the chosen threshold τ_i.

Attack Detectors: Bad-data Detector

Bad-Data Procedure:

$$\text{If } |r_{k,i}| > \alpha_i, \quad \tilde{k}_i = k, \quad i \in \mathcal{I}.$$

Design parameter: threshold $\alpha_i > 0$.

Output: alarm time(s) $\tilde{k}_i$.

- Alarm is triggered if the distance measure, $|r_{k,i}|$, for the i^{th} sensor exceeds the threshold α_i at the time instance k.

Attack Detectors: *NoisePrint*

- When the system is in steady state, the residual vector obtained from the system model is a function of sensor and process noise**.
- Using system state estimation, it is possible to extract the sensor and process noise characteristics of the given industrial control system.
- Machine learning is applied on the residual vectors to fingerprint the given sensor and process.
- Detector design:
  - Residual collection
  - Data chunking
  - Feature extraction

** NoisePrint: Attack Detection Using Sensor and Process Noise Fingerprint in Cyber Physical Systems. ASIACCS '18

Threat Model

Attack classification based on execution

Single-point Attack	Multi-point Attack	Stealthy Attack
Targets a single point in the system	Multiple simultaneous target points	Minute alterations to sensor data

Data Injection Attacks

Bias Injection Attack

- Goal is to deceive the control system by sending incorrect sensor readings.
- Sensor reading is biased by a value of δ_k.
- Sensor value under attack: $\bar{y}_k = y_k + \delta_k$

Stealthy Attack

- The attack vector δ_k chosen such that it stays inconspicuous.
- The residual does not change noticeably or exceed the thresholds of the detectors.

Attack Simulations

Attack ID	Description (Initial State / Attack State)
Stage 1	
Atk-1-s	LIT101 = 659mm / change level +1mm/sec
Atk-2-s	LIT101 = 659mm / LIT101 = 850mm
Atk-3-s	LIT101 = 659mm / LIT101 = 210mm
Atk-4-s	LIT101 = 679mm / LIT101 = 700mm
Atk-5-s	LIT101 = 1029mm / LIT101 = 700mm
Atk-6-s	LIT101 = 789mm / LIT101 = 789mm
Atk-7-s	LIT101 = 784mm / LIT101 = 600mm
Stage 3	
Atk-8-s	L < LIT301 < H / LIT301 = HH+
Atk-9-s	L < LIT301 < H / change level -1mm/sec
Atk-10-s	L < LIT301 < H / change level -0.5mm/sec
Atk-11-s	FIT301 = 0 m^3/hr / FIT301 = 2 m^3/hr
Atk-12-s	L < LIT301 < H / water leakage attack
Stage 4	
Atk-13-s	FIT401 = 0.48m^3/hr / FIT401 = 0m^3/hr
Atk-14-s	LIT401 < 1000mm, P401 = ON / LIT401 = 1000mm and P401 = ON
Atk-15-s	L < LIT401 < H, P301 = ON / LIT401 = 600mm and P301 = ON
Atk-16-s	L < LIT401 < H / LIT401 < L
Atk-17-s	LIT401 = 1005mm / LIT401 = 1005mm

→ Stealthy attack

→ Bias injection attack

Table 2: List of attacks simulated on SWaT

Performance Metrics

- True Positive Rate (TPR)*** – the number of times the method correctly raises alarms over the duration of the attack.
- False Positive Rate (FPR) or False Alarm Rate (FAR) – the number of times the method incorrectly raises alarms in the absence of any attack.
- Time Taken for Detection (TTD) – the time taken by the procedure to raise an alarm in the event of an attack.

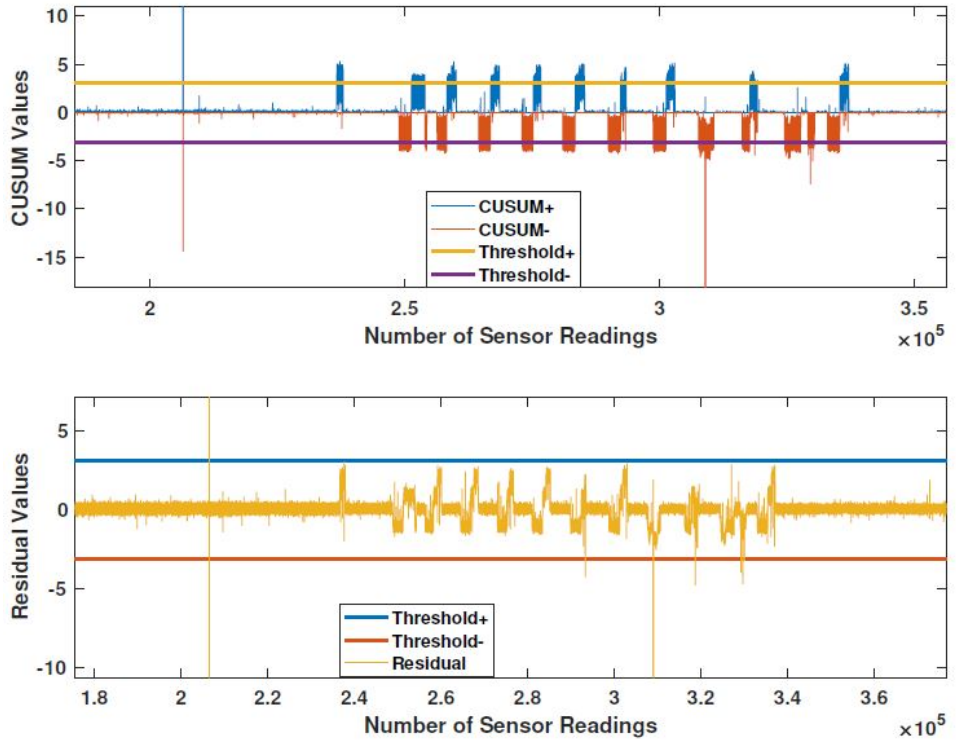
*** False Negative Rate (FNR) is an alternate way of expressing TPR: $\mathrm{FNR} = 100\ \% - \mathrm{TPR}$

Performance under Normal Operation

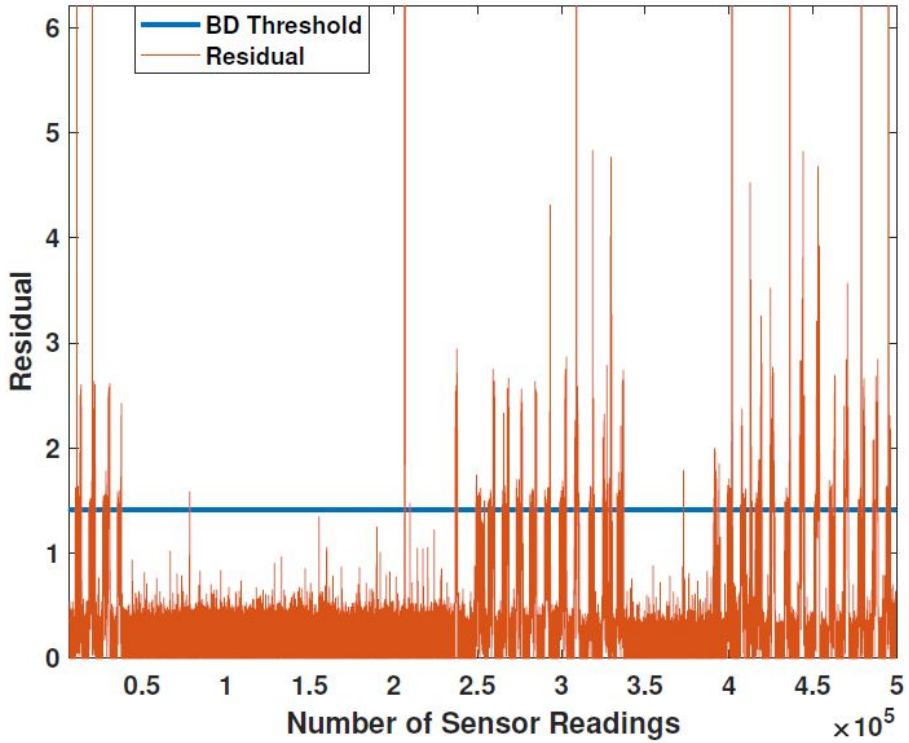
Sensor	FIT101	LIT101	FIT301	LIT301	FIT401	LIT401
CUSUM Detector						
Threshold	0.0149	3.1168	0.2209	0.5529	0.0156	0.5674
κ	0.0074	0.3117	0.0276	0.1382	0.0028	0.1135
FAR	5.54%	5.19%	5.34%	4.65%	4.02%	4.03%
Bad Data Detector						
Threshold	0.0205	1.4100	0.1184	0.4887	0.0108	0.4178
FAR	4.29%	5.32%	4.84%	4.56%	5.41%	5.42%
NoisePrint						
FAR	0%	1.29%	8.3%	2.44%	0%	0%

Table 3: False positives raised by the detectors under normal operation in SWaT

Performance under Normal Operation



(a) Bad-Data detection



(b) CUSUM detection

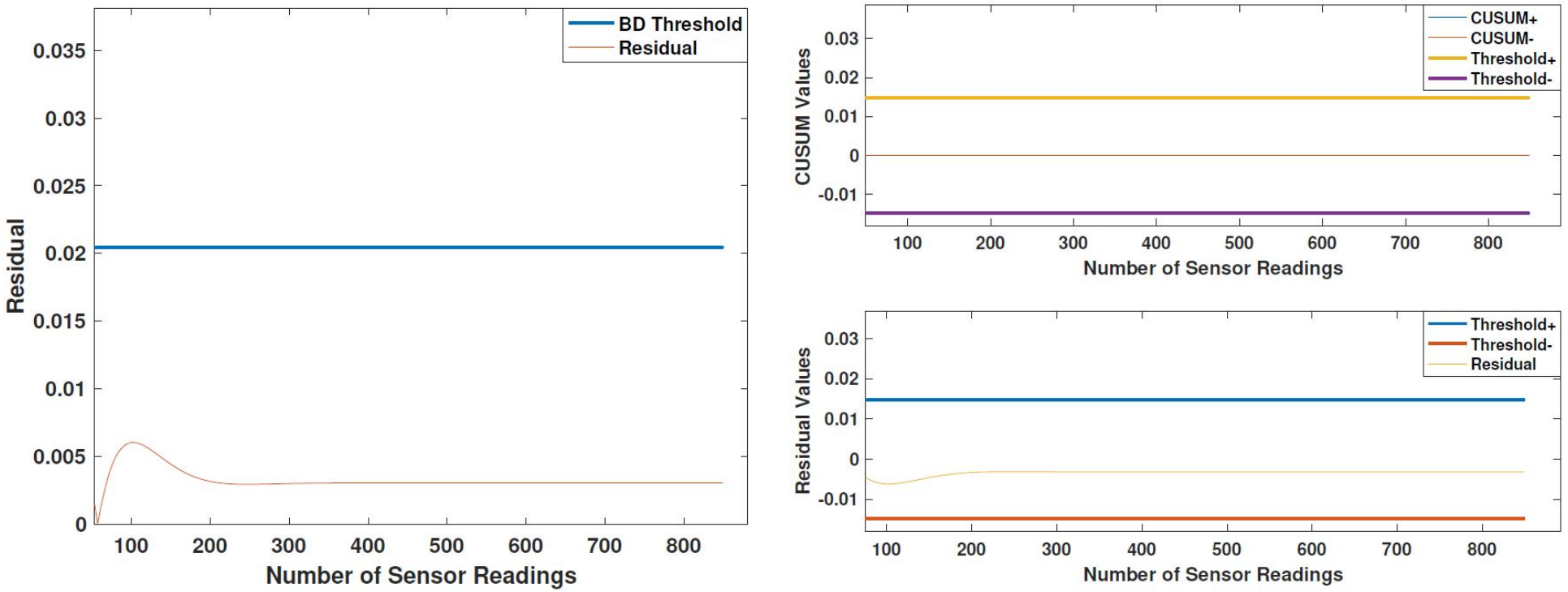
Statistical attack detection methods applied on the residual for level sensor (LIT-101) estimates from SWaT under normal operation

Performance under Attack

Attack	NoisePrint			CUSUM			Bad Data		
	TPR	FNR	TTD (s)	TPR	FNR	TTD (s)	TPR	FNR	TTD (s)
Single Point Attacks									
Atk-8-s	85.72%	14.28%	121.22	17.46%	82.54%	2	16.75%	83.25%	2
Atk-9-s	14.50%	85.50%	179	88.15%	11.85%	2	93.18%	6.82%	2
Atk-10-s	80.64%	19.35%	130.09	56.30%	43.70%	5	58.48%	41.52%	3
Atk-11-s	87.50%	12.50%	89.59	100%	0%	1	100%	0%	1
Atk-12-s	63.63%	36.37%	117.83	95.42%	4.58%	6	96.64%	3.36%	6
Atk-1-s	88.88%	11.12%	32.48	91.16%	8.83%	2	91.34%	8.66%	1
Atk-2-s	67.56%	32.44%	46.90	85.08%	14.92%	1	78.02%	21.98%	1
Atk-3-s	90.91%	9.09%	35.25	98.92%	1.08%	1	99.08%	0.92%	1
Atk-7-s	88.24%	11.76%	57.35	77.58%	22.42%	1	60.62%	39.38%	1
Atk-13-s	55%	45%	44.43	32.82%	67.18%	2	13.94%	86.06%	2
Atk-16-s	86.21%	13.79%	56.26	6.21%	93.79%	1	6.32%	93.68%	1
Multi-Point Attacks									
Atk-14-s	81.82%	18.18%	125.59	16.32%	83.68%	1	6.76%	93.24%	1
Atk-15-s	77.78%	22.22%	105.3	54.68%	45.32%	2	99.64%	0.36%	2
Atk-4-s	94.73%	5.26%	35.59	99.66%	0.34%	1	100%	0%	1
Atk-5-s	90.47%	9.53%	44.50	99.68%	0.32%	1	100%	0%	1
Stealthy Attacks									
Atk-17-s	80%	20%	67.03	0%	100%	**ND**	0%	100%	**ND**
Atk-6-s	75%	25%	174.84	0%	100%	**ND**	0%	100%	**ND**

Table 4: Attack Detection Performance on SWaT testbed

Performance under Attack



(a) Bad-Data detection

(b) CUSUM detection

Statistical attack detection methods applied on the residual for level sensor (LIT-101) estimates from SWaT under stealthy attack

Performance Remarks (Attack Detection)

Statistical Detectors	***NoisePrint***
• Successful detection of basic attacks, such as bias injections. • Faster detection time. • Fail under stealthy attacks.	• Better overall accuracy. • Able to detect stealthy attacks, since replication of process and sensor noise can be difficult. • Slower speed of detection.

General Comments/Challenges

- Practicality of model-based approach:
  - Testbeds used are small-scale and obtaining complete system models for them was a feasible task.
  - Larger industrial plants could be divided into several sub-systems (based on the processes taking place) and have multiple models corresponding for each.
- Obtaining a normal reference system model for the plants and sensors sensitive to environmental disturbances (e.g., for the WADI testbed) is a non-trivial task:
  - Noise from the environmental disturbances on the system's processes causes unpredictable deviations from its modelled behavior.

General Comments/Challenges

- Sensor faults under normal operation: hindered the creation of useful system models.
- Data availability and reliability:
  - Dataset for model creation obtained after the plants were run continuously under normal operating conditions.
  - However, unexpected results were obtained when the system models were tested when the plants were not running.

Conclusions

It is deduced that bias injection attacks on sensors that are quite similar to faults can be easily detected using statistical techniques like Bad-Data and CUSUM detectors.

However, it is observed that advanced stealthy attacks require more sophisticated detection methods, like *NoisePrint*.

While detection methods must be able to demonstrate accuracy, their attack detection speed is also a crucial metric for critical CPSs.

Acknowledgements

This work was supported by the SUTD start-up research grant SRG-ISTD-2017-124.

The authors thank the reviewers for their comments.

The authors express their gratitude to the iTrust research centre at Singapore University of Technology and Design for their research facilities, which have been extensively used in this work.

Thank You