Limitations of State Estimation Based Cyber Attack Detection Schemes in Industrial Control Systems

Chuadhry Mujeeb Ahmed
Singapore University of Technology
and Design, Singapore
Email: chuadhry@mymail.sutd.edu.sg

Sridhar Adepu
Singapore University of Technology
and Design, Singapore
Email: sridhar_adepu@sutd.edu.sg

Aditya Mathur
Singapore University of Technology
and Design, Singapore
Email: aditya_mathur@sutd.edu.sg

***Abstract*—An experiment was conducted on a water treatment plant to investigate the effectiveness of using Kalman filter based attack detection schemes in a Cyber Physical System (CPS). Kalman filter was implemented with Chi-Square detector. Random, stealthy bias, and replay attacks were launched and results analysed. Analysis indicates that stealthy false data injection and replay attacks cannot be detected by legacy failure detection methods.**

I. Introduction

Cyber Physical Systems (CPS) are embedded systems composed of computing as well as physical processes. With the advent of networked control systems to enable better operations and monitoring of the physical processes, these systems become part of cyber space. These advancements help manage critical infrastructures which are essential components of a smart city, like public transportation, smart grid and water treatment facilities but have created new challenges [1]. One such challenge arises in ensuring the security of the critical infrastructures. The nature of security threats and attacks is different from those found in corporate networks [2], [3]. Threat models have evolved significantly and the fact that any successful attack could be fatal as it is more than a computer being hacked and disturbing the physical process may result in unsafe scenarios. Besides cyber attacks, physical attacks are also possible.

The study reported here is to examine from a control theoretic perspective techniques for detecting attacks on a CPS. The objective is to investigate these techniques on an autonomous water treatment (SWaT) testbed (see Section II). In this context, an adversary alters sensor readings as in a Man-In-The-Middle attack. Using the knowledge of the structure and operation of the system, an adversary can launch an intelligent attack by changing sensor values within its operating limits but yet affecting the plant's performance. During such an attack, estimator like Kalman filter would best try to track sensor measurements and remove the noise. The estimator follows the altered values of the sensor and the difference (residue) between measured values and the estimation vanishes as soon as convergence is achieved.

Related Work: The impact of attacks on a CPS, and techniques to detect and mitigate their effects, has been reported. In [4] it is concluded that perfect estimation is not feasible when half of the sensors are under attack. One key assumption here is that a local controller has complete access to the system state. In [5] physical watermarking is used whereby a known noise sequence is added to control input and it's effect on sensor readings evaluated. The assumption is that unaware of the added noise, an adversary would not be able to alter sensor values without being detected. This method also uses the Kalman filter based χ^2 detection mechanism.

In [6], false data injection attack detection in smart grid systems using Kalman filter is studied. The method shows success against random and DoS attacks but fails in case of false data injection attack. Attacks on power grid for false data injection attacks have been studied [7]. Generalised false data injection attack is introduced where an adversary adds a bias to measured data which does not change the residue significantly and could not be detected. These mentioned works point to the limitations of estimation based detectors against stealthy attacks using simulation models. We expose limitations of bad data detection schemes in detecting strategic attacks by implementing these methods in an operational water treatment testbed. Attack vectors are systematically designed and carried out by compromising links between PLC and the SCADA workstation. We demonstrate the success of sophisticated attacks.

The remainder of the paper is organised as follows. Section II describes the architecture of SWaT testbed used in this experiment. In Section III system model along with Kalman filter and detector is explained. Attacker and attack models used in this work are explained in Section IV. Section V describes the experimentation setup and discusses the results. Section VI, concludes the work.

II. Architecture of the SWaT Testbed

SWaT [8] is a fully operational (research facility), scaled down water treatment plant producing 5 gallons/minute of

This work was supported by the National Research Foundation (NRF), Prime Minister's Office, Singapore, under its National Cybersecurity R&D Programme (Award No. NRF2014NCRNCR001-40) and administered by the National Cybersecurity R&D Directorate.

978-1-5090-2924-2/16/$31.00 ©2016 IEEE

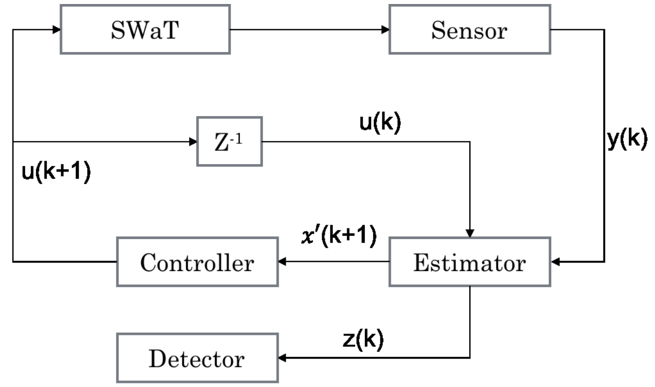


Fig. 1: System Model for attack detection using Kalman filter.

doubly filtered water, this testbed mimics large modern plants for water treatment [2].

Water treatment process: The treatment process in SWaT [8] consists of six distinct stages each controlled by an independent Programmable Logic Controller (PLC). Control actions are taken by the PLCs using data from sensors. Stage P1 controls the inflow of water to be treated by opening or closing a motorised valve MV-101. Water from the raw water tank is pumped via a chemical dosing station (stage P2, chlorination) to another UF (Ultra Filtration) feed water tank in stage P3. A UF feed pump in P3 sends water via UF unit to RO (Reverse Osmosis) feed water tank in stage P4. Here an RO feed pump sends water through an ultraviolet dechlorination unit controlled by a PLC in stage P4. This step is necessary to remove any free chlorine from the water prior to passing it through the reverse osmosis unit in stage P5. Sodium bisulphate (NaHSO3) can be added in stage P4 to control the ORP (Oxidation Reduction Potential). In stage P5, the dechlorinated water is passed through a 2-stage RO filtration unit. The filtered water from the RO unit is stored in the permeate tank and the reject in the UF backwash tank. Stage P6 controls the cleaning of the membranes in the UF unit by turning on or off the UF backwash pump.

Communications: Each PLC obtains data from sensors associated with the corresponding stage, and controls pumps and valves in its domain. PLCs communicate with each other through a separate network. Communications among sensors, actuators, and PLCs can be via either wired or wireless links. Attacks that exploit vulnerabilities in the protocol used, and in the PLC firmware, are feasible and could compromise the communications links between sensors and PLCs, PLCs and actuators, among the PLCs, and the PLCs themselves. Having compromised one or more links, an attacker could use one of several strategies to send fake state data to one or more PLCs.

III. Attack Detection Framework

In this section a model of the system and attack detection strategy using Kalman Filter is explained. The notion of attack detection in a CPS using Kalman filter essentially relies on legacy techniques designed for failure detection [9]. Kalman filter is a linear filter used to remove noise from the measurements and derive an estimate of the state being observed. Figure 1 shows the overall system model for the attack detection procedure. Kalman filter takes sensor values as an input and estimates the values for the state variables. The output from the Kalman filter and observed values is passed to a detector where both values are compared and evaluated against a precomputed threshold. If the difference is higher than the threshold computed from system observations in normal operations an alarm is triggered indicating an attack.

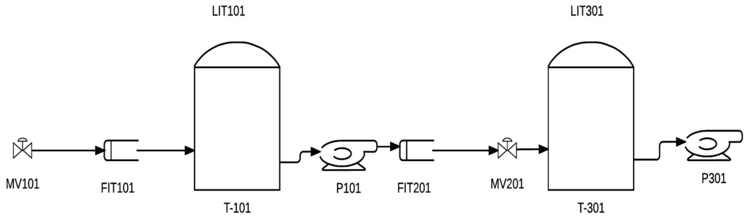


Fig. 2: Stage 3 of SWaT.

A. State Space Model

Attacks on stage three of the SWaT testbed were carried out and a level sensor in tank-3 (T301) is considered under attack. As shown in figure 2, for T301 we have constant input and output flow and hence the filling rate is known. Given the initial state of the level sensor and the rate of fill, the level of T301 level can be modelled as follows.

$$x(k+1) = Ax(k) + Bu(k) + w(k) \tag{1}$$

where $x(k) \in \mathbb{R}^{\mathrm{n}}$ is the state variable vector at time k, $u(k) \in \mathbb{R}^{\mathrm{p}}$ is the control input and $w(k) \in \mathbb{R}^{\mathrm{n}}$ is the process noise at time k.

Sensors are deployed to monitor several parameters in the water treatment plant. These include flow meters, level sensors, pressure gauge, and sensors to check the chemical properties of water. Measurements from these sensors can be written as an observation equation.

$$y(k) = Cx(k) + Du(k) + v(k), \tag{2}$$

where $y(k) \in \mathbb{R}^{\mathrm{m}}$ is the measurement from a sensor at time k and $v(k) \in \mathbb{R}^{\mathrm{m}}$ is the measurement noise at time k which is independent of $w(k)$.

B. Kalman Filter

Figure 1 is an overview of the system model where Kalman filter is used for estimation; the detector makes use of the residual to decide whether there is an attack. After δt time unit, values from the sensors are sampled and a decision. In SWaT, $\delta t = 1$. The estimator computes an estimate at each time step based on the previous reading up to $x(k-1)$ and sensor reading $y(k)$. The estimator provides $\hat{x}(k)$ an estimate of the state variable $x(k)$. Thus, an error term can be defined as,

$$e(k) = \hat{x}(k) - x(k) \tag{3}$$

where $\hat{x}(k|j)$ denotes the optimal estimate for $x(k)$ given the measurements $y_1, ..., y_j$. Let $P(k)$ denote the error covariance, $Cov(e(k)) = E[(x(k) - \hat{x}(k))(x(k) - \hat{x}(k))^T]$ and $\hat{P}(k|j)$ denote the estimate of $P(k)$ given $y_1, ..., y_j$. Prediction equation for state variable using Kalman filter can be written as,

$$\hat{x}(k+1|k) = A\hat{x}(k|k) \tag{4}$$

$$P(k+1|k) = AP(k|k)A^T + Q, \tag{5}$$

where $\hat{x}(k|k)$ is estimate at time step k using measurements up to time k and $\hat{x}(k+1|k)$ is the $(k+1)^{th}$ prediction based on previous k measurements. Similarly, $P(k|k)$ is the error covariance estimate up untill time step k. Q is the process noise covariance matrix. The next step in Kalman filter estimation is the time update step using Kalman gain K.

$$K(k) = P(k|k-1)C^T(CP(k|k-1)C^T + R)^{-1} \tag{6}$$

$$\bar{x}(k+1|k) = \hat{x}(k+1|k) + K(k)(y(k) - C\hat{x}(k+1|k)) \tag{7}$$

$$\bar{P}(k+1|k) = (I - K(k)C)\hat{P}(k+1|k), \tag{8}$$

where $\bar{x}(k+1|k)$ and $\bar{P}(k+1|k)$, are the updates for time step $k+1$ using measurements y_i from the i^{th} sensor and Kalman gain $K(k)$. R is the measurement noise covariance matrix. We can choose the initial state as $x(0) = x_0$ with $P(0) = E[(\hat{x}_0 - x_0)(\hat{x}_0 - x_0)^T]$. Kalman gain $K(k)$ is updated at each time step but after a few iterations it converges and operates in a steady state. Kalman filter is an iterative estimator and $\hat{x}(k|k)$ in equation 4 comes from $\bar{x}(k-1|k)$ in equation 7.

C. Detection Procedure

Failure detectors have been used along with Kalman filter to detect process anomalies. A detector often used is χ^2. Other detectors have also been proposed [10]. Here we look at the largest residue based detection method and the χ^2 based scheme. For a χ^2 detector the residual quantity $z(k+1)$ at time step $k+1$ is needed.

$$z(k+1) \triangleq y(k+1) - \bar{y}(k+1) \tag{9}$$

Equivalently,

$$z(k+1) \triangleq y(k+1) - C(A\bar{x}(k+1)). \tag{10}$$

The χ^2 test consists of following expression.

$$g(k) = z(k)^T \Sigma z(k), \tag{11}$$

where Σ is the covariance matrix of the residue vector $z(k)$. The χ^2 detector compares $g(k)$ with a precomputed threshold based on standard χ^2 table [6]. If $g(k) > threshold$ an attack is assumed to be detected. A second method also uses residue $z(k)$ but the threshold is computed from healthy sensor measurements. The system is run without any attack for some time and the residue calculated using the maximum value from vector $z(k)$ as a threshold. Then residue $z(k)$ is compared with threshold to identify an attack.

IV. Attacker and Attack Models

In this section types of attacks launched on SWaT testbed as well as attacker model are summarised. The attacker model essentially contains the objectives of an attack and the attacker's intentions. An attacker may choose the goals from a set of intentions [2], including performance degradation, disturbing a system property, and damaging a component. In our experiments, the attacker's intention is to overflow tank-3 (T301) and cause damage to the plant and its surroundings. Three attacks are modelled and executed. It is assumed that the attacker has knowledge of the system dynamics and knows the true sensor measurements $y(k)$ for all k and can modify sensor reading to arbitrary values y_a(k).

1) *Random Attack*: First, a failure like attack (random) attack is designed where the attacker's goal is to deceive the control system by sending incorrect sensor values. In this scenario sensor LIT-301 values are decreased, when the actual tank level is close to the high (H) mark. Doing so may make the controller believe that the attacked values are true sensor readings and maintain pump P-101 to its *ON* state, ultimately overflowing T301. The attack vector can be defined as,

$$y_a(k) = y(k) \pm \Delta(k), 0 < k < \tau, \tag{12}$$

where $\Delta(k)$ is the modification by attacker for τ units of time (period of attack).

2) *Stealthy Bias Injection Attack*: The second attack is more sophisticated and is is carried out by injecting a bias in sensor values over time to drive the system to an undesired state. In particular, during the tank (T301) filling process, the attack subtracts a small value from the level sensor (LIT-301) reading. If continued for some time, the tank may physically be filled but the aggregate bias in sensor readings will show much lower level to the controller which keeps the pumping potentially causing an overflow.

$$y_a(k) = y(k) \pm \Delta(k), 0 < k < \tau \tag{13}$$

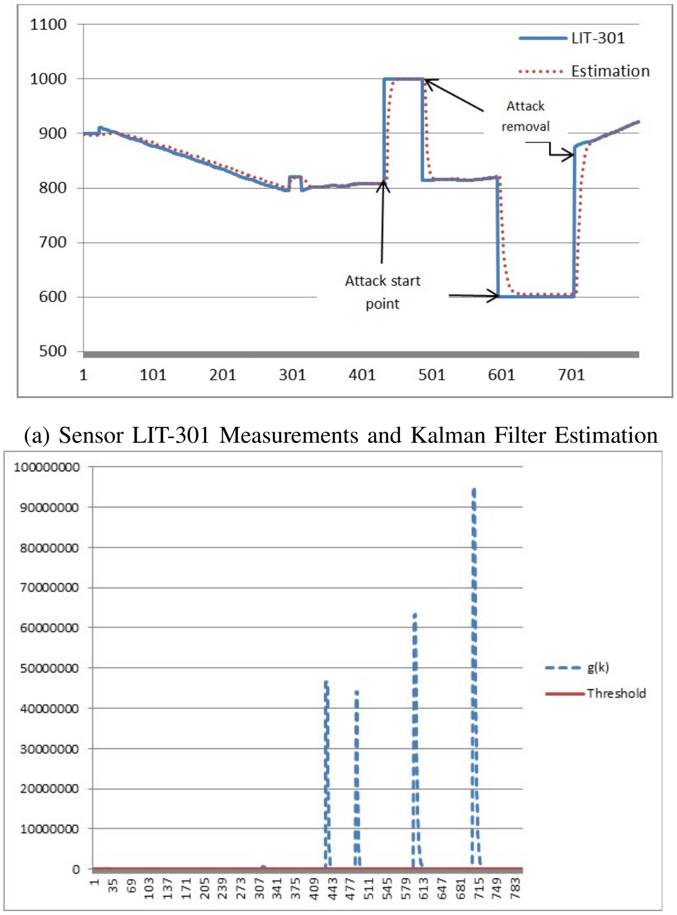
$$y(k+\tau) = y_a(k) \tag{14}$$

where $\Delta(k)$ is the modification by attacker for τ time. Where τ is time period for which attacker keep the bias and then iteratively update equation 14 and again execute attack as in equation 13 until attack objectives are achieved. The difference between this and random attack is that here $\Delta(k)$ is chosen as such that residue based detection method fails.

3) *Replay Attack*: In the third attack the attacker replays previously recorded sensor thereby moving the system into an incorrect state.

$$y_a(k) = y(k-\tau), 0 < k < \tau - 1 \tag{15}$$

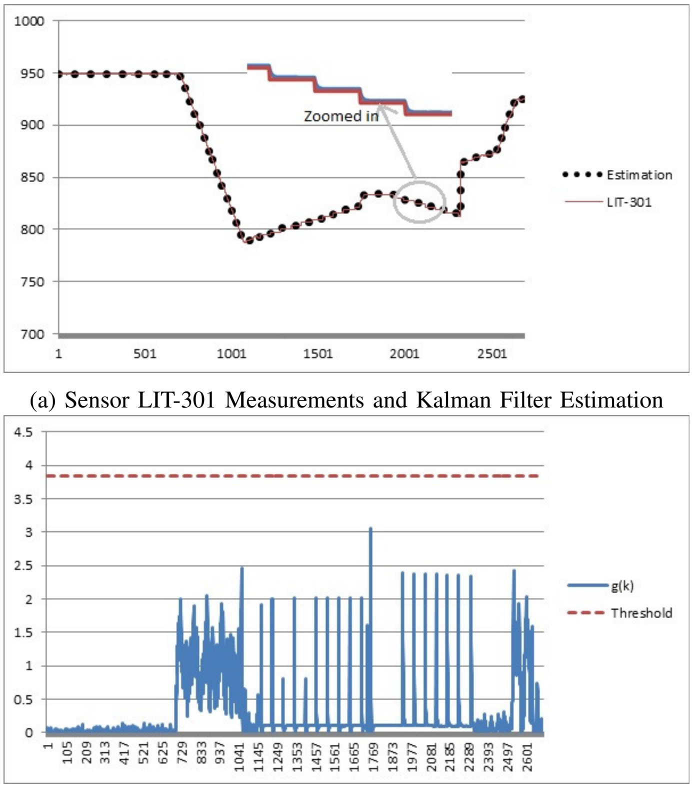
Here the attacker replays $y(k)$ measurements with τ delay. The intuition for failure of estimation based detection schemes evolves from the fact that if the injected



(a) Sensor LIT-301 Measurements and Kalman Filter Estimation

(b) Chi-Square Detection

Fig. 3: Random Attack



(a) Sensor LIT-301 Measurements and Kalman Filter Estimation

(b) Chi-Square Detection

Fig. 4: Stealthy Bias Attack

value and the previous steady state measurements differ in a small amount, then the residue would not get higher than the threshold. This happens because Kalman filter would try its best to follow the sensor measurements. If those measurements are changed suddenly by a large amount in the case of a random attack, the residue for few time steps would be high as the Kalman filter takes some time to converge to new induced (attacked) values.

V. Implementation and Performance Evaluation

The Kalman Filter, χ^2 and threshold based detector in the PLC (P3) of the six stage SWaT testbed, were implemented. Attacks were launched on a level sensor (LIT-301) of tank 3 (T301) in stage 3. As shown in the simplified version of the testbed in Figure 2, we can find the filling rate in the tank, using the input and output flow rates, that the information is what we needed to model the level in T301 as a sate variable ($x(k)$). Therefore, Kalman filter equation for prediction becomes,

$$\hat{x}(k+1|k) = \hat{x}(k|k) + |(inflow - outflow)|, \qquad (16)$$

where $A = 1$ is considered as only single state, i.e., the water level in T301 is modelled. Hence all the matrices, $A, C, Q, R, andP$ become scalars. Arbitrary initial values for $x(0)$ and $P(0)$ are used. R is found using data from the level sensor. Q is taken as a small value because we know accurately the system model for T301 based on system. The system is run under normal operating conditions to obtain its normal behaviour profile. Then all three attacks are launched one by one, as explained above.

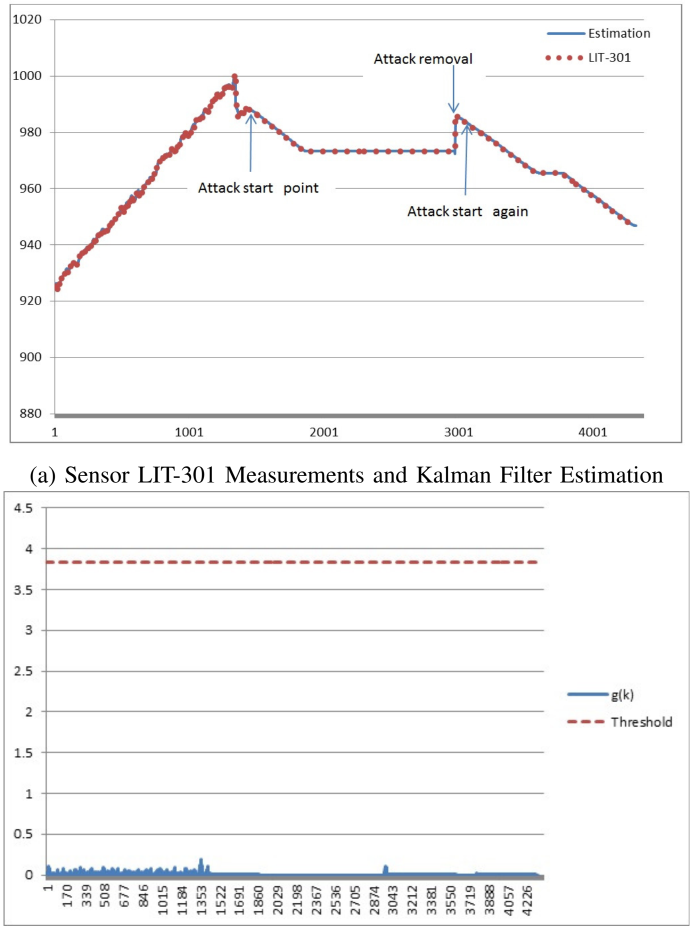
Prior to launching the attacks, SWaT is brought to a steady state, i.e., it operates under normal conditions for some time so that once an attack is launched the response could be observed and compared with that under normal conditions. To achieve attacker's objectives attacks are performed when system was in a specific state. Attacks are carried out by manipulating the values of a particular *tag* (a memory location in PLC to store sensor data). Readings from level sensor LIT-301 (level sensor in tank T301) were manipulated. This was done by attacking the level 2 network in SWaT, i.e., the link between PLCs and SCADA workstation.

The results for sensor measurements (LIT-301) and Kalman filter estimation of their values under normal conditions are not shown here due to limitations of space. The plant was run for several hours to obtain values of $\bar{x}(k)$ and the residue $z(k)$. For simple threshold based detection method the largest residue was used as a threshold and the residue values compared under attack with this threshold to distinguish between normal operation and operation under attack. The results from largest residue test were found to be in accordance with the χ^2 detector. For the χ^2 detector, the threshold from standard table was found such that the error rate is less than 5% for a single degree of freedom (single state variable) [6].

Figure 3 shows how the detector successfully triggers an alarm when $g(k)$ values cross the threshold. Since the attack was executed by injecting a random value, the Kalman filter eventually converges making the controller believe that the attacked value is the real sensor reading and tracks $y_a(k)$. Essentially this attack is detected because it resembles a sensor failure with abrupt change in sensor data.

Kalman filter based fault detection techniques are effective but detecting a strategic attack is not possible. The sensor is under attack for a long time as specified in the plot but $g(k)$ shoots up instantly and the residue has lower values for most of the duration of attack because of the convergence of Kalman filter. $g(k)$ values are high because χ^2 uses variance of $z(k)$, which is high in this case. Figure 4 shows the plots for stealthy bias attack where an attacker subtracts a small value from the real sensor (LIT-301) reading in such a way that Kalman filter output converges without it's residue going above threshold. If this kind of attack carried out for some time, it will lead to an overflow in tank T301. Small bias injection exploits the fact that Kalman filter considers these small changes as noise and converges fast without attack being detected. Figure 4a shows a zoomed-in part of the plot during attack. Here it is seen that a small bias is subtracted every two minutes. When an attack was removed after some time, there was significant difference between the real and estimated (attacked) values. Figure 4b shows that $g(k)$ always stayed below threshold.

Replay attack results are shown in Figure 5. An attacker recorded the readings for few complete cycles of the process related to T301 and launched the replay attack at an appropriate time driving the system to an undesired state. Sensor measurements for LIT-301 are replayed and since this is kind of hard attack to detect, residue and χ^2 plots are no different than those found under normal operation.



(a) Sensor LIT-301 Measurements and Kalman Filter Estimation

(b) Chi-Square Detection

Fig. 5: Replay Attack

VI. Conclusion

Limitations of state estimation based attack detection schemes for CPS are discussed in this article. Implementation on a testbed in a realistic setting offers valuable insights. It is observed that such techniques are not effective in detecting stealthy false data injection and replay attacks. Only random attacks that lead to sensor data as if it is faulty, get detected. This is a well known use of linear filters for dynamic systems. It is concluded that Kalman filter and threshold based detection schemes should not be employed as a defence mechanism against strategic attacks.

References

[1] Edward A. Lee. Cyber physical systems: Design challenges, http://www.eecs.berkeley.edu/Pubs/TechRpts/2008/EECS-2008-8.html. Technical Report UCB/EECS-2008-8, EECS Department, University of California, Berkeley, Jan 2008.

[2] S. Adepu and A. Mathur. An investigation into the response of a water treatment system to cyber attacks. In *Proceedings of the 17th IEEE High Assurance Systems Engineering Symposium, Orlando*, January 2016.

[3] B. Zhu, A. Joseph, and S. Sastry. A taxonomy of cyber attacks on SCADA systems. In *Internet of Things (iThings/CPSCom), International Conference on and 4th International Conference on Cyber, Physical and Social Computing*, pages 380–388, 2011.

[4] H. Fawzi, P. Tabuada, and S. Diggavi. Secure estimation and control for cyber-physical systems under adversarial attacks. *IEEE Trans. Autom.Control*, 59(6):1454–1467, 2014.

[5] Yilin Mo, S. Weerakkody, and B. Sinopoli. Physical authentication of control systems: Designing watermarked control inputs to detect counterfeit sensor outputs. *Control Systems, IEEE*, 35(1):93–109, Feb 2015.

[6] K. Manandhar, Xiaojun Cao, Fei Hu, and Yao Liu. Detection of faults and attacks including false data injection attack in smart grid using Kalman filter. *EEE Transactions on Control of Network Systems*, 1(4):370–379, Dec 2014.

[7] Y. Liu, P. Ning, and M. Reiter. False data injection attacks against state estimation in electric power grids. In *Proceedings of the 16th ACM Conference on Computer and Communications Security*, pages 21–32, 2009.

[8] SWaT: Secure Water Treatment Testbed, 2015. https://itrust.sutd.edu.sg/wp-content/uploads/sites/3/2015/11/Brief-Introduction-to-SWaT_181115.pdf.

[9] B. Brumback and M. Srinath. Chi-square test for fault-detection in kalman filters. *IEEE Trans. Autom. Control*, June 1987.

[10] H. L. Jones. *Failure Detection in Linear Systems*. PhD thesis, M.I.T., Cambridge, 1973.