SANDIA REPORT
SAND2018-12994
Unlimited Release
Printed August 2018

Development of Defenses against False Data Injection Attacks for Nuclear Power Plants

Yeni Li and Hany S. Abdel-Khalik, Elisa Bertino, and Arvind Sundaram

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from

U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@osti.gov
Online ordering:	http://www.osti.gov/scitech

Available to the public from

U.S. Department of Commerce
National Technical Information Service
5301 Shawnee Rd
Alexandria, VA 22312

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.gov
Online order:	https://classic.ntis.gov/help/order-methods/

SAND2018-12994
Printed August 2018
Unlimited Release

Development of Defenses against False Data Injection Attacks for Nuclear Power Plants

Yeni Li, Hany S. Abdel-Khalik, Elisa Bertino, and Arvind Sundaram
School of Nuclear Engineering, Purdue University
Sandia National Laboratories
P. O. Box 5800
Albuquerque, New Mexico 87185-MSXXXX

Abstract

With the recent successful attempts against the digital control systems of critical infrastructures, there is a need to develop new defense strategies that recognize two important realities, 1) state-sponsored attackers can rely on a number of techniques including espionage, social engineering, and brute force techniques, etc. to gain access to the raw data used to control system behavior, 2) attackers can falsify operational data in manners that do not trigger conventional outlier/anomaly detection techniques in order to go undetected, which is referred to as false data injection attacks. Therefore, there is a strong need to explore another class of defense measures, referred to as physical process defense, serving as a new line of defense in the event existing defenses relying on information protection measures are breached. This physical process defenses utilize the physics and engineering models of the system to build unique signatures for genuine system behavior. If successful, the signatures would be able to detect attacks that falsify the operational data and render them harmless before they can inflict physical damage on the system. This report is focused on exploring the feasibility of physical process defenses for nuclear reactors, and their associated functional requirements to maximize their resiliency against state-sponsored, or equivalent, attackers.

TABLE OF CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

Motivated by the growing frequency and level of sophistication of cyber-attacks against critical infrastructures, specifically focusing on nuclear power plants, this work explores the feasibility and functional requirements of a new class of defense methods relying on the use of physics simulation of the system, referred to as physical process defenses. The idea is to create a signature for the system/entity to be protected rather than the incoming attack. The major findings of this study are:

- Given that most of the technical know-how on nuclear plants design and operation is in the open literature, attackers can develop self-learning techniques to approximate the reactor's dynamical state, even if some of the design details are missing.
- The self-learning of system behavior must rely on physics models that emulate system behavior; meaning that pure neural network-based models will not be sufficient to learn system behavior in a manner that can produce reliable predictions for the wide range of reactor conditions. In the event some of the modeling details are missing, attackers can supplement their lack of knowledge by continuously monitoring reactor measurements and employing estimation/inference techniques to determine the missing details. This can be done using either active or passive monitoring. In passive monitoring, the learning techniques rely on observing reactor measurements collected by the network of sensors without interfering in reactor operation, whereas in active monitoring, the attackers injects small perturbations in the control network to affect changes to reactor state to improve the self-learning process or to prove their ability to control reactor state.
- Once the dynamical behavior of the reactor is learned, attackers could falsify the data displayed to the operator in a manner that does not trigger existing plant consistency checks, which are typically based on conservation principles. The ability to falsify data displayed to the operator may be used to launch two types of attacks. The first is designed to initiate an event which is typically identified by a number of well-known scenarios, e.g., reduction in heat removal event. The second attack is designed to alter normal (as-designed) reactor response to an event, by changing additional parameters used by the control system to guide the system during an event, e.g., trip set points.
- Defense measures relying on creating signatures for reactor behavior may be introduced in either passive or active manners. Absent any obscurity measures, the passive defense is not expected to be resilient as it will be based on the sensors data and the physics models, which are both available to the attackers. By obscuring the mathematical definition of the signatures, passive defenses can be effective.
- Active defenses are expected to be more resilient in the event attackers have full access to the defense strategy. Active defenses introduce small perturbations to the actuators commands and the sensors readings in order to satisfy two conditions, first they are small enough, i.e., within the noise, such as not to impact system behavior, and second, their definition is based on randomized application of data mining techniques, where the randomization affects the choice of the subset of components (including both dominant and weak components) used for building signatures. This ensures that brute forcing techniques will not be feasible approach to break the defense algorithm, since there is no clear sequence of steps to follow such as the case with encryption techniques which are widely known, with only few pieces missing, i.e., the encryption keys.

1. INTRODUCTION

The increased frequency and level of sophistication of cyberattacks taking place in recent years against digital control systems of critical infrastructures[1] have heightened concerns over such attacks being directed towards nuclear power reactors. Digital control systems are typically referred to as the SCADA[2], short for supervisory control and data acquisition systems, which are used to manage, supervise, and operate a wide range of industrial systems, including nuclear power plants, fossil plants, chemical plants, water treatment facilities, etc. SCADA systems continuously collect performance data about the systems using distributed network of sensors and continuously issues commands to actuators to keep the system operating per design specifications. Access to SCADA systems is currently being protected using perimeter defenses (e.g., routers, firewalls, cryptography, etc.), which are designed to stop unauthorized access. It is recognized that research is currently underway to assess the robustness and resilience of already-in-place defense measures, such as "perimeter defenses" and "information security" measures, which are designed to stop unauthorized access to the SCADA traffic. While this research is criticality needed, our work will focus on building another layer of defense, assuming that existing defenses have already been bypassed. This new layer of defense is designed to protect the SCADA traffic from malicious manipulation, which could be done via systematic falsification of the performance data and/or modification of the commands to actuators[2]. This new layer is designed to protect the system at the basic process level, and is thus based entirely on the physical principles governing the system behavior and derives its strength from the unique design and operational characteristics of the engineering system. This layer of defense is referred to hereinafter as physical process defense.

In designing physical process defenses, one must take into account a number of unescapable realities for the class of digital control systems of interest, such as those employed in nuclear reactors, a) the technical know-how on most critical infrastructures such as nuclear reactors, fossil plants, water treatment facilities, electric power grid, etc., exists worldwide and is accessible to the attackers; b) security by obscurity is never an effective approach, because most breaches involve insiders implying that any strategy relying on hiding or obfuscating the details of its operation will eventually be hijacked; c) defense techniques that rely on secret passphrases, private keys, etc. can eventually be bypassed with enough computing resources that are available to state-sponsored attackers; d) reliability methods based on probabilistic and dynamic risk assessment, are incapable of distinguishing between malicious/deliberate accidents (as manufactured by the attackers) and normal accidents that could result from equipment failure, especially when all information displayed to the operators are falsified by the attackers [3].

Particularly we focus in this work on how much the attacker can do in light of the above realities. Thus, the work will follow three thrusts. ***In the first thrust***, we explore whether an attacker equipped with knowledge about reactor system can predict the reactor dynamical state. To do that, we employ a simplified point kinetics models with a number of unknown parameters and explore how inference techniques may be used to fully determine the dynamical behavior of the system. ***In the second thrust***, we explore whether the predictive models used by the attackers need to be based on physics simulation, or they can be constructed using machine learning techniques, which are purely data driven. The data employed by this study, emulating a number of critical reactor observables, are generated using a thermal hydraulics model of the secondary and primary loops of a representative reactor. **The third thrust** provides recommendations on the types of attacks that may be initiated and provides ideas on how to develop monitoring systems that are capable of detecting intrusive false data injection attacks.

With regard to the first thrust, the concept of online self-learning is not new to the nuclear engineering community. In fact, there exists a strong arsenal of algorithms which have been developed for condition monitoring and online senor validation[4][5][6]. These algorithms employ a form on online self-learning techniques to characterize the correlations between the various signals to determine the reliability and robustness of signals against normal statistical noise and load fluctuations. In the hands of the attackers, these methods can be exploited to design attack signals that can bypass basic statistical checks, hence making data deception attacks a real threat in modern control systems. This represents the goal of this thrust, where online learning algorithms will be employed to learn the reactor dynamic behavior. Taking this first step is essential to designing defensive strategies that can anticipate the attackers moves. More importantly, it is to alert the community that defensive methods based on approximate physics models could be bypassed by the attacker who can approximate the models in an online mode during a lie-in-wait period. For illustration, we employ a simplified point kinetics model and show how an attacker, once gaining access to the reactor raw data, i.e., instrumentation readings, can inject small perturbations to learn the reactor dynamic behavior. In our context, this equates to estimating the reactivity feedback coefficients, e.g., Doppler, Xenon poisoning, etc. We employ a non-parametric learning approach that employs alternating conditional estimation in conjunction with discrete Fourier transform and curve fitting techniques to estimate reactivity coefficients. An Iranian model of the Bushehr reactor is employed for demonstration.

2. THRUST 1: SELF-LEARNING OF REACTOR STATE

The basic premise of physical process defenses is that instead of constructing signatures for the incoming attacks, one would be developing signatures for the genuine (i.e., unaltered) operational characteristics of the entity to be protected. This class of defense techniques is sometimes denoted by other researchers as physics-based or model-based defenses. While the premise is reasonable, one must question the effectiveness of such signatures if the attacker has access to approximate models. This thrust will address this concern and will demonstrate whether an attacker armed with technical knowledge about the system can in fact develop a model whose predictions are indistinguishable from the defender's models.

To demonstrate, a point kinetics model for a reactor core is modeled but the model parameters are assumed to be unknown. Inference techniques are employed to learn reactor parameters in an online mode. In particular we use a combination of least-squares, discrete Fourier transform, and nonparametric estimation to build an inference model for the model parameters. A short description of the nonparametric estimation technique employed is given in a separate section. Also a short description of a generic industrial control system layout is given next.

2.1. Industrial Control System

A genetic instrumentation and control (I&C) network can be abstracted into 4 major parts as illustrated in Figure 1: (1) physical response from a physical model, e.g., decrease in neutron flux , or coolant temperature, etc., this response is denoted as p_n; (2) Sensors are employed to detect p_n and generate a response signal, y_n, which is delivered and visualized on the interface between I&C system and practitioners; (3) controllers receive the response y_n from sensors and perform physics or statistics checks or analysis, based on which, give control commands; (4) actuators convert the control commands into physical changes u_{n+1} in the physical system, e.g., pressure changes, movement of control rods, or boron concentration, etc.

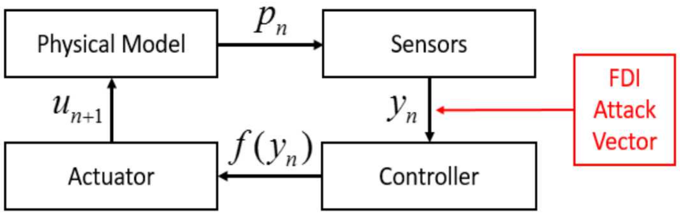

Figure 1. Generic Industrial Control System

2.2. Nonparametric Approach: Alternating Conditional Expectation (ACE) algorithm

Despite the startling growth in computer power, rendering high number of executions using high fidelity simulation is far beyond the reach of foreseen increase in computer power. Thus, construction of simulation emulators is considered as an essential requirement to complete engineering analyses. An emulator, by definition, is any mathematical construct that attempts to capture the main features of the high fidelity simulation. For example, a model of simplified physics assumptions may be considered an emulator, sometimes referred to as low-fidelity model. The lower fidelity may be a result of using simplifications in the physics equations, or due to numerical approximations, such as the use of coarse versus fine mesh to discretize the model equations. These may be referred to as physics-based emulators, as they attempt to retain the physics principles that underpin the behavior of the system. Another class of emulators employ function approximation techniques, wherein a parametric representation is used to approximate the model behavior over the envisaged range of model application. This approach is also referred to as response surface approximation, where an assumed functional form, typically with unknown features such as undetermined coefficients, is fitted against the model behavior at a number of training points. The fitting is done via a minimization search that identifies the surface features that minimize the discrepancy between the model predictions and the assumed surface predictions at the training points. An excellent example of this class of methods is the commonly used least-squares-based polynomial fitting approach. With different surfaces, a wide class of methods have been proposed over the years. Examples include the use of radial basis functions, polynomial chaos expansion, orthogonal polynomials, etc. In the statistics community, this type of function approximation is typically referred to as supervised learning. Another class of methods that have gained a lot of prominence in the data mining community is the so-called unsupervised learning methods, which employ nonparametric methods for emulator construction. Nonparametric methods preclude the need for parametric surface representation. Instead the emulator uses the data directly to make predictions.

The ACE is provided with training datasets, $\{(x_1, x_2, x_3, y)\}$. Then we get transforms of each quantity, $\phi_i(x_i)$ and $\theta(y)$. The algorithm process can be expressed as below[4]:

```
1. Initiate all data, $\theta(y) = y / \|y\|$, $\phi_i(x_i) = 0$, $\|y\| \equiv [E(y)^2]^{1/2}$
2. $e^2(\phi,\theta) = E[\theta(y) - \Sigma\phi_i(x_i)]^2$
3. Iterate until $e^2(\phi,\theta)$ fails to decrease:
4. For $k = 1$ to $p$, do:
     $\phi_k{}'(x_k) = E[\theta(y) - \Sigma_{i \neq k}\phi_i(x_i) | x_k]$;
     Replace $\phi_k(x_k)$ with $\phi_k{}'(x_k)$
   End For Loop;
   End Inner Iteration Loop;
     $\theta'(y) = E[\Sigma_{i=1}^{p}\phi_i(x_i) | y] / \|E[\Sigma_{i=1}^{p}\phi_i(x_i) | y]\|$
     Replace $\theta(y)$ with $\theta'(y)$
   End Outer Iteration Loop;
   $\theta$, $\phi$ are solutions, mentioned as transforms.
5. End ACE Algorithm.
```

2.3. Physical model: Point Kinetics

This study employs a point kinetic reactor model which is described by Equations (1)-(4), for Bushehr reactor, see Ref[5] for a full description of the model. The model is based on four differential equations that follow the evolution of reactor flux or power, precursors concentration, Iodine, and Xenon concentrations. Xenon decays radioactively and is produced from both of fission and the decay of Iodine, which is generated from fission.

$$\frac{dP}{dt} = \frac{\rho_{net} - \beta_{eff}}{\Lambda} P + \lambda_{eff} C \tag{1}$$

$$\frac{dC}{dt} = \frac{\beta_{eff}}{\Lambda} P - \lambda_{eff} C \tag{2}$$

$$\frac{dI}{dt} = \gamma_I \Sigma_F P - \lambda_I I \tag{3}$$

$$\frac{dXe}{dt} = \gamma_{Xe} \Sigma_F P + \lambda_I I - \lambda_{Xe} Xe - \bar{\sigma}_{Xe} XeP \tag{4}$$

In equation (1), ρ_{net} denotes net reactivity, which demonstrates how neutron source and feedback effects working on the system, where $\bar{\sigma}_{Xe}$ is an effective value for Xenon absorption cross section, expressed in equation (6).

$$\rho_{net} = \rho_{ext} - \alpha_P [P(t) - P_0] - \frac{\bar{\sigma}_{Xe}}{\upsilon \Sigma_F} [Xe(t) - Xe_0] \tag{5}$$

$$\bar{\sigma}_{Xe} = \frac{\sigma_{Xe}}{\Sigma_f E_f V} \tag{6}$$

where V is the reactor volume. In addition, the coefficients designed values are listed in Table 1.

Table 1. Point kinetics designed parameters

Symbol	Quality	Value
$P(t)$	core power	P_0 =3000MW
$C(t)$	precursor concentration	
ρ_{ext}	external reactivity injected into the core	
ρ_{net}	net reactivity of the core	
α_P	power coefficient of reactivity (temperature dependent feedback)	0.48×10^{-11} W^{-1}
$I(t)$	Iodine concentration	
$Xe(t)$	Xenon concentration	
β_{eff}	effective delayed neutron fraction	700×10^{-5}
λ_{eff}	effective precursor decay constant	7.841×10^{-2} s^{-1}

Λ	neutron mean generation time in the core	32×10^{-6} s
υ	average neutrons produced by fission	2.45
Σ_f	effective one group cross section for the core	0.77×10^{-2}
γ_I	fission yield for Iodine	6.386×10^{-2}
λ_I	Iodine decay constant	2.875×10^{-5} s^{-1}
γ_{Xe}	fission yield for Xenon	0.228×10^{-2}
λ_{Xe}	Xenon decay constant	2.092×10^{-5} s^{-1}
σ_{Xe}	neutron capture cross section for Xenon	2.7×10^{-18} cm^2
E_f	Energy released per fission	320×10^{-13} J
V	Core volume	27.8 m^3

Since Xenon has an exceedingly large neutron absorption cross section, the neutron flux changes as Xenon concentration varies. With respect of this fact, during a period of reactor operation, a rivalry between the addition and removal rates of Xenon results in a periodic variation of neutron flux, Xenon and Iodine concentration. However, this oscillation is undesirable, since it can induce some difficulties e.g., a dead time for reactor restart, violation thermal limitations at local areas, or hiding attacker's false data injection, etc. In order to model this oscillation behavior accurately, the physical model and related parameters above are necessary. In our approach, these significant parameters can be estimated in a surrogate model, by introduction of perturbations to the reactor.

3. SIMULATION RESULTS

For this study, three parameters are selected to demonstrate the proposed approach, these parameters are the power feedback coefficient α_P , the neutron capture cross section of Xenon σ_{Xe}, and the fission cross section Σ_f. Figures 2 through 4 show results of a parametric study in which each of these parameters is perturbed by a small amount. As evident, these parameter perturbations result in changing oscillations amplitude, phase, and dying speed. To identify the parameters only by a brute force least-squares solver is proved to be inadequate.

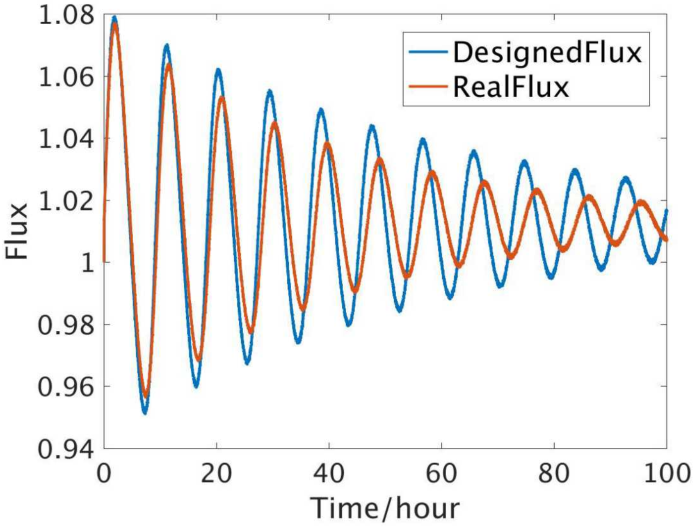

Figure 2. Flux Sensitivity due to Parameter Σf

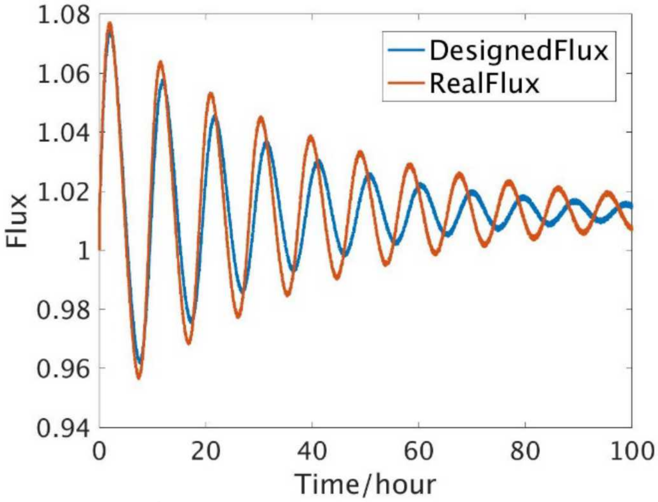

Figure 3. Flux Sensitivity due to Parameter σxe

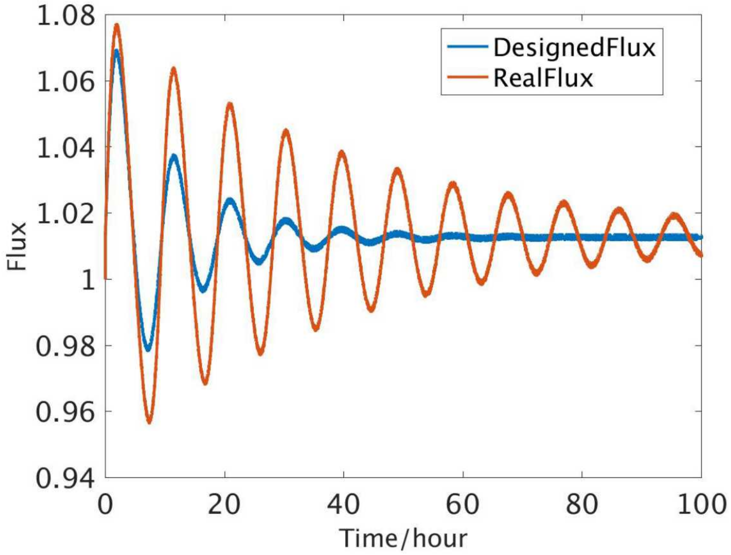

Figure 4. Flux Sensitivity due to Parameter α_p

3.1. Computational Procedures

The following inference approach is employed:

1. Starting with estimates for the k parameters ($k = 3$), generate an estimate for the flux profile by solving Equation (1) through (4). Let N refer to the dimension of the flux profile, i.e., the number of components, one component per time step.
2. Generate M randomized perturbations of the parameters, and calculate the corresponding flux profiles.
3. Calculate fast Fourier transform of the M flux profiles .
4. Using scatter plots and simple variance measures, identify the dominant Fourier coefficients associated with each parameter, where dominance implies strong sensitivity to the input parameters.
5. Combine all identified Fourier coefficients into a K component vector.
6. This reduces the inverse problem to one with k input parameters and K output responses. The goal is to identify the best transfer function relating inputs and outputs.
7. Apply the ACE (Alternating Conditional Expectation) algorithm to help identify the best input-output transfer functions. For this work, given the smoothness of the coefficients variations with the parameter perturbations, a 3rd order polynomial is employed.
8. For a given flux shape, one can update the parameters by first identifying the K Fourier coefficients, and inverting the transfer function in step 7 to determine necessary adjustment for the parameters.
9. With the fitted functions for transforms and inputs, a numerical solver is employed to find a new value for input, given the transform values of responses as well as initial estimation guess.

Figure 5 through 7 show the Fourier coefficients for a number of perturbations where a single parameter is perturbed at a time.

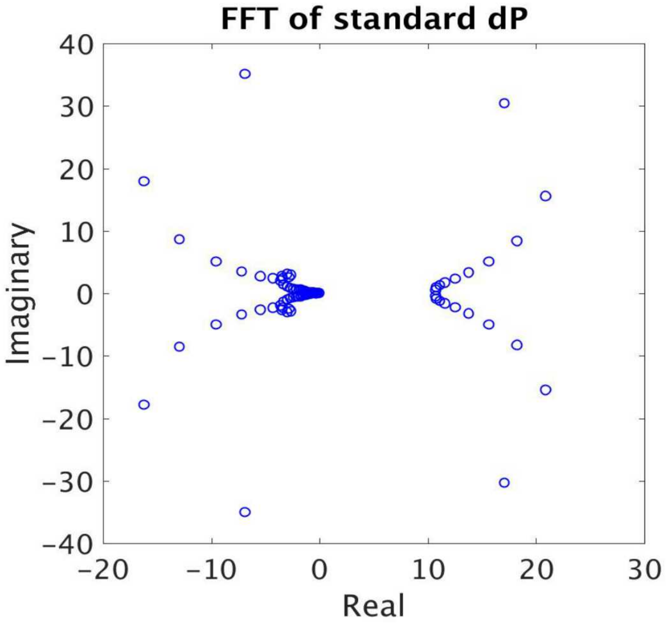

Figure 5. Fourier Coefficients with Perturbed α_p

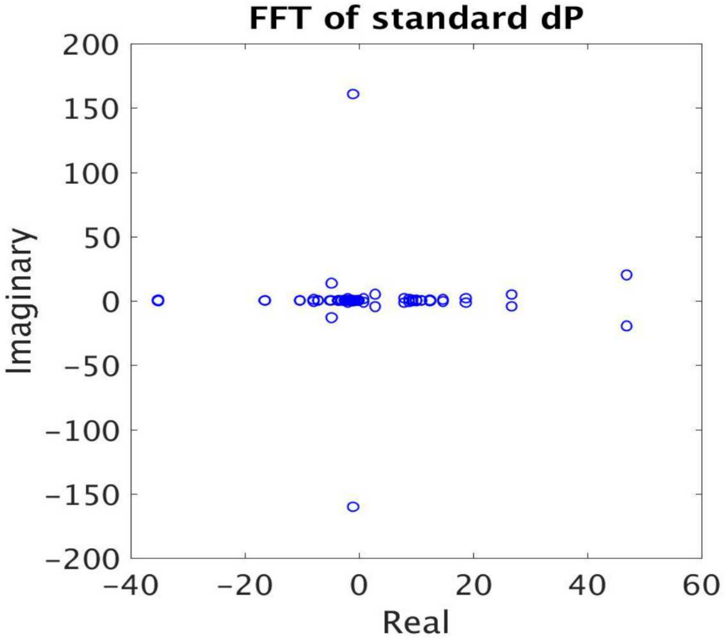

Figure 6. Fourier Coefficients with Perturbed Σ_f

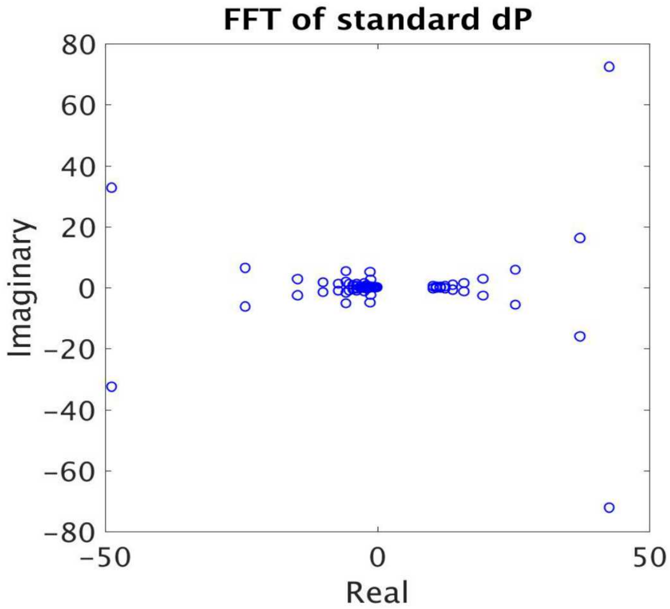

Figure 7. Fourier Coefficients with Perturbed σ_{xe}

Figures 8 through 10 show results of the parametric study of select number of Fourier coefficients versus each of the three parameters.

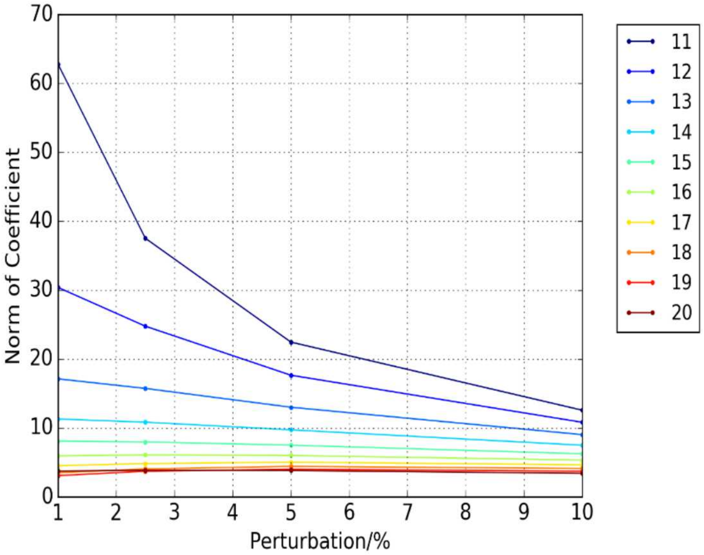

Figure 8. Coefficients Variation with Perturbed α_p

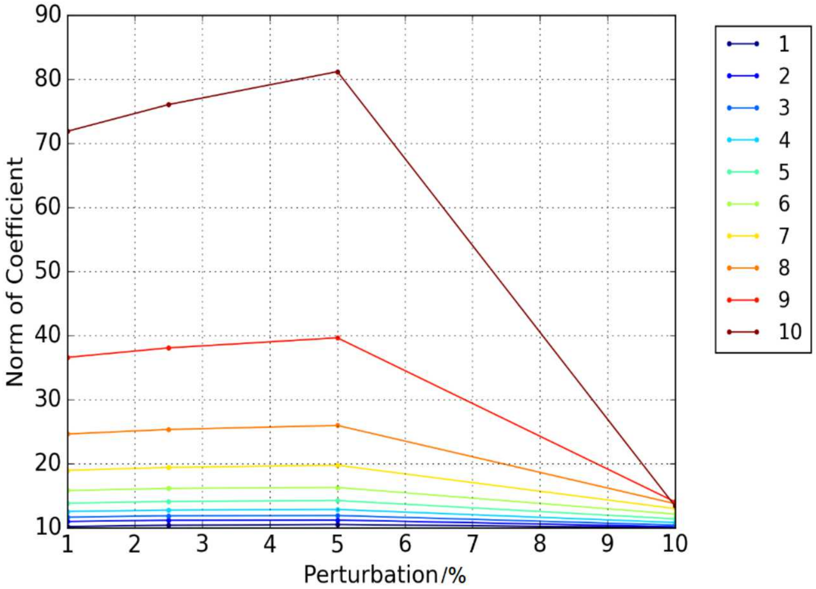

Figure 9. Coefficients Variation with Perturbed Σ_f

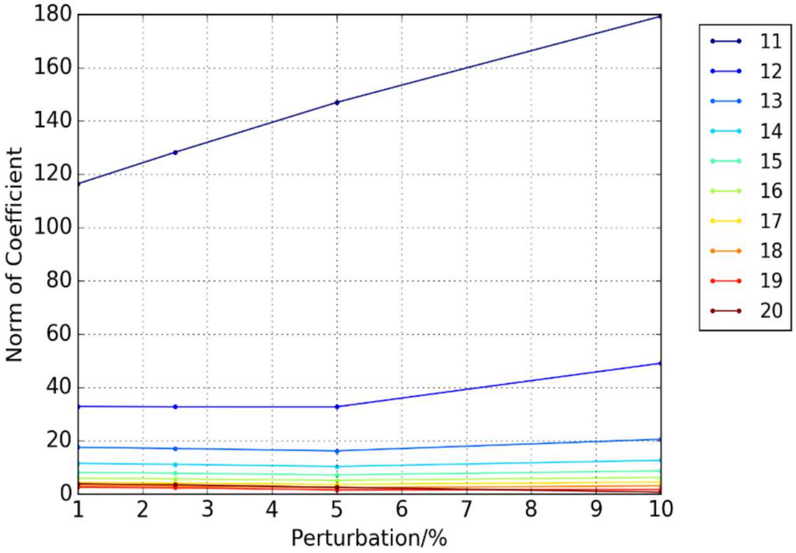

Figure 10. Coefficients Variation with Perturbed σ_{xe}

The figures above show that variations of the different Fourier parameters, from which we select a subset of dominant coefficients, coefficient 10, 11 and 12, as responses to apply ACE algorithm. ACE is a nonparametric method meaning that no prior parametric representation is required to describe the dependence of the Fourier coefficients on the input parameters. Instead, this dependence is learning directly from the data. To evaluate the accuracy of this application of ACE, we divide the sample into 2 parts, one for training the model of ACE, meanwhile the other for testing the accuracy of trained ACE model. Training data are distributed at 960 random points in the 3 dimensional space of the parameters mentioned before. A 3×960 matrix **X** stores the relative values of parameters; and the corresponding responses, are three selected coefficients from fast Fourier transform, which are stored in three 1×960 vectors $\mathbf{Y_1}$, $\mathbf{Y_2}$ and $\mathbf{Y_3}$. Independent from the training data, there are 40 random points in testing data which are stored in a 3×40 input matrix and three 1×40 vectors for storing responses.

We apply the algorithm in based python developed function by Nick Touran[6]. Given training data, this function gives transformations of each perturbed quantity and responses, from which we can identify the general relationship between the quantities and their corresponding transformation, employ interpolation to predict the responses of test cases, and apply curve-fitting to obtain the transformation function afterwards.

The evaluation calculation process is shown in Figure 11, and functions in the process are defined as below. Then we can get the residuals by the difference between y_{new} and y .

$$\begin{gathered} f_i(x_i) = f_{\text{interpolation}}(x_i, \phi_i(x_i)); \\ \phi = \Sigma\phi_i(x_i) \\ h(\phi) = h_{\text{interpolation}}(\phi, \theta); \\ g(\theta) = g_{\text{interpolation}}(\theta, y); \end{gathered}$$

$$\left.\begin{matrix} x_0 \\ x_1 \\ x_2 \\ x_3 \end{matrix}\right\} \xrightarrow{f_i(x_i)} \left.\begin{matrix} \phi_0 \\ \phi_1 \\ \phi_2 \\ \phi_3 \end{matrix}\right\} \to \phi \xrightarrow{h(\phi)} \theta \xrightarrow{g(\theta)} y_{new};$$

Figure 11. Evaluation Calculation Process

Figure 12 shows an example of the nonparametric dependencies discovered by ACE for one of the Fourier coefficients. With the known nonparametric relationship generated from ACE, We employ least square to functionalize the variables and the corresponding transforms. Then we apply data for testing to the fitted formulation, which is shown in Figures 13-15.

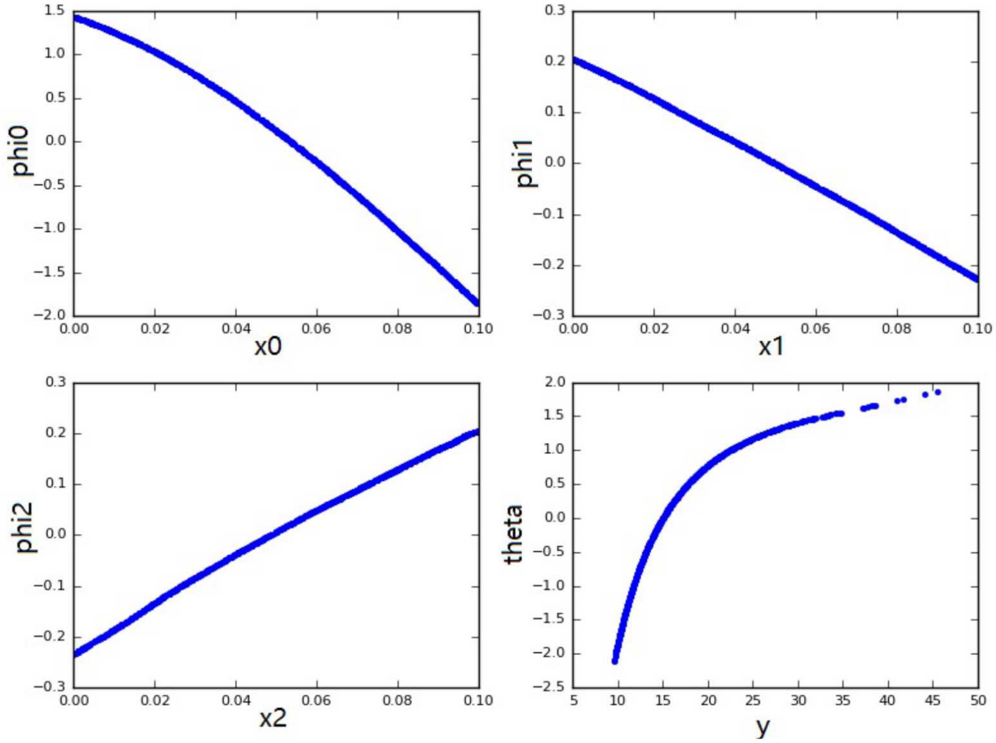

Figure 12. Transformation Plot of Poefficient 10

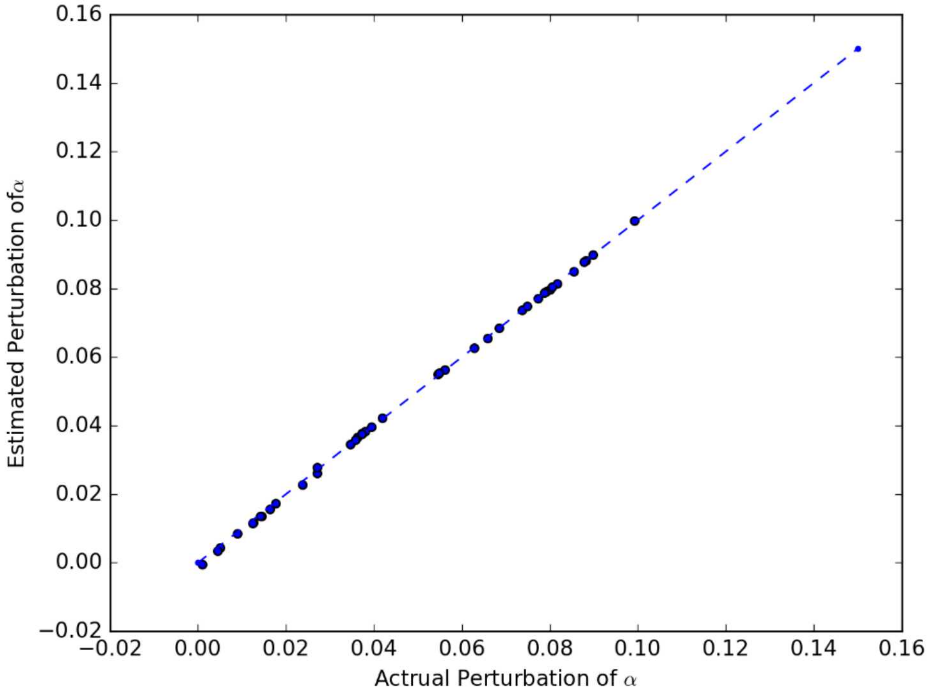

Figure 13. Estimated vs. Actual Perturbation of α_p

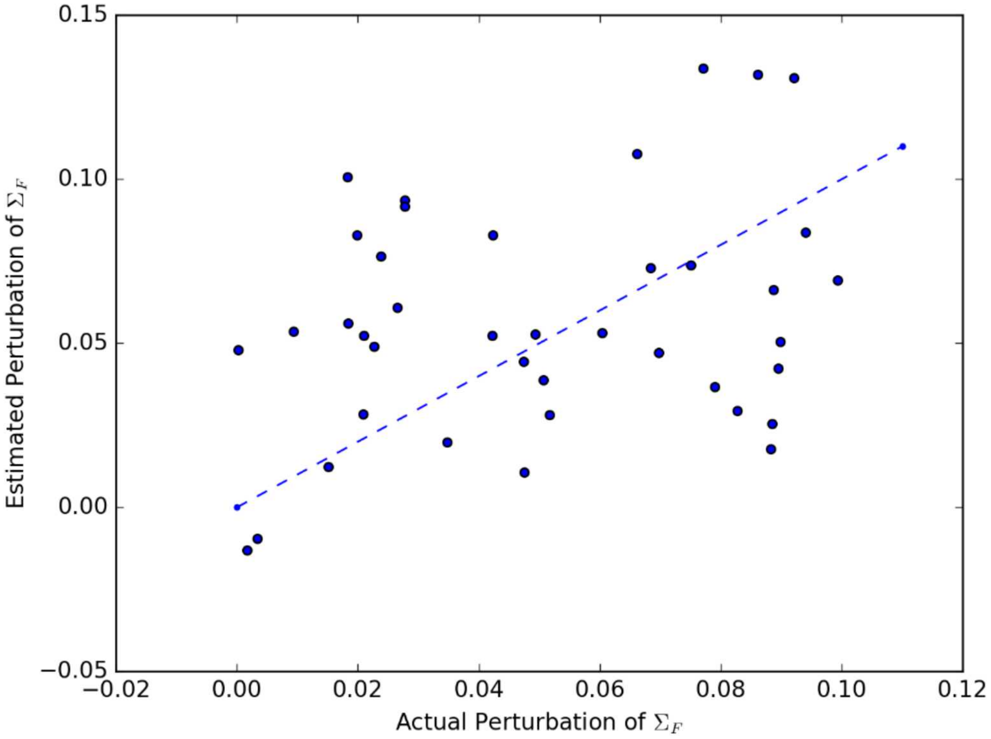

Figure 14. Estimated vs. Actual Perturbation of Σ_f

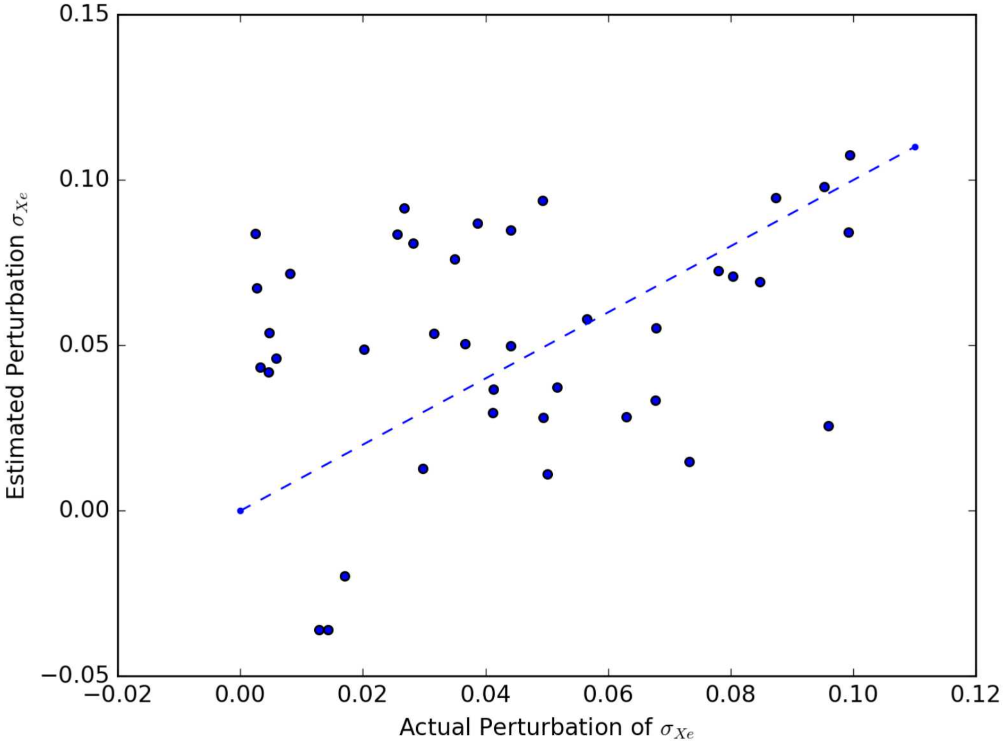

Figure 15. Estimated vs. Actual Perturbation of σ_{xe}

Figure 16 compares the original flux and the flux reconstructed using the estimated parameters, which represents the model to be used by the attackers. Results indicate that the inference approach produces highly accurate estimation of the flux shape, implying that the attacker can later use this model to inject false data into the I&C network while going undetected by the conventional outlier/anomaly detection techniques.

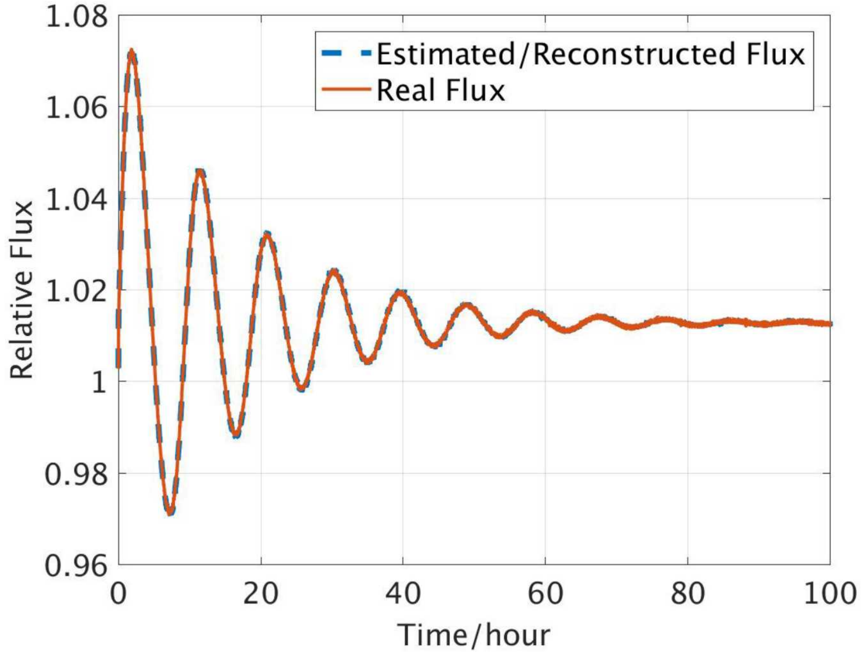

Figure 16. Comparison of Estimated Flux and Real Flux

4. THRUST 2: DATA DRIVEN LEARNING OF REACTOR STATE

This thrust poses the question of whether data driven techniques, not aided by physics models governing reactor behavior, may be used to estimate reactor state in an online mode. To achieve that, we employ the RELAP5 thermal-hydraulics model to design representative behavior for a reactor under a number of operating scenarios to judge the ability of data driven techniques to learn reactor states. Data-driven techniques may be thought of as black-box learning techniques which assumes complete ignorance of the physics models governing the observed behavior. They deal only with the inputs and outputs and attempt to build predictive models. Examples include response surface methodologies and function approximation techniques, which may be applied in either parametric or nonparametric fashions. Parametric techniques assume a surface with a number of unknown coefficients, such as polynomial fitting, and employ a minimization procedure to identify the coefficients that maximize the predictability of the assumed surface, i.e., minimize the discrepancies between measured and predicted responses using the assumed surface. In nonparametric techniques, such as alternating conditional expectation, explained earlier, no surface is assumed, and the best functional shape is determined directly from the data. Both approaches are typically applied after some dimensionality reduction is applied to explore any dependencies between the data to improve the performance of the minimization procedure.

4.1. RELAP5 Model Descriptions

The RELAP5 model simulates a PWR with two primary coolant loops by 139 volumes, 142 junctions, and 83 heat structures. Heat structures are used to represent heat transfer from fuel rods, U-tubes, pressure vessel wall, vessel downcomer wall, core shroud, and internals in the upper head and lower and upper plena. This model prints output every 20 seconds from 0 to 5000 seconds. The nodalization of the model is illustrated in Figure 17[7] and 18[7].

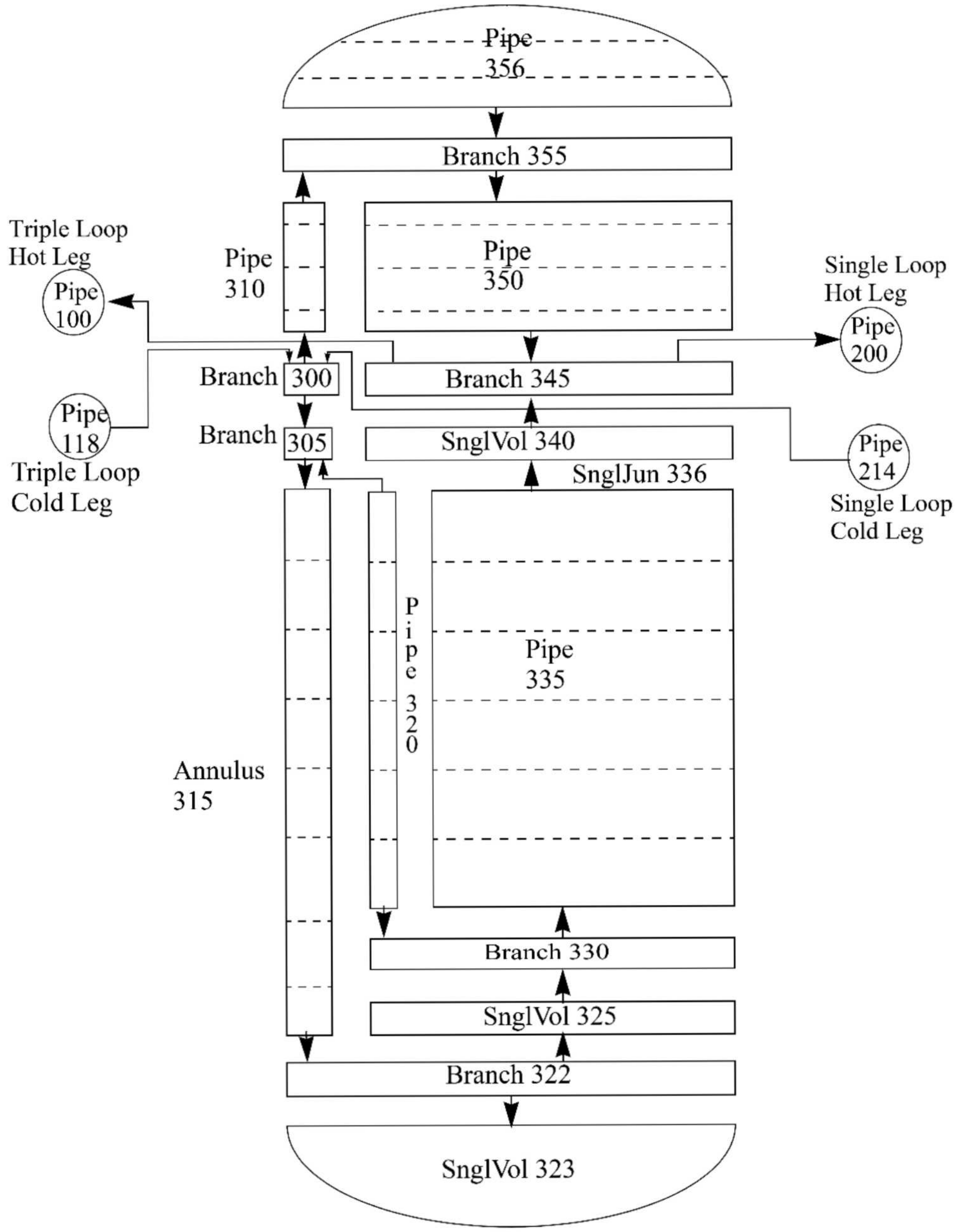

Figure 17. RELAP5 Nodalization: Vessel Model

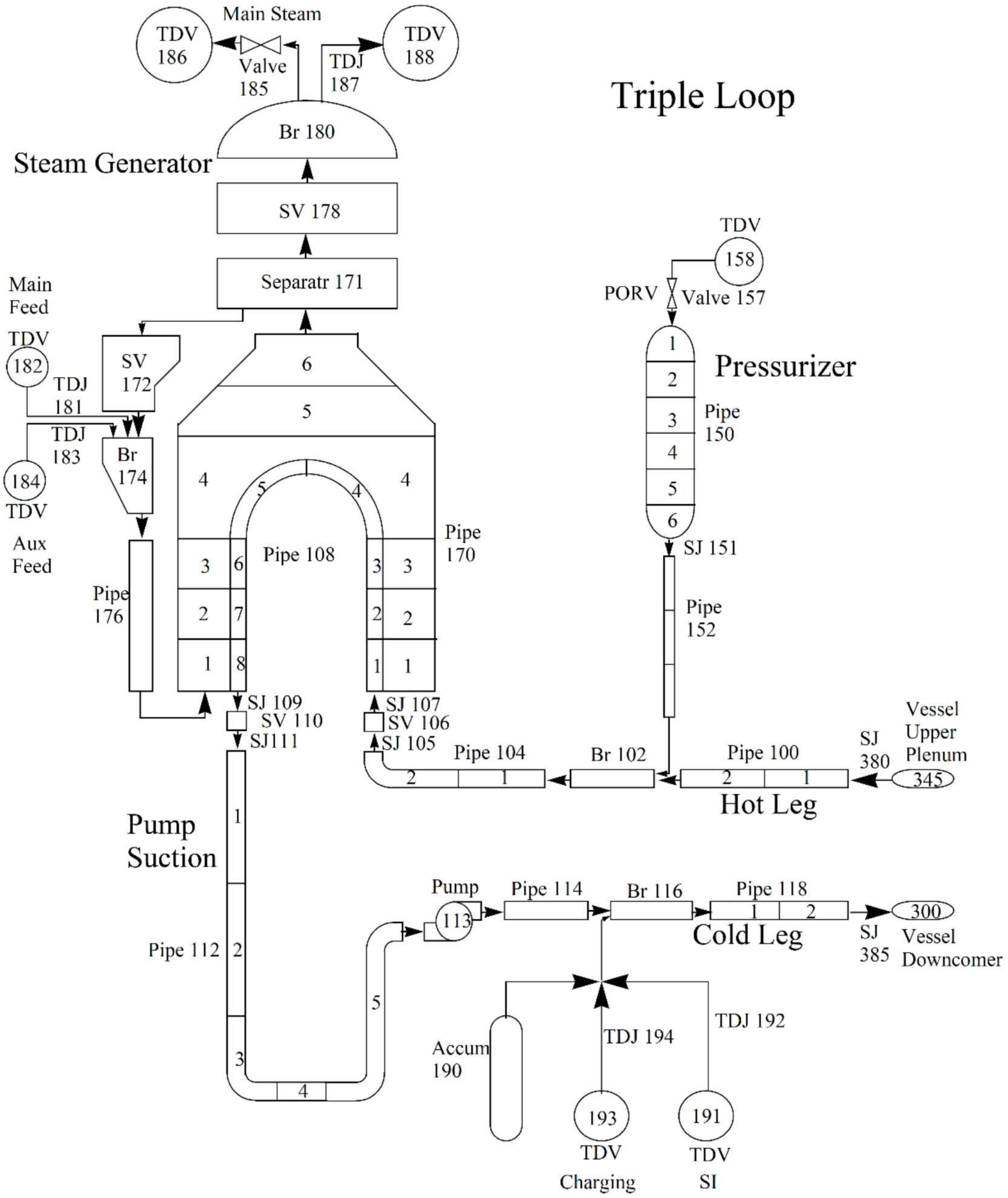

Figure 18. RELAP5 Nodalization for PWR: Loop Model

4.2. Reducibility of RELAP5 Model

Before applying data driven techniques on the physical model, we check the reducibility of this model. We change two control variables to simulate the scenarios which could happen when false data are injected in the control system of nuclear power plant. The changed two variables are steam demand and feed water flow rate in the lumped loop. These two variables are perturbed around their reference values to simulate normal variations during operation. The reference values have been selected to introduce oscillation in the predicted response, which is an undesirable behavior during operation, but selected here to measure the ability of data-driven techniques to predict system behavior when behavior deviates from normal steady state behavior. As responses, 5 physical quantities are selected: pressure in pressurizer, temperature in pressurizer, liquid level of pressurizer, liquid level of steam generator and total power. The simulation results for 9 different cases are stored in a matrix with 45 rows, representing 5 responses per simulation, and the number of columns represent the number of time steps. The goal here is to gauge the level of correlations between the responses over a range of operating conditions. Rank revealing decomposition, e.g., SVD, is applied to determine the effective rank of the resulting matrix. Figure 19 plots the rank as a function of the maximum reconstruction errors. Examples of the reconstructed responses, pressurizer level, power, and steam generator levels, with rank 10 are given in the following figures. Assuming a tolerance of 0.001, that's 0.1% of the nominal values, the rank is higher than the number of responses, indicating that the correlations between the responses over time, which must be captured by the data-driven approach employed. If one employs a weaker criterion, say 1%, the rank drops significantly to about 5-6, which implies one cannot predict one response from knowledge of all other responses. These results are interesting to both attackers and defenders because the noise level will determine the tolerance needed to predict system responses. With higher noise, it becomes easier for the attacker to learn a model that is indistinguishable from the defender's model. To the defender however, this may represent a possibility to develop signatures using the high order correlations which are below the noise level.

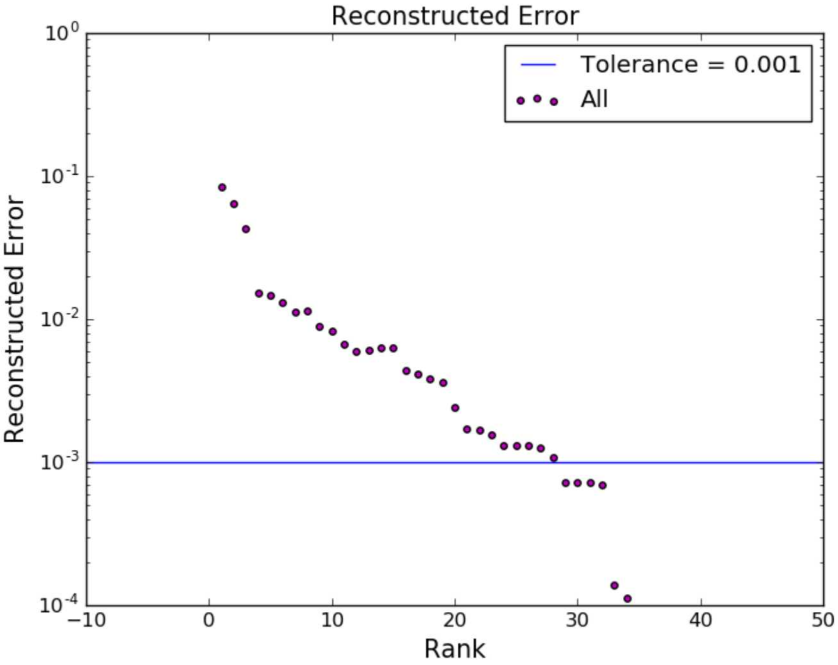

Figure 19. Reconstructed Error From RFA

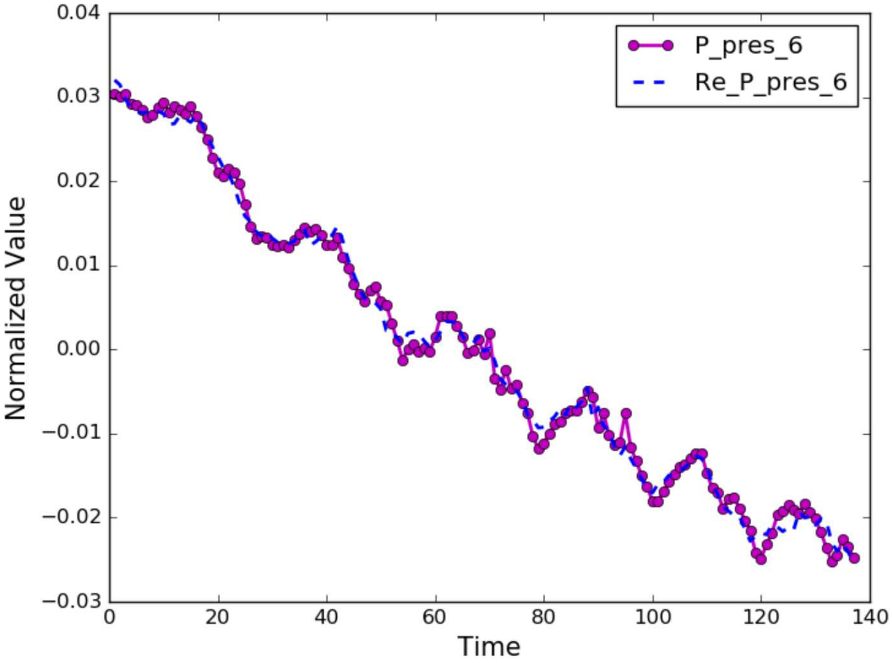

Figure 20. Comparison of Reconstructed and Real Pressure in Pressurizer

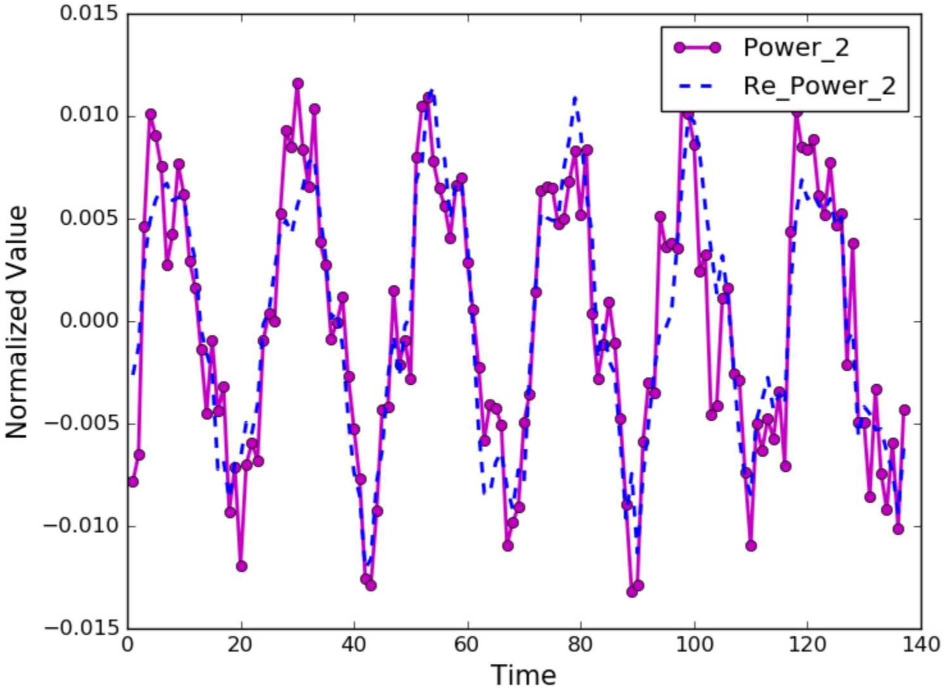

Figure 21. Comparison of Reconstructed and Real Power

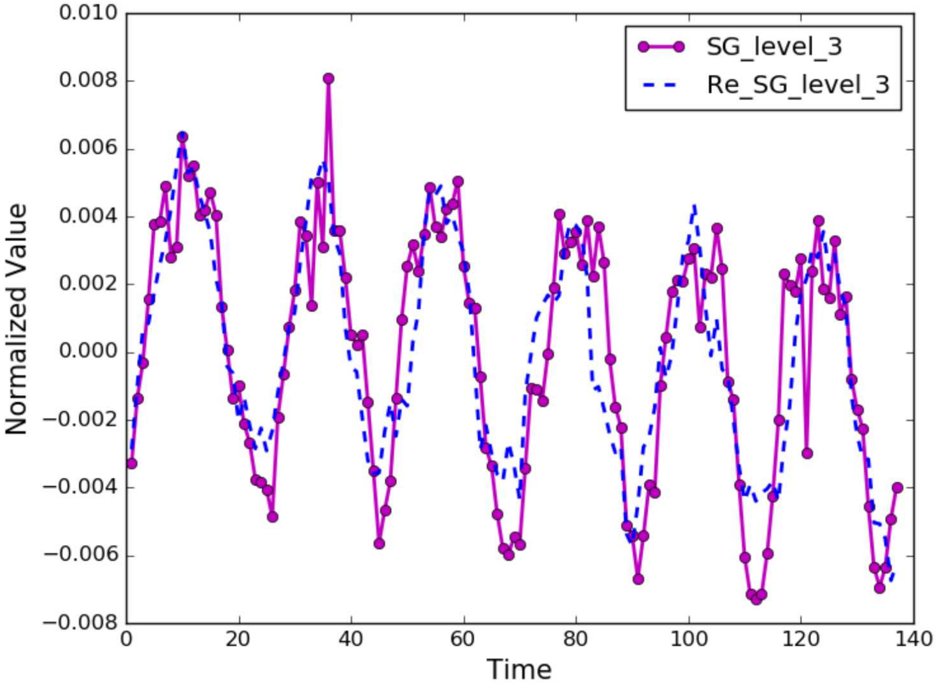

Figure 22. Comparison of Reconstructed and Real Liquid Level of SG

In order to explore the dependence of the responses variations on various conditions, such as inlet feed water flow rate and temperature, one can derive features from the responses variations directly and explore their dependencies on such conditions. For example, a common approach is to calculate the components of the responses variations along some prominent basis, such as Fourier transform basis, or a basis directly derived from the responses variations, such as the singular vectors from the singular value decomposition, or principal components of principal component analysis. For our application, we explore deriving features using both the right singular vectors and the fast Fourier transform. Figures 23 through 25 show the results for the singular vectors, and Figures 26 and 27 for the Fourier transform. Results indicate that while some patterns may exist, it is difficult to use this information to reconstruct the full responses variations. Many attempts have been done to explore this but all results were negative, indicating that the data cannot be reduced, and their dependencies cannot be captured using response surface models, which are not informed by the physics.

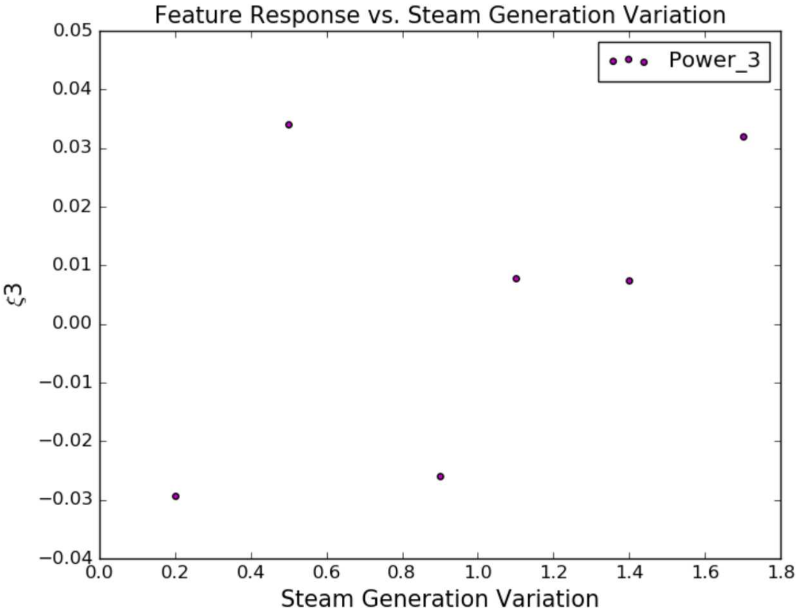

Figure 23. ξ of Power vs. Steam Generation Variation

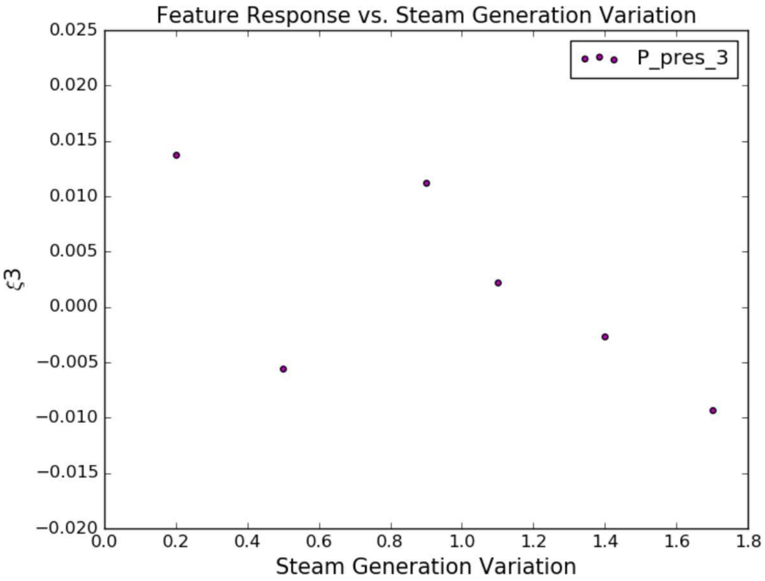

Figure 24. ξ of Pressure in Pressurizer vs. Steam Generation Variation

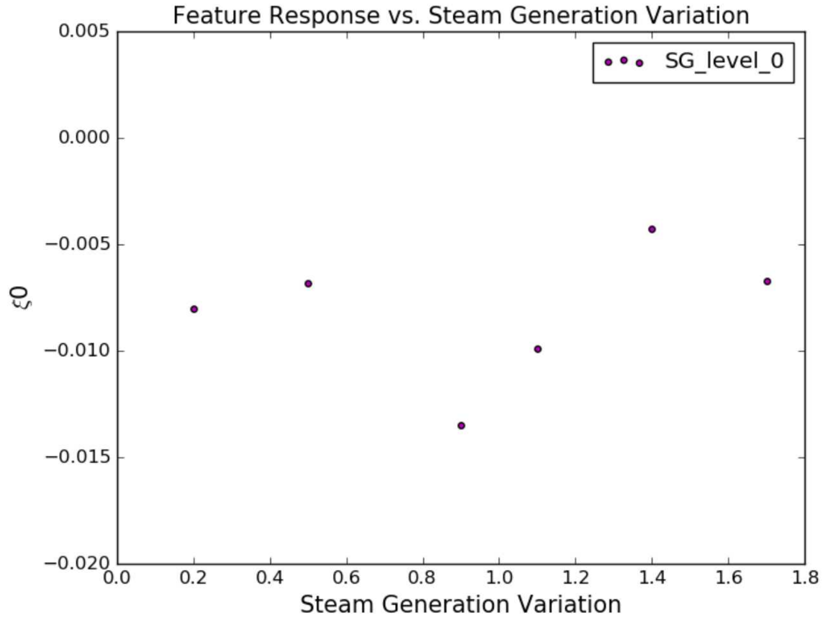

Figure 25. ξ of Liquid Level in SG vs. Steam Generation variation

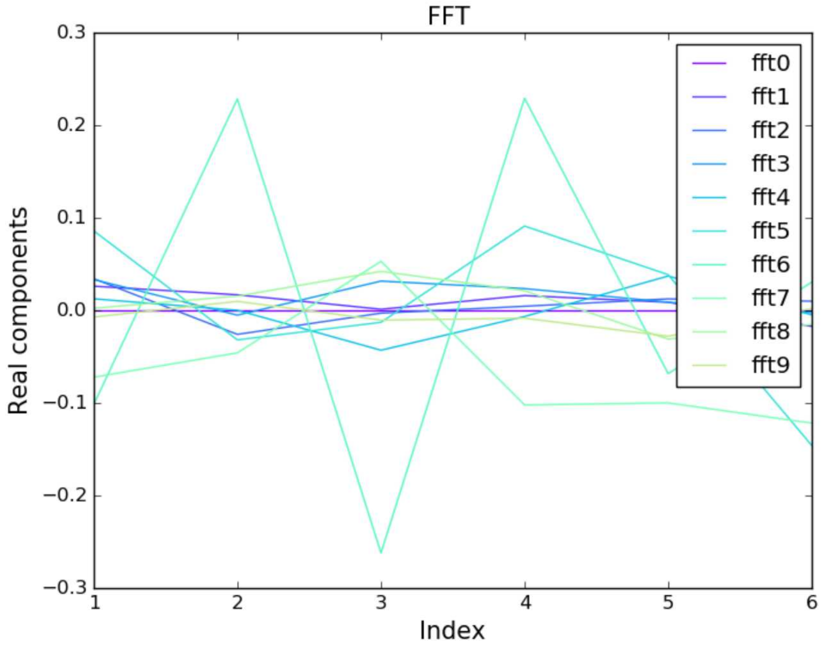

Figure 26. Fast Fourier Transform Coefficients of SG Liquid Level in Steam Generation Variation Cases

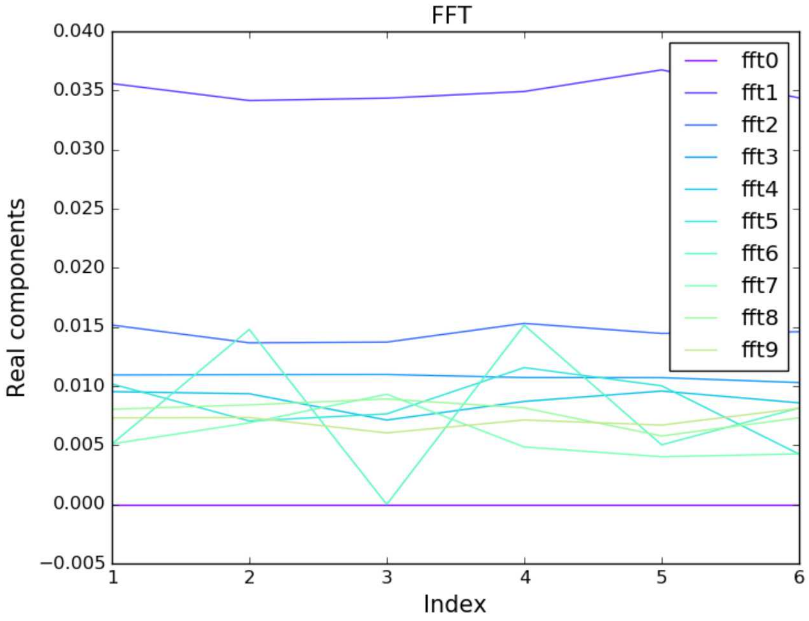

Figure 27. Fast Fourier Transform Coefficients of Temperature in Pressurizer in Steam Generation Variation Cases

5. THRUST 3: RECOMMENDATIONS

In this thrust we provide two types of recommendations on the types of attack scenarios that may be initiated by the attackers, and on the defense strategies needed to identify FDI attacks.

5.1. Recommendations on the Threat Scenarios

Two principal FDI threat scenarios (see Figure 28) may be envisioned, both designed to generate control commands that divert the reactor state from its normal range of operation for malicious purposes. The first threat scenario is designed to initiate an event, which can be done by falsifying sensors readings, and the second threat is to alter the normal (i.e., as-designed) reactor response to an event, which can be done by changing the controller logic.

An example of the first threat is when the attacker attempts to falsify a reduction in the heat removal in a BWR core by injecting false flowrate readings. To go undetected by the plant computer's checks, the attacker changes all other sensors' readings which are correlated with the core flowrate (e.g., neutron detectors signals used to estimate reactivity, outlet core temperature, water and steam levels, etc.). These correlations can be established via conservation principles or determined, as discussed earlier, via self-learning using condition monitoring techniques. Following that, the attacker lets the reactor respond to the event per design procedures (e.g., via actuation of fine motion control rods).

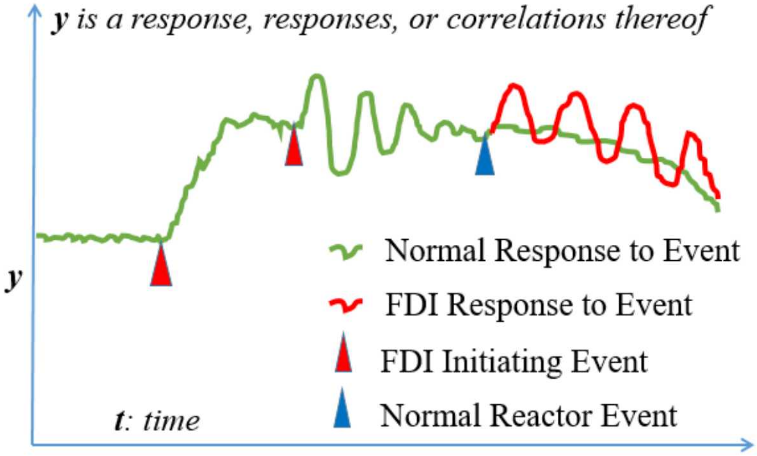

Figure 28. FDI Threat Scenarios

In the second threat, the attacker attempts to alter the normal plant response by changing for example the steam generator set point for trip or by trapping the reactor in an oscillatory regime that would otherwise be damped by the natural reactor response. For example, one could introduce undamped density wave oscillation (DWO) in a BWR. A DWO is a wave of traveling voids that produces a lagging pressure drop. Natural reactor response can dampen these oscillations because an increase in the flow increases pressure drop which reduces the flow thereby killing the oscillatory behavior. The attacker could interfere with this natural behavior by changing the phase shift (via falsifying measured flowrate and all correlated sensors, e.g., OPRMs) until it reaches 180°, where pressure drop feedback becomes positive, which results in unstable oscillations. Left undetected for long time leads to fuel clad failure.

5.2. Recommendation on the Defense Strategy

Based on the findings of this study, we propose a new formulation of the cybersecurity defense problem against false data injection attacks, based on the concept of active monitoring of system behavior. This defense relies on building mathematical signatures of the entity to be protected as functions of its unique historical operational and design characteristics which are expected to serve as fingerprints such that no two entities can be the same. An initial step towards this goal is to focus on measuring the strength of the mathematical signatures developed and what would it take for an attacker to break them, assuming they have access to everything about the system, including system design, raw signals data and the physics models employed by the control system, and the defense algorithms used to build the signatures.

Active monitoring relies on the physical understanding of the engineering system to establish a new measure of security that derives its strength from the unique design and operational characteristics of the reactor system. This approach is essential because the important accidents in a nuclear reactor have a very short response time, making it difficult for operators to enact proper counter measures, especially when their displayed information is falsified by the attack. In active monitoring, the data traffic of the digital control system are actively perturbed based on physics-based understanding of system behavior. These active perturbations represent small inconspicuous distortions to the data traffic that do not impact system behavior but make it nearly impossible for the attacker to go undetected when attempting to learn and/or modify system behavior. The main value of this approach is to allow for early detection of threats during their initial lie-in-wait period where attackers typically excite the system with small perturbations, e.g., commands to actuators of system components, to learn the system's dynamical behavior. Detecting the attack at early stage is very important given the short response time for most important accidents.

To demonstrate the resilience of this approach, the raw sensors data, the computer physics codes used to stimulate the system as well as access to the operators-displayed information is to be assumed exposed to the attacks at varying levels of access up to full-fledged access. In the full-fledged access, the attacker is assumed to have access to the algorithms used to design the data distortions used by the defender's active monitoring system. This is an important assumption in the development of any security approach, because research has proven that security by obscurity is never an effective strategy. For each access level, the attacker success criterion will be based on their ability to devise strategies that learn system behavior and identify engineering vulnerabilities that can be exploited to launch attacks that take the system outside its per-design operational boundaries. Part of the attackers' success will be measured by their ability to deceive operators to stop them from taking corrective measures that minimize the impact of the attack. The success criterion of the defense algorithm is to be measured by several factors including the ability to distinguish between malicious and normal deviations that could potentially lead to accidents, the ability to detect intrusions during their lie-in-wait periods before active attacks are initiated, and the ability to render such attacks harmless, i.e., the ability of the control system and/or the operators to automatically or manually shut down the system without violating any of the safety limits.

The concept of active monitoring is markedly different from passive monitoring which is commonly employed to check equipment reliability and predict early failure. In passive monitoring, the goal is to continuously monitor control system traffic (including both sensors readings and commands to actuators) to determine whether their behavior is consistent with expected variations. Passive monitoring can be ultimately bypassed by technically-able adversaries, as they are expected to possess the system's know-how and can develop computer models that mimic behavior to high degree of accuracy, which can be used to falsify SCADA traffic without alerting operators. Active monitoring, however, introduces small perturbations to SCADA traffic, designed to be small enough as not to impact system performance, but can be leveraged to generate new signatures that are known only to the defender, and thus can be used to detect intrusion. These perturbations can be identified using a number of mathematical techniques, collectively referred to as reduced order modeling (ROM) techniques, which can identify perturbations with negligible impact on system performance. The signatures represent mathematical functions of the all the data comprising the SCADA traffic, including sensors readings and commands variations over space and time, which can be harvested using data mining techniques.

Active monitoring allows for early detection of intrusion that attempts to learn system behavior during an initial lie-in-wait period. For sufficiently complex and stealth attacks, the attackers typically excite the system with small perturbations initially to learn system behavior before launching their attack. These perturbations are selected to have small impact on system behavior and designed to be consistent with normal operational manoeuvers, to avoid detection by operators. Active monitoring will detect these intrusion attempts early on as the attacker's introduced perturbations will not be consistent with those developed by the first module of the active monitoring software.

The data mining algorithms will perform the following functions

1. Identify active perturbations. This module is designed to execute the system's engineering model many times in an off-line mode to search for the optimum perturbations using reduced order modeling techniques. The engineering model is not part of this invention and will be system-dependent. ROM techniques are well-established in the literature, and are typically used to reduced complexity of a given model by identifying perturbations with maximal impact on system behavior. In this invention, ROM is used to search for the perturbations with weak impact on system performance.

2. Identify Signatures. This module will employ conventional data mining techniques to identify in an off-line mode mathematical relationships between the identified

perturbations (generated by the first module) and sensors variations over the combined spatial-temporal phase space.

3. Detect Signatures. This module will be executed in an online mode to compare the signatures identified by the second module to the online SCADA traffic.

The immediate goal of active monitoring is to develop measures by which one can distinguish between genuine behavior, even under accident conditions, and behavior that has been modified by potential adversaries. This monitoring strategy has a wide range of other applications. For example, it may be used as a watermarking tool for a general software predictions. Watermarking is typically used to authenticate authorship of a static representation of a file, e.g., a document, or a picture, but one may think of the need to create watermarking capability for dynamic data, such as videos, or software predictions. In many cases, adversaries may resort to reverse engineering to predict the internal mechanics of a captured component, e.g. drone. This will be based on building approximate models for the system, based on the best available data on system behavior. If one has a capability capable of authenticating the field data before they are used to issue commands to the various system components, one can develop another layer of defense to ensure that captured components are designed to self-destroy if captured by adversaries, thereby protecting the technology. This can be achieved by inserting small perturbations into the field data of the component, that do not impact its performance, but can be used to authenticate the source of the data, i.e., whether produced by an authorized version of the simulation software, or an approximation thereof. We demonstrate the application of such idea using the following model.

An OpenFOAM model for a pipe is employed to calculate the speed of a liquid in a pipe, subject to perturbations in the boundary conditions, the inlet speed from two side openings, see Figure 29. After executing the model for a wide range of conditions, the dominant behavior is evaluated, as captured along the principal components of the velocity field, these are the components of the velocity vector along the left singular vectors, evaluated using singular value decomposition. Small perturbations, referred to as the active perturbations, determined as randomized linear combination of the most dominant directions are added to the velocity vector, such as to render it indistinguishable from the original value. The singular value decomposition of the perturbed velocity field, assumed to be accessible to the attacker, is repeated, and the components along the entire spectrum of singular vectors is compared to the components along the unperturbed singular vectors. Figure 30 shows that one can design such perturbations in a manner that preserves the dominant part of the spectrum, implying that the attackers will not be able to detect that the data have been modified from their original values. The differences in the spectrum start to show later in the spectrum, where the variations are small enough to be within the noise level. Noise-canceling coding schemes, e.g., dirty paper coding, can then be used to insert hidden information along these components to serve authenticate the process used to generate the simulation results. We believe this technology could have a wide range of applications beyond the cybersecurity problem, which will be investigated in future work.

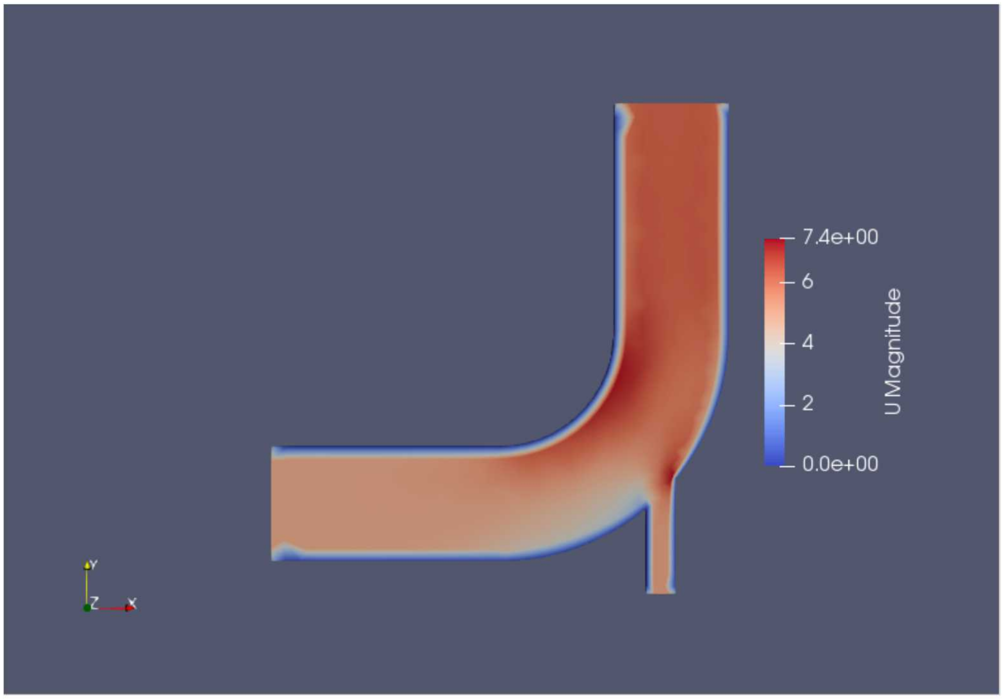

Figure 29. Schematic of Elbow Model in OpenFOAM

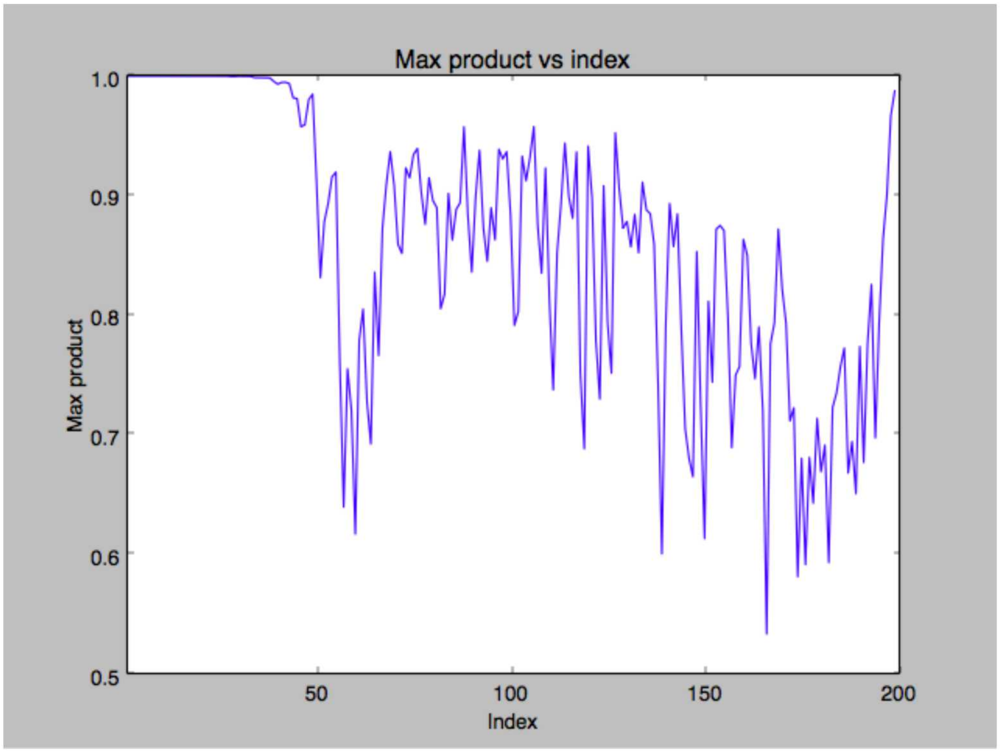

Figure 30. Similarity Between Perturbed and Unperturbed Singular Vector Spectrum

6. CONCLUSION

This work served an important goal, that is to explore the seriousness of intrusive attacks into the SCADA systems of a critical infrastructure such as a nuclear power plant, whose design details and operational characteristics are widely known. The study confirmed that technically-able attackers can indeed learn system behavior if equipped with physics models that approximate system behavior, and further, no reliable learning can be achieved using black-box data-driven techniques, such as neural networks. Data mining technique prove useful in determining unknown features of the physics model, which may include some proprietary data that are not available in the public domain, such as dimensions, compositions, etc. In doing so, the data mining procedure will be carefully guided by physics models that closely approximate system behavior. Again, this is possible, because enough knowledge about these systems is available in the public domain. Moreover, if the attacker does not have access to the physics models used by the defender, we believe that it is possible to develop signatures that employ the higher order correlations between field data. However, we don't believe this is a valid strategy, because the software used to model most critical infrastructures can only be protected using obscurity measures, and typically these types of software are exposed to a large number of code developers, and hence can eventually be acquired by persistent attackers. Hence, passive monitoring techniques are not expected to be resilient enough to build signatures for system genuine behavior. The study provides recommendations on the types of attacks that may be initiated by attackers once have gained access to the SCADA field data, and proposes defense strategies based on active monitoring techniques. In active monitoring, small perturbations are actively inserted in the SCADA field data, in a manner similar to what the attacker does, but for the purpose of authenticating whether the SCADA field data have been manipulated.

ACKNOWLEDGEMENT:

The research team would like to acknowledge the additional source of funding received from DOE office of nuclear energy university program in FY2018; award number DE-NE0008705. The team would like to thank Dr. Katrina Groth, then of Sandia National Laboratory, for her contributions during the first year of the project and for participating in winning another grant from DOE to further expand the scope of the current project. The team would also like to thank Professor Said Abdel-Khalik of Georgia Tech for his contribution on the design of threat scenarios, and Dr. Lon Dawson for his comments and feedback during the second year of the project.

REFERENCES

1. H. M. Hashemian, *Maintenance of Process Instrumentation in Nuclear Power Plants*, Springer, 2006 .
2. P. Fantoni and A.Mazzola, *Applications Of Autoassociative Neural Networksfor Signal Validation In Accident Management*, vol. 37, no. 2, pp. 1040–1047, 1990.
3. P. Fantoni, *Experiences And Applications Of Peano For Online Monitoring In Power Plants*, Progress in Nuclear Energy, Vol. 46, No. 2-3, pp. 206-225, 2005.
4. L. Breiman and J. H. Friedman, *Estimating Optimal Transformations Multiple Regression and Correlation*, Journal of the American Statistical Association September 1985, vol. 80, no. 391, Theory and Method.
5. M. Zarei, R. Ghaderi, and A. Minuchehr, *Progress in Nuclear Energy Space independent xenon oscillations control in VVER reactor : A bifurcation analysis approach*, Progress in Nuclear Energy, vol. 88, pp. 19–27, 2016.
6. N. W. Touran, https://github.com/partofthething/ace, 2012.
7. *RELAP5 / Mod3 . 3 Code Manual Volume III: Developmental Assessment*, Assessment, vol. III, 2010.

Email—Internal

Name	Org.	Sandia Email Address
Technical Library	9536	libref@sandia.gov

Sandia National Laboratories