INL/RPT-22-69871

Light Water Reactor Sustainability Program

A Novel Data Obfuscation Method to Share Nuclear Data for Machine Learning Application

November 2022

U.S. Department of Energy

Office of Nuclear Energy

INL/RPT-22-69871

A Novel Data Obfuscation Method to Share Nuclear Data for Machine Learning Application

Ahmad Al Rashdan (Principial Investigator)[1]
Arvind Sundaram[2]
Tyler J Lewis[2]
Hany Abdel-Khalik[2]

[1]Idaho National Laboratory
[2] Purdue University

November 2022

Prepared for the
U.S. Department of Energy
Office of Nuclear Energy

ABSTRACT

The nuclear power industry is data-rich and these data introduce a tremendous potential for automation and cost savings. On the other hand, research organizations, among other stakeholders, have very capable methods and solutions often developed using simulated or synthetic data due to the lack of real data. One cause of this disconnect is data privacy. Data privacy is of paramount importance in all industries, but especially the nuclear industry due to the risks associated with the malicious use of data (e.g., loss of competitive edge, reverse-engineering of proprietary systems, national security concerns). However, for the data to be usable by research organizations, its inference characteristics need to be maintained. This challenge motivated the data obfuscation method called deceptive infusion of data (DIOD). DIOD is a novel data-masking paradigm that specifically addresses the above concerns with existing data-masking techniques. Fundamentally, DIOD ensures that the information content of the masked and the proprietary data are identical through the information-theoretic guarantee of mutual information, while also disassociating the identity of the masked data from the proprietary system. This one-way (i.e., masking of data) operation is irreversible and allows the analyst to arrive at identical conclusions using the masked data without permitting successful reverse-engineering.

In this effort, DIOD is applied and demonstrated using two use cases for regression. One use case targeted a physics-based model generated from a simple, noise-free point-kinetics (PK) model with one delayed neutron group; the second targeted a process that resembles an actual nuclear power plant process. The first use case was applied to three scenarios in which power was used to predict the PK parameters. Those parameters were “well-posed,” “ill-posed,” and “reduced ill-posed.” All three scenarios were concealed by electrical load data. The results validated that the DIOD procedure preserves mutual information between the original and masked data. The second use case used a red team-blue team exercise where the blue team created process data from a simulation with anomalies included. The blue team masked the data with another process data set using DIOD and shared it with the red team. The red team attempted to identify anomalies in the masked data, and to reverse-engineer the masked data to decipher the identity of the proprietary system. The anomalies were discoverable, but the identity of the system was not revealed, indicating a successful demonstration of DIOD use.

ACKNOWLEDGEMENTS

The authors wish to thank the Light Water Reactor Sustainability (LWRS) program for funding this effort.

CONTENTS

ABSTRACT iii

ACKNOWLEDGEMENTS v

ACRONYMS x

1. INTRODUCTION 1

2. METHOD 4

3. USE CASES 6
   - 3.1 Physics Use Case 6
     - 3.1.1 Description of Physics Use Case 6
     - 3.1.2 DIOD Implementation in Physics Use Case 10
   - 3.2 Process Use Case 16
     - 3.2.1 Description of Process Use Case 16
     - 3.2.2 System Anomalies in Process Use Case 18
     - 3.2.3 DIOD Implementation in Process Use Case 19

4. CONCLUSION 29

5. REFERENCES 30

FIGURES

Figure 1. A data warehouse enables experience and methods to be transferred among NPPs and outside organizations, allowing the establishment of industry-wide solutions. 2

Figure 2. Data obfuscation is a potential solution for enabling data sharing among plants and research organizations. 2

Figure 3. Sample power ratio in time. 7

Figure 4. Original vs. predicted parameters given original data in the well-posed case. 8

Figure 5. Original vs. predicted parameters given original data in the ill-posed case. 9

Figure 6. Original vs. predicted parameters given original data in the reduced ill-posed case. 10

Figure 7. ERCOT hourly electrical load. 11

Figure 8. Magnified ERCOT hourly electrical load. 11

Figure 9. Sample *P* data, masked with DIOD. 12

Figure 10. Original vs. predicted parameters given masked data in the well-posed case. 13

Figure 11. Original vs. predicted parameters given masked data in the ill-posed case. 14

Figure 12. Original vs. recovered parameters given masked data in the reduced ill-posed case. 15

Figure 13. A simple process system simulating water pumped into a draining reservoir. 16

Figure 14. Volumetric flow through the downstream valve. 17

Figure 15. Power of the flow downstream from the reservoir. 18
Figure 16. Enthalpy flow from reservoir with anomalies highlighted. 19
Figure 17. A simple process system simulating water flowing through a series of heated pipes [14]. 20
Figure 18. Enthalpy flow through Pipe 6 in Figure 17. 21
Figure 19. Temperature of fluid of Pipe 8 in Figure 17. 21
Figure 20. Masked power of flow downstream the reservoir. 22
Figure 21. Response $y3D$ showing anomalous regions. 23
Figure 22. Response $y7D$. 24
Figure 23. Response $y3D$ / $y4D$ depicting anomalous regions. 25
Figure 24. Response $y3D$ / $y7D$ depicting anomalous regions. 25
Figure 25. Response $y3D$ / $y7D$ depicting features of the anomalous region in the interval (5900,6400). 26
Figure 26. Response $y3D$ / $y7D$ depicting features of a potential anomalous region. 26
Figure 27. Fifth left singular vector depicting the anomalous regions identified earlier. 27

TABLES

Table 1. Parameter values for generated PK data. 6
Table 2. Mutual information between input and output parameters. 15
Table 3. Characteristics of each anomalous valve opening. 19

ACRONYMS

AI/ML	artificial intelligence and machine learning
DIOD	deceptive infusion of data
ERCOT	Electric Reliability Council of Texas
LWRS	Light Water Reactor Sustainability
NPP	nuclear power plant
PK	point kinetics
PI	proportional-integral

A NOVEL DATA OBFUSCATION METHOD TO SHARE NUCLEAR DATA FOR MACHINE LEARNING APPLICATION

1. INTRODUCTION

The nuclear power industry has been operating for decades and storing a huge amount of useful data for automation of various activities at nuclear power plants (NPPs). On the other hand, research organizations, among other stakeholders, have very capable methods and solutions often developed using simulated or synthetic data due to the lack of real data. This gap motivated the development of a data warehouse to serve as a data hub for NPPs, vendors, research organizations (e.g., universities), standards committees and professional societies, compliance organizations, and various other stakeholders (Figure 1). The data warehouse will host a library of proven methods for NPPs to use and easily access (Figure 2). This would enable NPPs to validate the developed methods on the data warehouse's servers, then download the methods' algorithms and code for local use. The data warehouse would also enable data sharing, including allowing the data to be transferred to a user once the NPP has issued specific authorization or the data have been sanitized (i.e., obfuscated).

Another benefit of the data warehouse is that it enables data integration. Given artificial intelligence and machine learning (AI/ML) has proven very useful to the nuclear power industry, data integration is considered a key enabler for AI/ML. The potential to integrate and leverage data across the entire industry via a data warehouse would afford multiple benefits to AI/ML development, including increased statistical power, higher frequencies of low base-rate behaviors, as well as enhanced verification.

The specific data privacy challenge associated with data sharing for integration in nuclear energy required research into methods of data obfuscation. Data privacy is of paramount importance in all industries, but especially the nuclear industry due to the risks associated with the malicious use of data (e.g., loss of competitive edge, reverse-engineering of proprietary systems, national security concerns). The concern of data privacy has been investigated for decades, starting with data-masking techniques such as substitution, shuffling, encryption, etc., for data warehouses, and more recently, differential privacy and fully homomorphic encryption [1–3].

The methods suited for data warehouses are generally not applicable to industrial data analysis and/or do not preserve the physical correlations necessary for AI/ML tools to be effective. For example, omitting all but the last four digits of social security numbers is not applicable to time-series data from the sensor of an industrial control system. Traditional encryption with decryption keys is intended to protect the data in transit to an analyst; however, it does not protect the data from the analysts themselves, instead relying on administrative red tape, non-disclosure agreements, etc., to prevent the analyst from reverse-engineering the data and publicizing the findings. Homomorphic encryption, while promising and allowing for the mathematical manipulation of data directly in the encrypted form, is in its infancy and is limited in application, typically reduced to multiplication and addition operations in a constrained analytical environment [2]. Furthermore, the massive overheads in encryption render it unscalable to the size of process data commonly encountered in industry. Last, differential privacy relies on the privacy-utility tradeoff by injecting artificial noise (typically Laplacian) into the data collected to provide plausible deniability to the source while preserving group statistics [3]. However, the effect of the injected noise on industrial data is typically detrimental to AI/ML algorithms as it degrades the quality of the data, and injecting vast amounts of noise to obscure trends and patterns renders the data unusable.

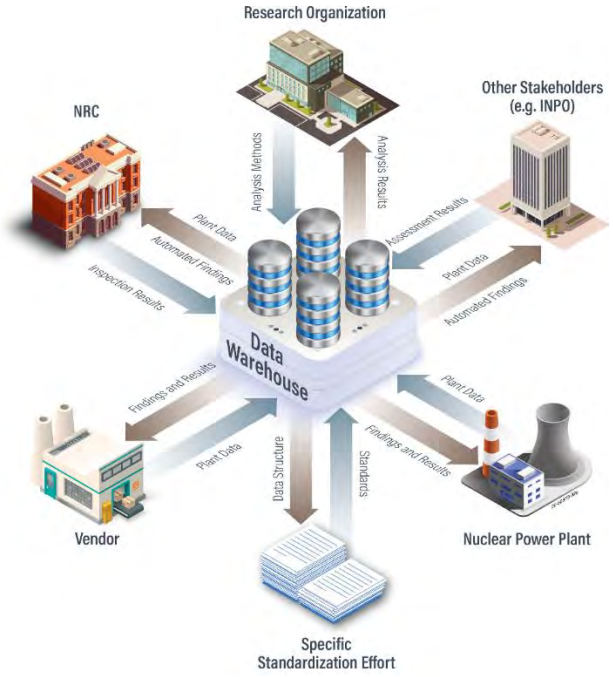


Figure 1. A data warehouse enables experience and methods to be transferred among NPPs and outside organizations, allowing the establishment of industry-wide solutions.

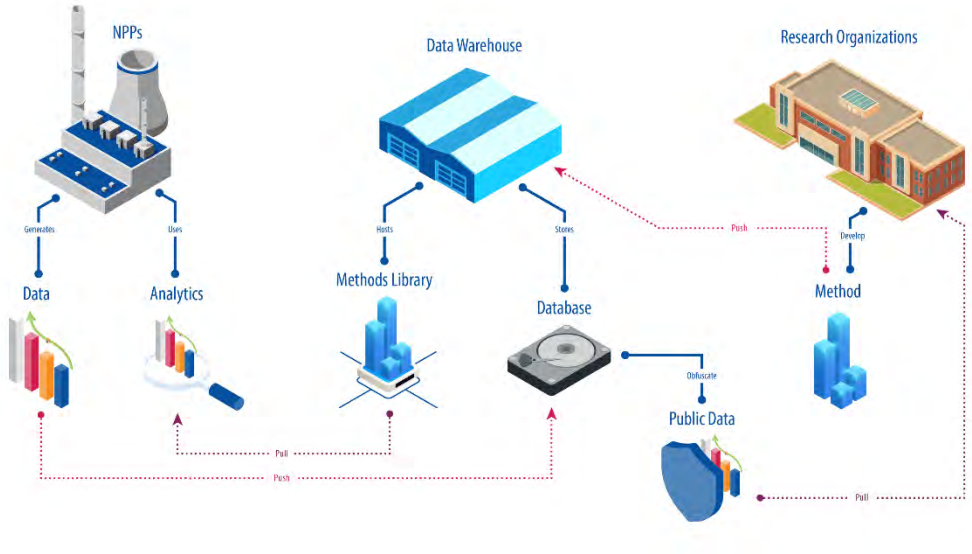


Figure 2. Data obfuscation is a potential solution for enabling data sharing among plants and research organizations.

In this effort, a new data obfuscation method is created. Deceptive infusion of data (DIOD) is a novel data-masking paradigm which specifically addresses the above concerns with existing data-masking techniques. Fundamentally, DIOD ensures that the information content of the masked and the proprietary data are identical through the information-theoretic guarantee of mutual information, while also disassociating the identity of the masked data from the proprietary system. Mutual information is employed to validate the claim of identical inference here. This one-way (i.e., masking of data) operation

is irreversible and allows the analyst to arrive at identical conclusions using the masked data without permitting successful reverse-engineering. Also, DIOD addresses the computational burden on data-rich industrial systems by introducing a highly scalable implementation after an initial one-time reduced-order modeling cost that is typically performed by domain experts for most industrial systems.

The remainder of the report is organized as follows: Section 2 presents the DIOD method. Section 3 consists of two use cases to demonstrate DIOD for regression. One use case targeted a physics-based model, the second targeted a process that resembles an actual NPP process. The preservation of inference properties using DIOD are demonstrated in both use cases. The second use case outlines a red team-blue team exercise where the blue team is composed of experimentalists generating proprietary data with anomalies from a proprietary system and performing the DIOD procedure, while the red team is composed of analysts that are tasked with identifying anomalies in the masked data and attempting to reverse-engineer the masked data to decipher the identity of the proprietary system.

2. METHOD

DIOD decomposes the target data into two sets of metadata: fundamental metadata, the metadata relating to the proprietary system identity; and inference metadata, the metadata relevant for AI/ML applications. The fundamental metadata are typically composed of the underlying differential equations, system geometry, material properties, etc., that are fixed across a set of experiments, whereas the inference metadata are composed of the operational regimes, varying parameters of interest, and temporal and/or sensor correlations that depend on the target AI/ML application. This decomposition is given by:

$$y^P(x,\alpha) \approx \sum_{i=1}^{r} \psi_i^P(x)\phi_i^P(\alpha) \tag{1}$$

Here, $y^P(x,\alpha)$ is the proprietary system data approximated as the sum of r dyads, where $\psi_i^P(x)$ is the proprietary system fundamental metadata corresponding to a parameter x and $\phi_i^P(\alpha)$ is the proprietary system inference metadata corresponding to some process parameter α.

The goal of the DIOD methodology is to replace the fundamental metadata of the proprietary system with that of the fundamental metadata of another generic system, $\psi_i^G(x')$, to disassociate the identity of the data from the proprietary system, while preserving the inference metadata. This is achieved by decomposing the generic system to extract the generic system fundamental metadata $\psi_i^G(x')$:

$$y^G(x',\alpha') \approx \sum_{i=1}^{r} \psi_i^G(x')\phi_i^G(\alpha') \tag{2}$$

and using it through the so-called concealment kernel $k(x',x)$:

$$k(x',x) = \sum_{i=1}^{r} \psi_i^G(x')\psi_i^{P*}(x) \tag{3}$$

$$y^D(x',\alpha) = k(x',x) * y^P(x,\alpha) \triangleq \sum_{i=1}^{r} \psi_i^G(x')\psi_i^{P*}(x)\psi_i^P(x)\phi_i^P(\alpha) = \sum_{i=1}^{r} \psi_i^G(x')\phi_i^P(\alpha) \tag{4}$$

ψ_i^* is the conjugate of ψ_i and is used to eliminate the dependence on x in y^{D} (i.e., remove the fundamental metadata). Given that the masked data y^{D} only possess the fundamental metadata of the generic system and the transformed inference metadata of the proprietary system, it is impossible to guess the identity of the source since infinite possibilities exist. For instance, if a proprietary first-order system of equations (such as a point-kinetics [PK] model) is transformed into a generic system of equations (spring-mass-damper model), any reverse-engineering efforts would only inform the adversary of the spring-mass-damper model, providing no clues to the first-order or the stiff nature of the simple PK model. Furthermore, the invariance of mutual information to invertible transformations and extraneous metadata implies that transformations on the inference metadata and discarding of irrelevant inference metadata is possible to further fine-tune the masking procedure to the target AI/ML application. In summary, Eq. 1–4 show that any reverse-engineering efforts to identify the system are expected to lead to the generic system fundamental metadata. However, any inference efforts will have identical performance on both the proprietary and the masked data since they carry the same information content.

Additionally, Eq. 1–4 are highly scalable requiring an initial one-time cost to develop a library of concealment operators corresponding to the fundamental metadata of various proprietary and generic systems. Multiple data sets carrying the same information content may then be generated through repeated applications of Eq. 4, fusing the inference metadata (or transformations of it) with the

fundamental metadata of multiple generic systems. This effectively creates a benchmark data set where the masked data consist of the same information but appear to have come from various systems. In theory, an ideal AI/ML algorithm is expected to perform identically on all the sets of data due to the identical information content. A detailed discussion of the method can be found in [4].

3. USE CASES

This section presents two use cases to demonstrate the use of DIOD to regression problems, which are commonly used for time-series type of data. In the first use case, PK equations are used as a physics-based experiment in which the data need to be obfuscated. In the second use case, process data from a simulated process that represent real nuclear power plant data are used. In the first use case, the obfuscation is performed using concealment operators that are from a totally different system. In the second use case, another process (i.e., similar type of data) is used to obfuscate the original process. In both cases, the aim is to establish a relationship between the observations of a system or process and its input parameters, assuming the data owner is reluctant to share the observations directly for fears of misuse. To circumvent this issue, the owner of the data provides the masked version of the data using the DIOD procedure, and the true relationship can be identified by reversing the operations known only to the owner.

The process use case utilizes a red team-blue team setup wherein the blue team generates the sensitive and generic data, injects anomalies into the proprietary system, and performs the DIOD procedure. The masked data are then handed to the red team, which is tasked with detecting the various anomalies while simultaneously attempting to reverse-engineer the masked data (i.e., recover the identity of the proprietary system and potentially the sensitive data themselves). It is assumed that the red team is aware of the DIOD procedure and its mathematical framework for the target AI/ML application without knowing the specific transformations used.

3.1 Physics Use Case

3.1.1 Description of Physics Use Case

For this experiment, the proprietary system data are generated from a simplified, noise-free PK model with one delayed neutron group [5, 6] as shown below in Eq. 5–6.

$$\frac{dP}{dt} = \frac{\rho - \beta}{\Lambda} P(t) + \lambda C(t) \tag{5}$$

$$\frac{dC}{dt} = \frac{\beta}{\Lambda} P(t) - \lambda C(t) \tag{6}$$

Here, ρ is the initial reactivity inserted into the system, β is the total fraction of delayed neutrons, Λ is the mean lifetime of prompt neutrons, λ is the one-group average half-life of neutron precursors, and each sample of P is aggregated to form the proprietary system data. For this experiment, each parameter is sampled from a uniformly random distribution with mean values shown in Table 1, and an uncertainty of 10% for each parameter. A sample power profile is generated below in Figure 3.

Table 1. Parameter values for generated PK data.

Parameter	$\boldsymbol{\rho}$	$\boldsymbol{\beta}$	$\boldsymbol{\Lambda}$ **(seconds)**	$\boldsymbol{\lambda}$ **(seconds)**
Mean Value	0.0005	0.0065	$1 * 10^{-5}$	0.08
Uncertainty	$\pm$0 .00005	$\pm$ 0.00065	$\pm\ 10^{-6}$	$\pm$0.008

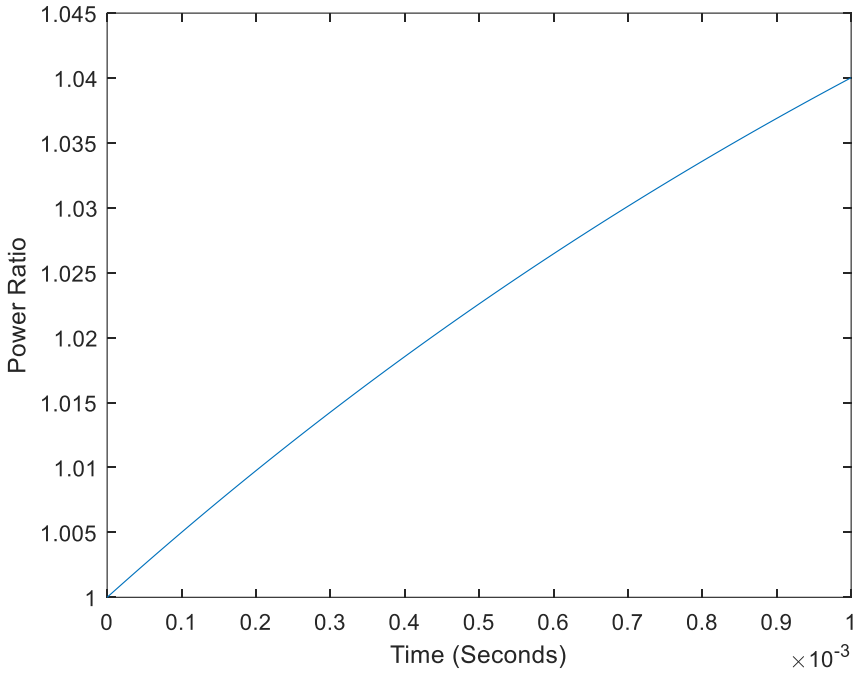


Figure 3. Sample power ratio in time.

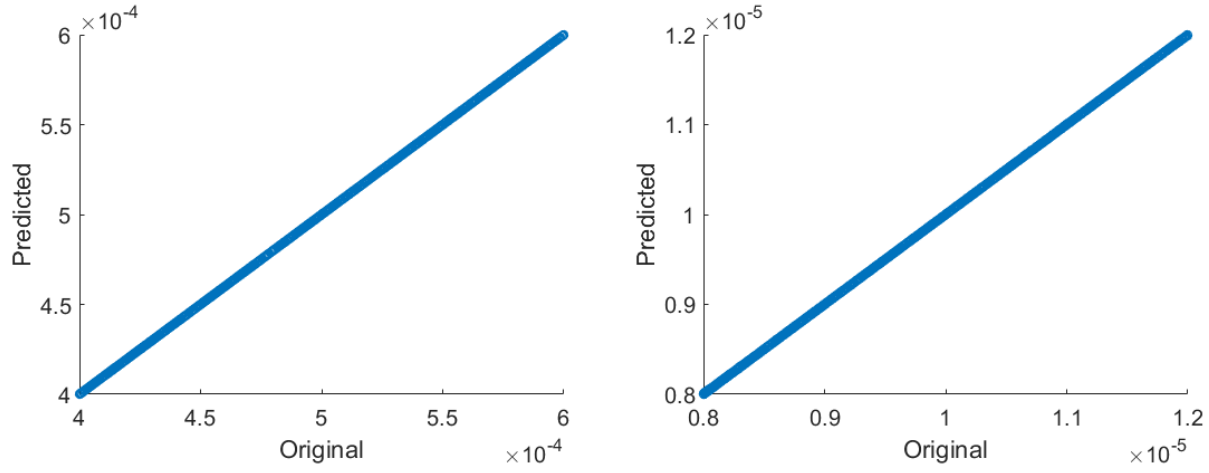
For this numerical experiment, a simple PK model of a nuclear reactor is simulated with the initial ρ, λ, β, and Λ as inputs and P and precursor-related concentration as output, via Eq. 5–6. The inverse problem is set up to solve for the PK parameters as output given the P data as input using a neural network. The time-series of P in time is used to provide sufficient degrees of freedom to generate the four PK parameters. Although there are four PK parameters, it is noted that the P data are uniquely determined by three combinations, namely, $\frac{\rho-\beta}{\Lambda}$, $\frac{\beta}{\Lambda}$, and λ. This implies that the inverse problem is ill-posed as all four parameters cannot be uniquely determined from the P distribution without additional constraints.

For the purposes of this report, the numerical experiment is subdivided into three cases: (1) a case where only two of the four PK parameters are varied, rendering the inverse problem well-posed, (2) a case where all four PK parameters are varied limiting the degree of inference and rendering the inverse problem ill-posed, and (3) a reduced form of the ill-posed case where three combinations of parameters are varied. For all cases, 50,000 samples are generated and randomly partitioned into 90% training samples, 5% validation samples, and 5% testing samples. The results are evaluated by comparing the fit PK parameters against their true values.

The experiment begins by evaluating the former case, formulated as a well-posed problem with perfect recoverability where the variables β and λ are fixed, while ρ and Λ are varied uniformly as described in Table 1. From this trial, it is observed that both ρ and Λ are recovered perfectly, as shown in Figure 4.

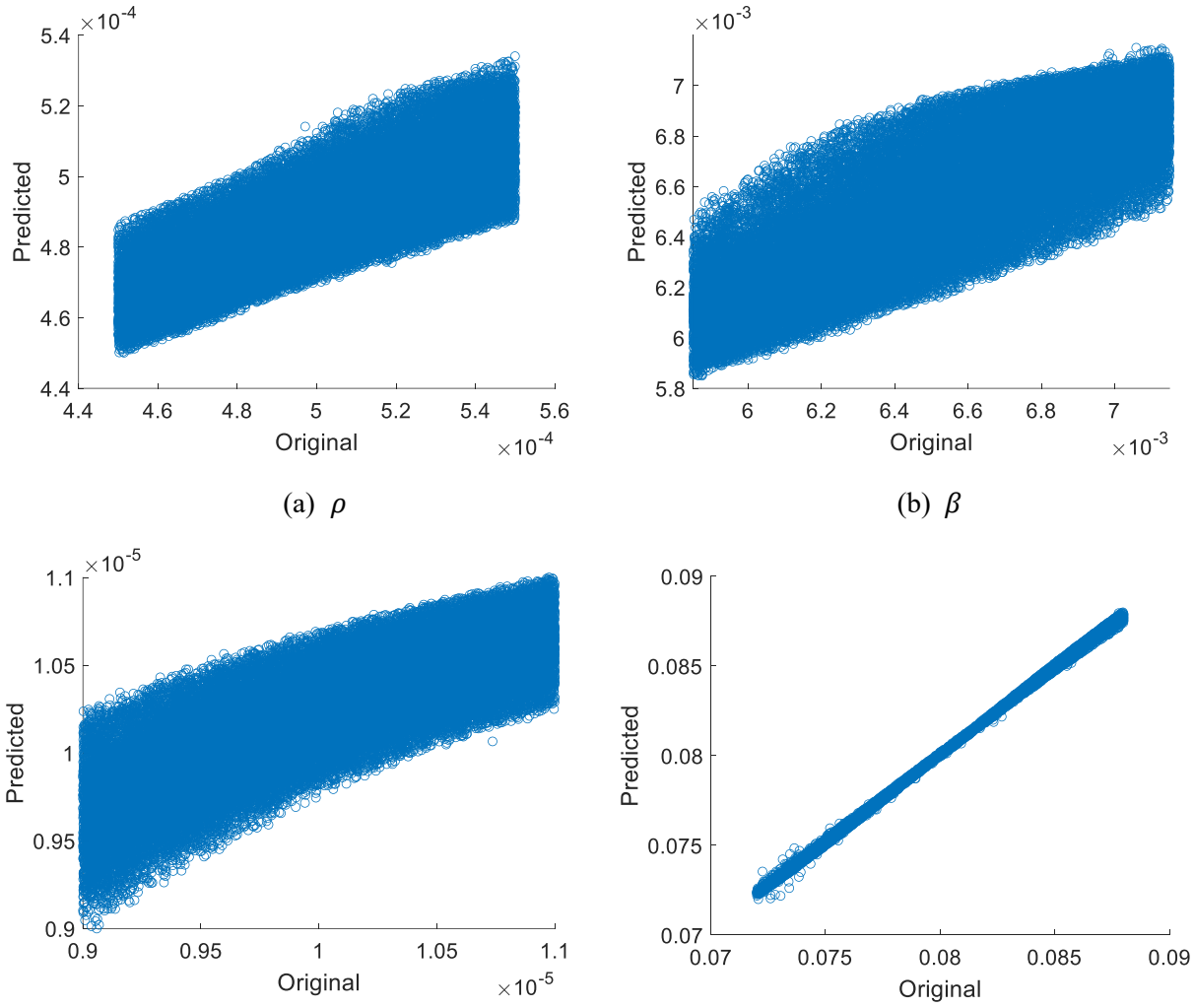
The experiment is then extended to the ill-posed case, representative of most realistic systems, where all four parameters are allowed to vary. However, the inverse problem only allows recoverability of, at most, three combinations of the parameters, resulting in the neural network applying additional assumptions/constraints such as minimization of error to arrive at one of infinite solutions. Figure 5 displays a large bound of uncertainty from this effect that will only be reduced by changing the experiment (i.e., generating PK parameters with less than 10% uncertainty per Table 1). This is an example of bias introduced by the inference procedure specifically and artificially exaggerating the degree of relationship between the input and output variables (i.e., it artificially inflates the mutual information between the input and the neural network output).

If three variables are used instead, namely $\frac{\rho}{\Lambda}$, $\frac{\beta}{\Lambda}$, and λ, they are perfectly recoverable from P without any bias from the inference procedure as shown in Figure 6. This denotes the limit of inference for the given inverse problem, after which the individual four PK parameters can only be determined with additional constraints imposed by the inference procedure (minimizing mean-squared error, L1 norm, regularization, etc.).



(a) ρ (b) Λ

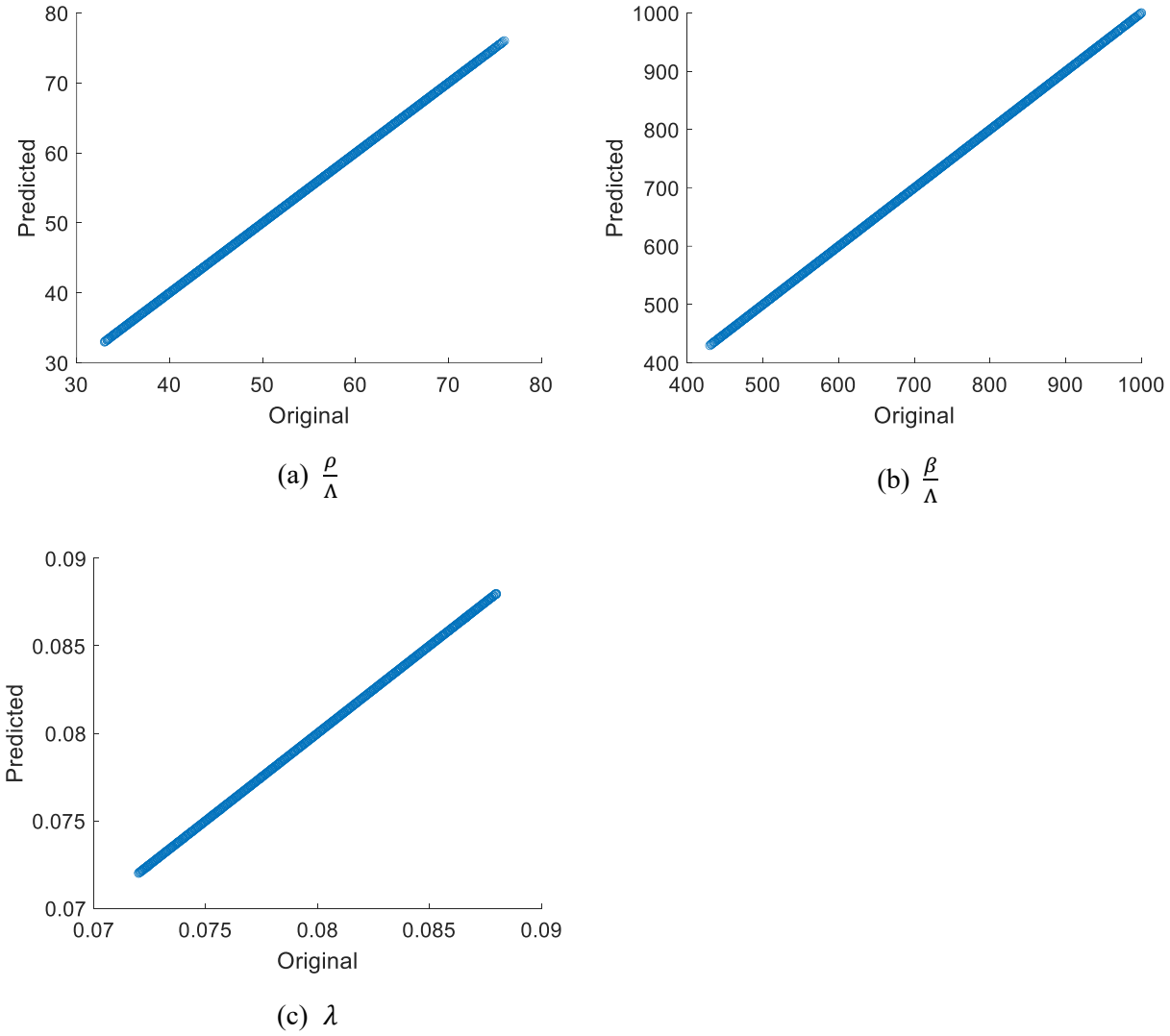
Figure 4. Original vs. predicted parameters given original data in the well-posed case.



(a) ρ (b) β

(c) Λ (d) λ

Figure 5. Original vs. predicted parameters given original data in the ill-posed case.



(a) $\frac{\rho}{\Lambda}$ (b) $\frac{\beta}{\Lambda}$

(c) λ

Figure 6. Original vs. predicted parameters given original data in the reduced ill-posed case.

3.1.2 DIOD Implementation in Physics Use Case

The generic data considered for this section of the manuscript are electrical load data from the Electric Reliability Council of Texas (ERCOT) [7]. ERCOT provides a multitude of free-access data involving the power grid, economic demand, generation, relevant documents, etc., encompassing the majority of the Texas power grid. These data are usable for many analyses (e.g., surrogate data generation, regression, classification training). For this specific task, the 2013 electrical load demand is considered, which is shown in Figure 7.

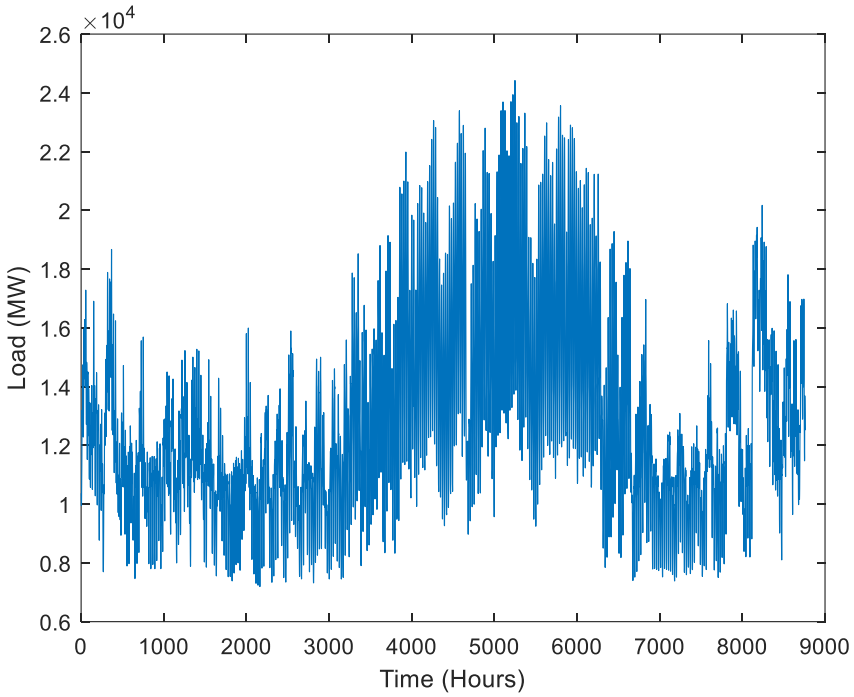
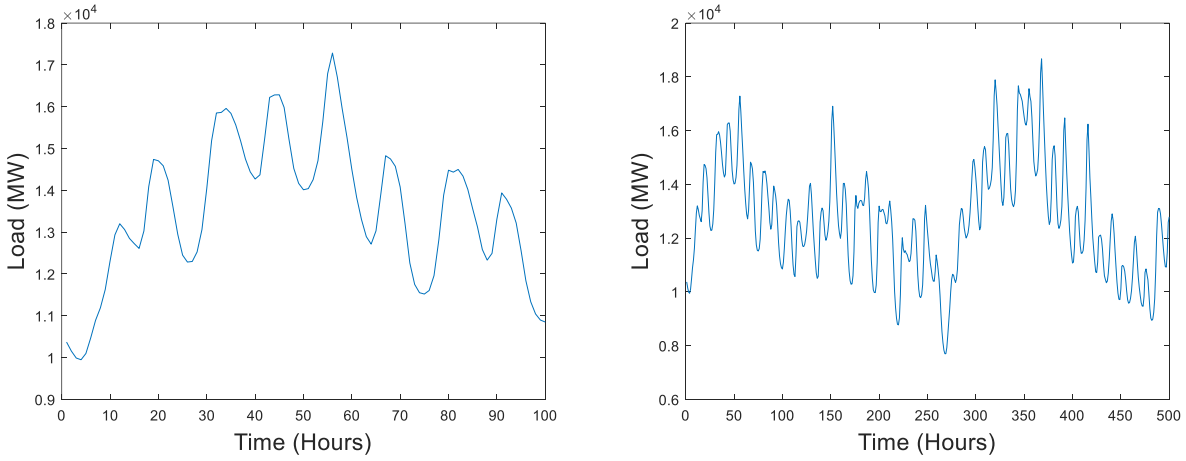


Figure 7. ERCOT hourly electrical load.

This time series, measured hourly over a year, was selected due to the presence of strong seasonality that shares a large correlation with a regular business cycle. For example, the left subplot of Figure 8 shows only the first 100 hours of the 2013 load data, and clearly reflects 12-hour and 24-hour cycles, and displays a larger trend which can be seen in the right subplot of Figure 8. The ERCOT data are dominated by seasonal behavior, in sharp contrast to the simple exponential curve, which may result in masked data shown in Figure 9 that are also dominated by seasonal behavior.



(a) Hours 1-100 (b) Hours 1-500

Figure 8. Magnified ERCOT hourly electrical load.

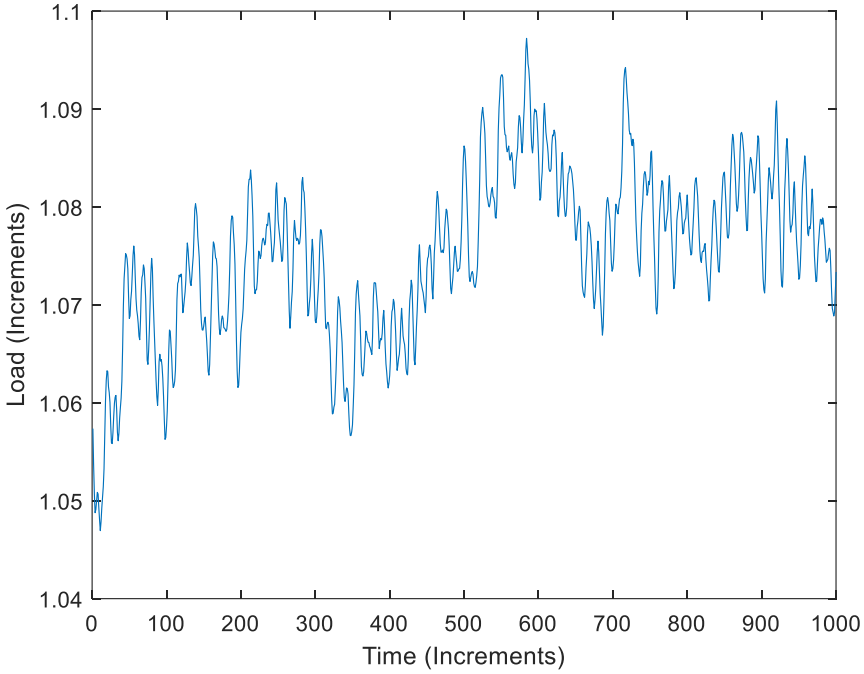


Figure 9. Sample P data, masked with DIOD.

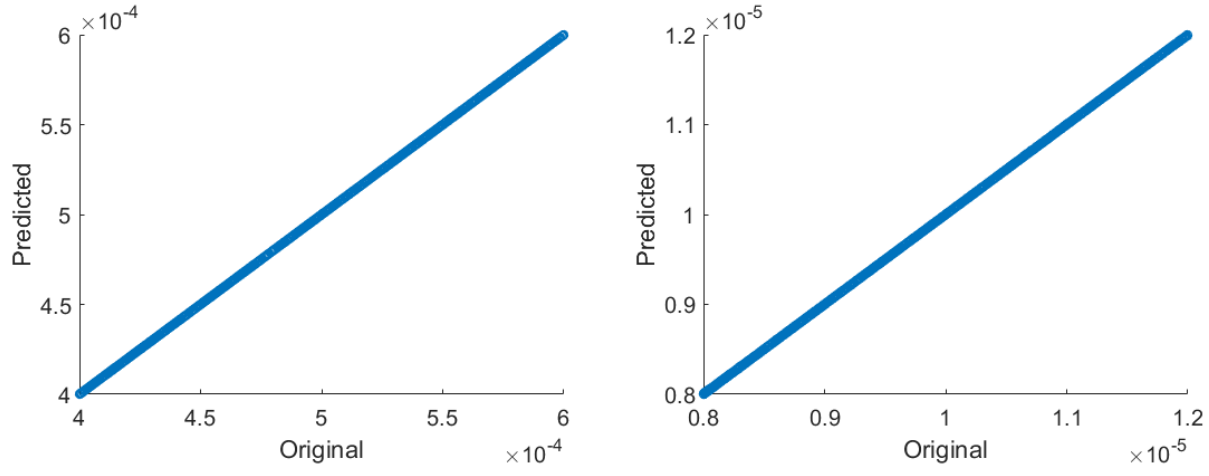
While a different behavior is not strictly required by DIOD, the level of masking (i.e., a complete change in the shape of the data) is convenient for investigating the regression case for this section. That is, if the masked data are also exponential, then the experiment carries a risk of bias (i.e., may lead the reader to ask how the success of DIOD for regression cannot also be attributed to the similarity between the sensitive and masked data). DIOD is implemented by decomposing the given P data representing the proprietary system data, y^P, and the generic data, y^G, composed of ERCOT electrical load data and applying Eq. 1–4:

$$y^P \approx \sum_{i=1}^{r} \psi_i^P(x)\phi_i^P(\alpha) \tag{7}$$

$$y^G = \sum_{i=1}^{r} \psi_i^G(x')\phi_i^G(\alpha') \tag{8}$$

$$y^D = \sum_{i=1}^{r} \psi_i^G(x')\phi_i^P(\alpha) \tag{9}$$

Here, $\psi_i^G(x), \phi_i^G(\alpha), \psi_i^{\mathrm{P}}(x'),$ and $\phi_i^P(\alpha)$ are the fundamental and inference metadata of the generic and proprietary system, respectively, obtained through reduced-order modeling techniques (such as principal component analysis [8], singular value decomposition [9], dynamic mode decomposition [10], singular spectrum analysis [11]) and y^D is the masked data obtained after applying the DIOD procedure. To validate identical performance, a neural network is trained to fit the masked data, y^D, to the four PK parameters, as performed earlier. The results are shown in Figure 10 for the well-posed case, Figure 11 for the ill-posed case, and Figure 12 for the reduced ill-posed case.



(a) ρ (b) Λ

Figure 10. Original vs. predicted parameters given masked data in the well-posed case.

To confirm the validity of DIOD, recall that the inference properties must be preserved, as measured by mutual information. Since the mutual information for the well-posed case is perfect, and for the ill-posed case is undetermined (due to the lack of constraint on the neural network and the consequential noisy results), only the reduced ill-posed case is analyzed further. Given the perfect recoverability of λ, the degree of relationship between the input and output in this case is truly quantified by the mutual information between the two-dimensional variables $[\frac{\rho}{\Lambda}\ \frac{\beta}{\Lambda}]$ and $[\rho\ \beta]$. However, due to computational difficulties in estimating mutual information among multidimensional variables and the analytical solvability of the univariate case, Table 2 estimates the mutual information between the individual terms as a proxy of the true information content. The analytically derived values are compared to the numerical data using a k-nearest-neighbors estimator [12]. It is observed that the nearest-neighbors estimate is within the uncertainty bounds for both the sensitive and masked data, thus validating that the DIOD procedure preserves mutual information. Alternatively, it may also be seen from Eq. 7–9 that the procedure amounts to an invertible transformation, which preserves the mutual information.

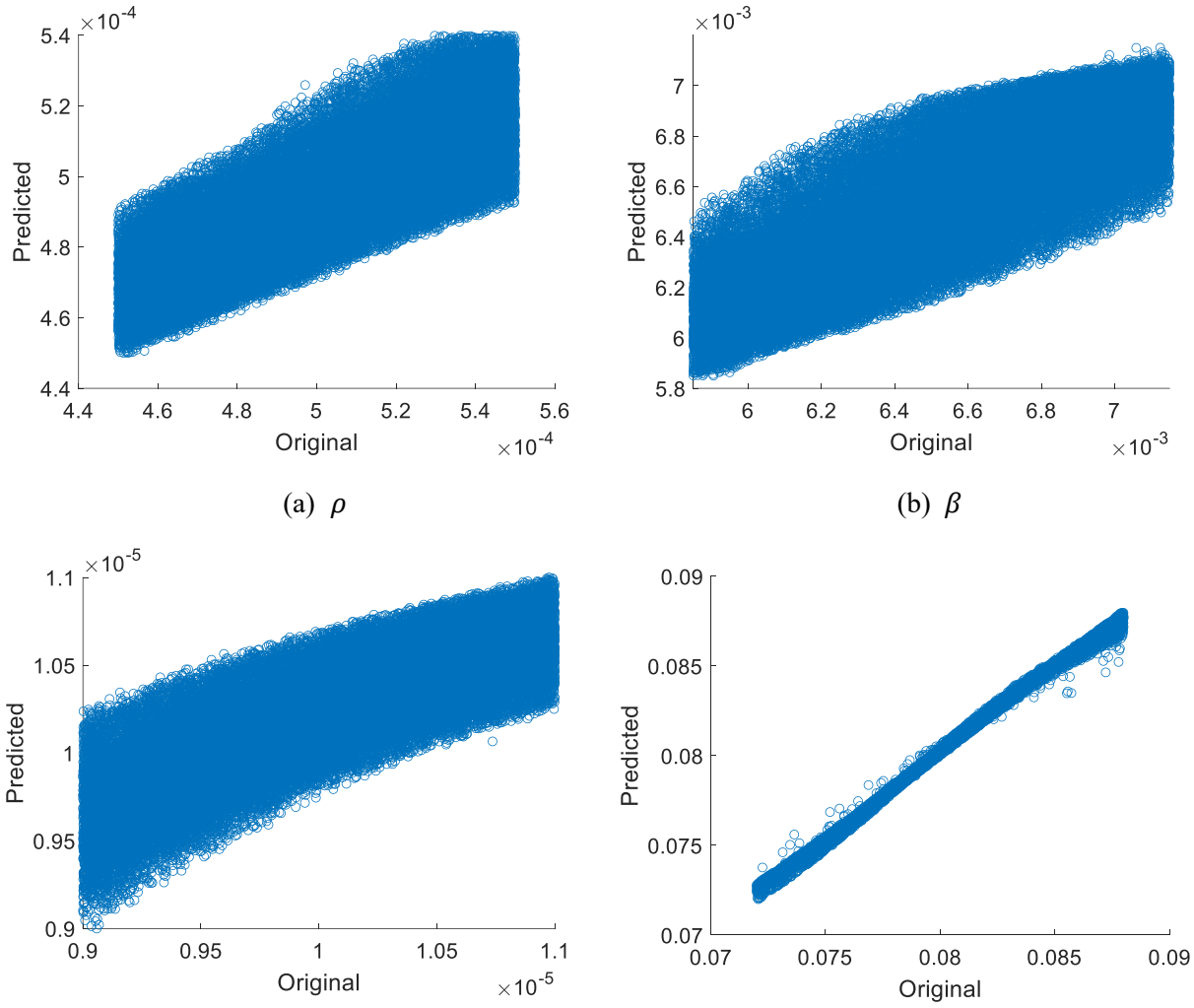
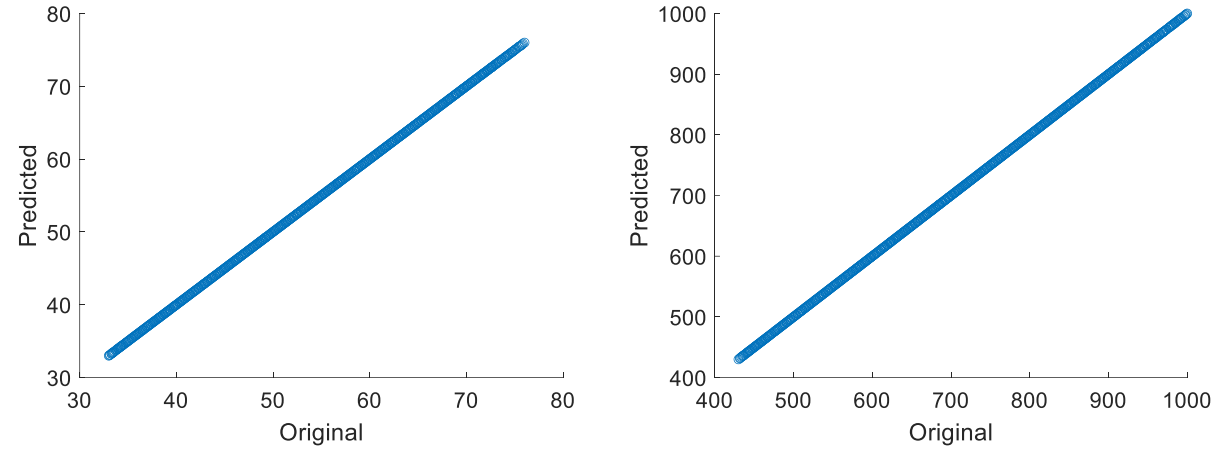
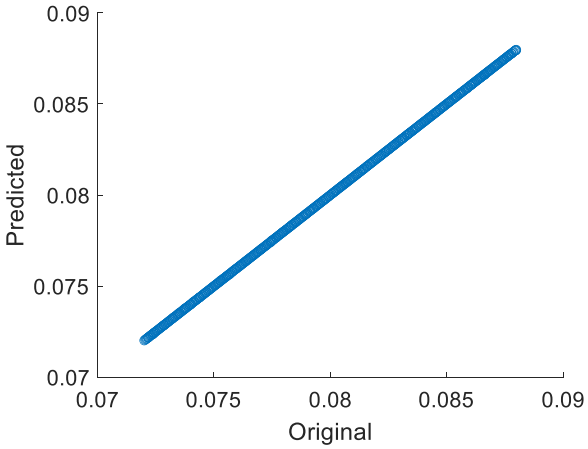


Figure 11. Original vs. predicted parameters given masked data in the ill-posed case.



(a) $\frac{\rho}{\Lambda}$

(b) $\frac{\beta}{\Lambda}$



(c) λ

Figure 12. Original vs. recovered parameters given masked data in the reduced ill-posed case.

Table 2. Mutual information between input and output parameters.

	Sensitive Data (Recovered)	**Masked Data (Recovered)**	**Sensitive Data, (True, KNN)**	**Sensitive Data, (True, Analytical)**
$I(\rho, \frac{\rho}{\Lambda})$	$0.492 \pm .0002$	$0.492 \pm .0002$	0.493	0.5008
$I(\Lambda, \frac{\rho}{\Lambda})$	$0.499 \pm .0004$	$0.499 \pm .0005$	0.500	0.5008
$I(\beta, \frac{\beta}{\Lambda})$	$0.508 \pm .0002$	$0.508 \pm .0002$	0.507	0.5008
$I(\Lambda, \frac{\beta}{\Lambda})$	$0.503 \pm .0011$	$0.501 \pm .0003$	0.500	0.5008

3.2 Process Use Case

3.2.1 Description of Process Use Case

The second use case of this effort simulates a process. In addition to obfuscating the data and showing how the mutual information is preserved, this use case demonstrates a practical application in which anomaly detection is performed on the proprietary and masked data. The proprietary and generic data are generated via an OpenModelica simulation based on well-known, simple processes [13,14]. The proprietary system is shown in Figure 13, wherein water is pumped into a reservoir via controller-activated pump which responds to reservoir pressure (i.e., the water level) containing two process valves, upstream (Valve 1) and downstream (Valve 2) of the reservoir, respectively [13]. The water is provided by an infinite source and is drained into an infinite sink. The system also contains four anomalous valves open at various times, which will be elaborated upon in the next section. The upstream and downstream valves are open for the full simulation, but allow a variable mass flow over time (randomly and according to a Gaussian distribution). The power provided by the pump also varies randomly.

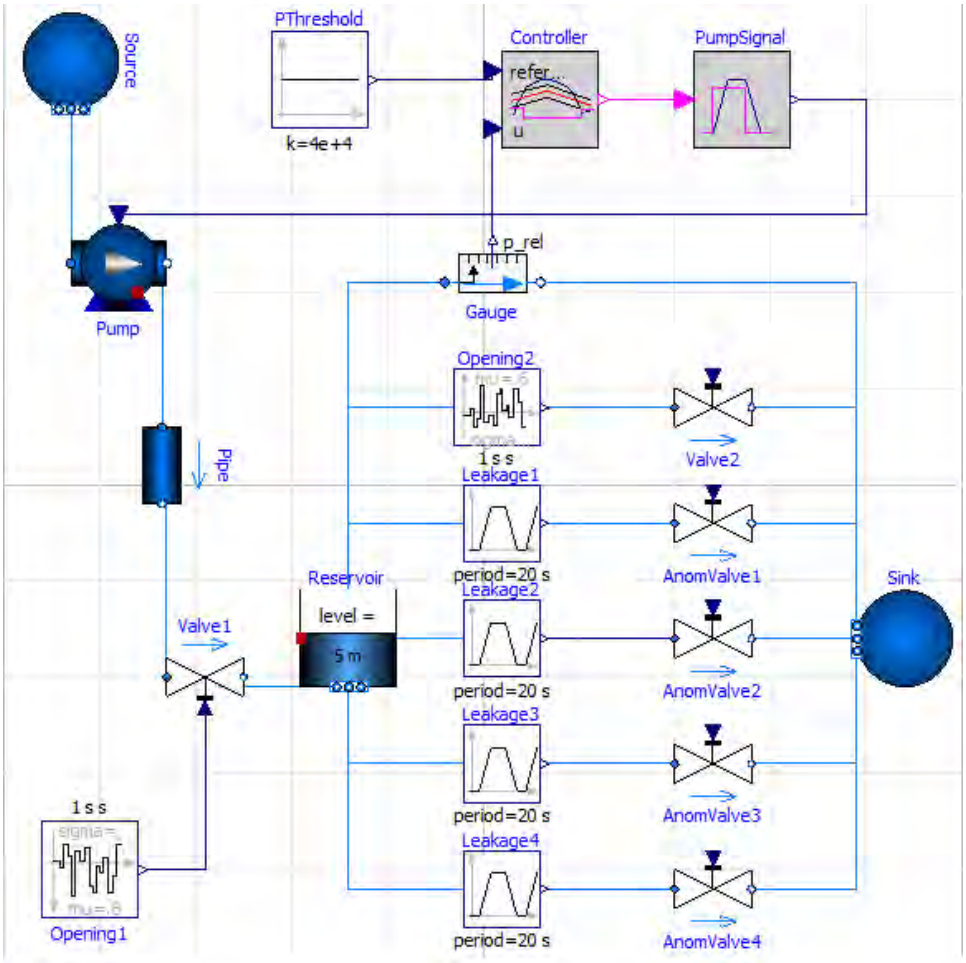


Figure 13. A simple process system simulating water pumped into a draining reservoir.

From this system, seven variables are arbitrarily chosen to comprise the sensitive data set from the proprietary system, as shown below.

1. Volumetric flow through the downstream valve
2. Power of the flow (or flow work on the total area of the pipe) downstream of the reservoir
3. Enthalpy of fluid out of the reservoir over time
4. Mass flow through the pipe upstream of the reservoir (after the upstream valve)
5. Pressure in the bottom of the reservoir, which acts as the trigger for the pump via the controller component
6. Change in the pressure at the bottom of the reservoir measured at the downstream valve over time
7. Volumetric flow through the upstream valve.

In post-processing, Gaussian noise was added to each variable to simulate both process and sensor noise. Two of the resulting variables are visualized below in Figure 14 and Figure 15.

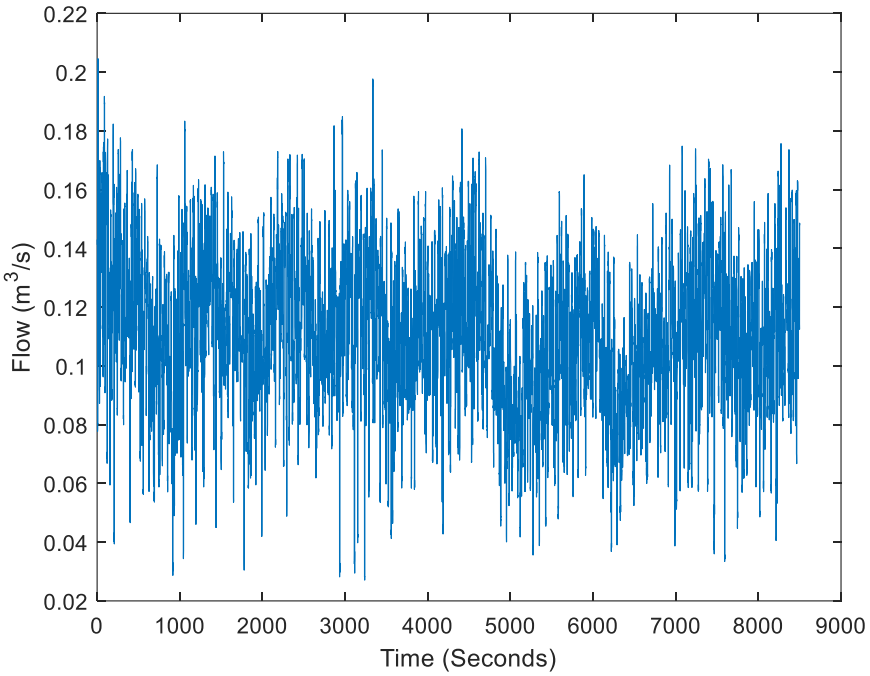


Figure 14. Volumetric flow through the downstream valve.

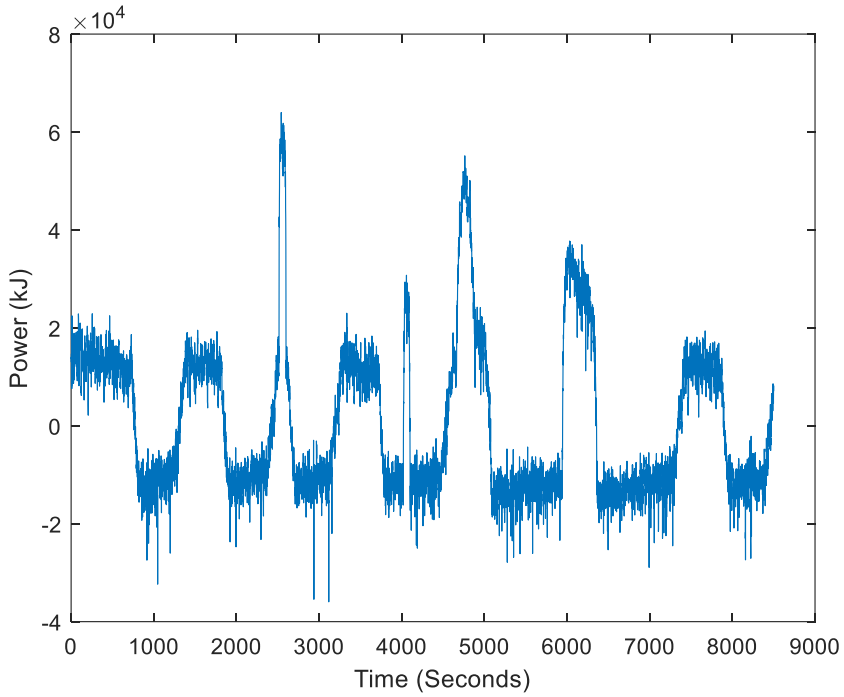


Figure 15. Power of the flow downstream from the reservoir.[a]

3.2.2 System Anomalies in Process Use Case

In this use case, anomalies are introduced into the proprietary system via process-based changes as well as post-processing data manipulation. A total of five anomalies are simulated—four describing the opening and shutting of valves (process based) and one describing sensor drift (post processing). The four process anomalies simulate a leak in the reservoir by opening anomalous valves as shown in Figure 13 (i.e., an interruption that is not part of the process) at various magnitudes and periods of time. This results in four distinct anomalous regions, shown in Table 3. Each anomaly represents a different variation of a leak included in the simulation to examine whether certain types of process anomalies are more detectable in the DIOD-masked data than others. For comparison, the mass flow of the non-anomalous valves is 1000 kg/s.

The post-processing anomaly was applied to simulate sensor drift wherein only one of the seven sensors of the proprietary system experiences a downward linear trend (i.e., drift) to its signal for a short period; the arbitrary variable chosen was enthalpy flow from the reservoir. The enthalpy flow from the reservoir is plotted in Figure 16 with the five anomalies highlighted (corresponding to times that any of the four valves were open, or sensor drift was implemented). Each of the five anomalies, particularly the process anomalies, are visible in the sensitive data.

[a] Negative power shown in Figure 15 refers to a net flow into the reservoir for a given time-step.

Table 3. Characteristics of each anomalous valve opening.

	Mass Flow Through Valve (kg/s)	Duration (Sec.)
Anomalous Valve 1	1000	20
Anomalous Valve 2	400	20
Anomalous Valve 3	1000	100
Anomalous Valve 4	400	100

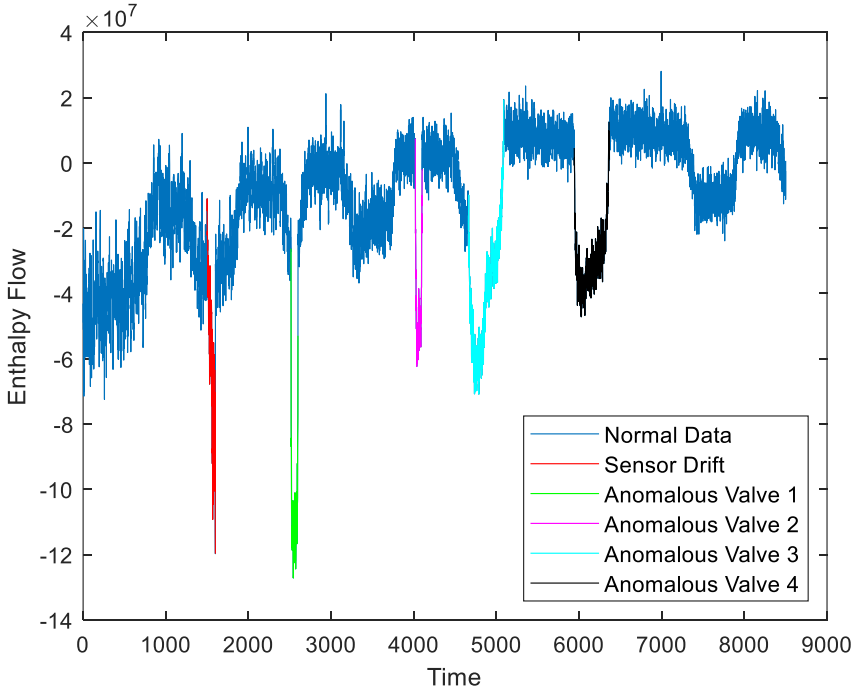


Figure 16. Enthalpy flow from reservoir with anomalies highlighted.

3.2.3 DIOD Implementation in Process Use Case

A generic system is used to generate five variables from a similar, but independent, simulation of a simple system of heated pipes as shown in Figure 17 [14]. Fewer variables are chosen for the generic data than for the sensitive data to assess whether the DIOD masking procedure can be detected by the red team. In other words, the masked data set consists of two unmasked variables and five masked variables from the proprietary system, of which two sample variables are plotted in Figure 18 and Figure 19. The variables arbitrarily taken from the generic system are:

1. Fluid temperature at the exit of Pipe 8
2. Enthalpy flow out of Pipe 6
3. Friction force experienced by the fluid in Pipe 6
4. Fluid density over time in Pipe 4
5. Change in pressure across Valve 1.

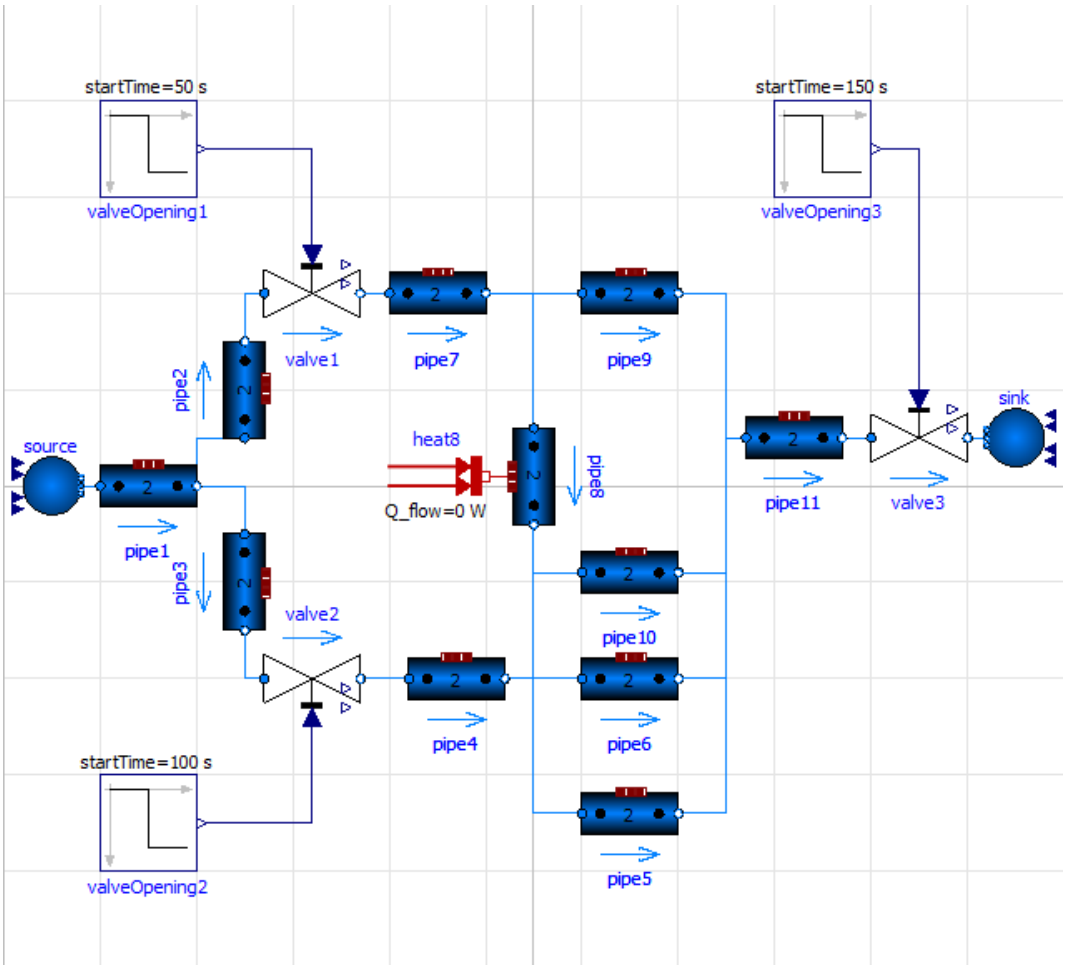


Figure 17. A simple process system simulating water flowing through a series of heated pipes [14].

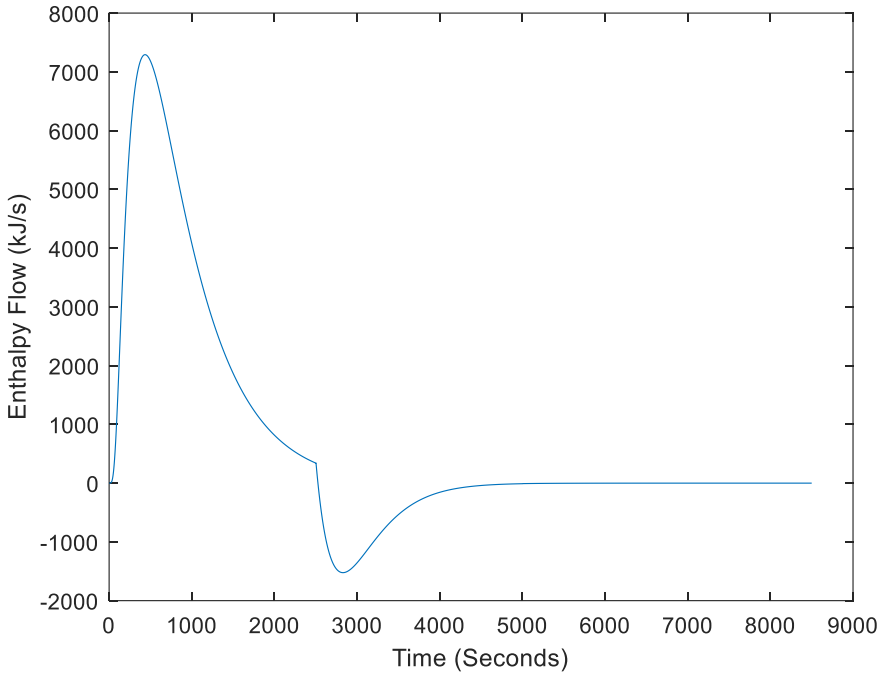


Figure 18. Enthalpy flow through Pipe 6 in Figure 17.

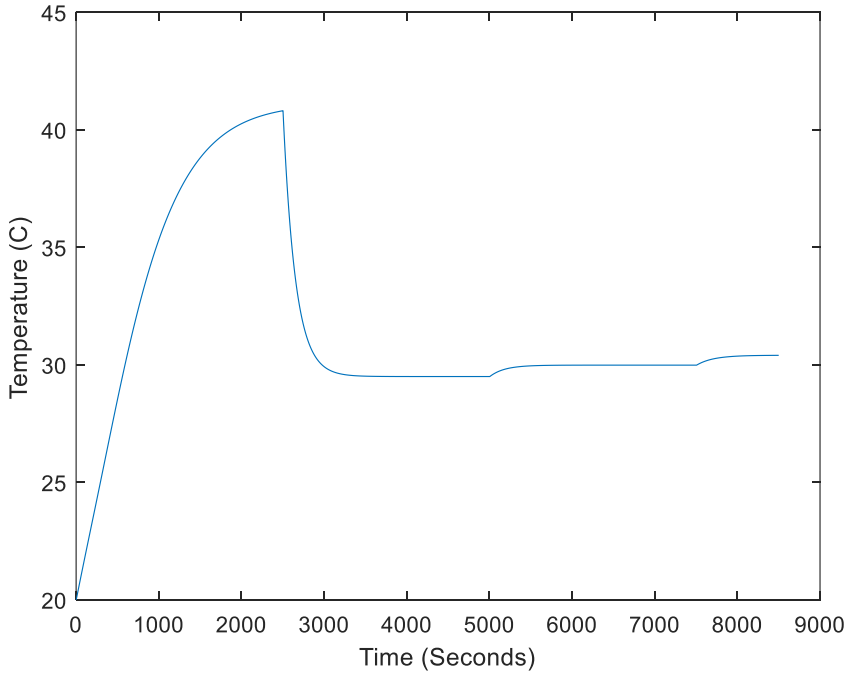


Figure 19. Temperature of fluid of Pipe 8 in Figure 17.

In this section, the DIOD procedure is implemented on the sensitive data by modifying Eq. 7–9. In the current experiment, the difference in scale of the proprietary system data may provide clues to its origin, necessitating the masking of the inference metadata. However, to preserve the mutual information

of the data with respect to anomaly detection, an invertible transformation of the inference metadata across sensors at each time-step is permitted, denoted using the function $f(\cdot)$. All other variables retain their original meaning (as in Eq. 7–9), but they are now applied to the new proprietary and generic system.

$$y^P = \sum_{i=1}^{r} \psi_i^P(x)\phi_i^P(\alpha) \tag{11}$$

$$y^G = \sum_{i=1}^{r} \psi_i^G(x')\phi_i^G(\alpha') \tag{10}$$

$$y^D = \sum_{i=1}^{r} \psi_i^G(x') f\left(\phi_i^P(\alpha)\right) \tag{12}$$

A sample of the DIOD-masked data is shown below in Figure 20. As mentioned above, only five of the seven variables are utilized to form y^D. Specifically, the volumetric flow through the downstream valve (y_1^P) and upstream valves (y_7^P) are not obfuscated by DIOD, and these variables are provided to the red team without masking.

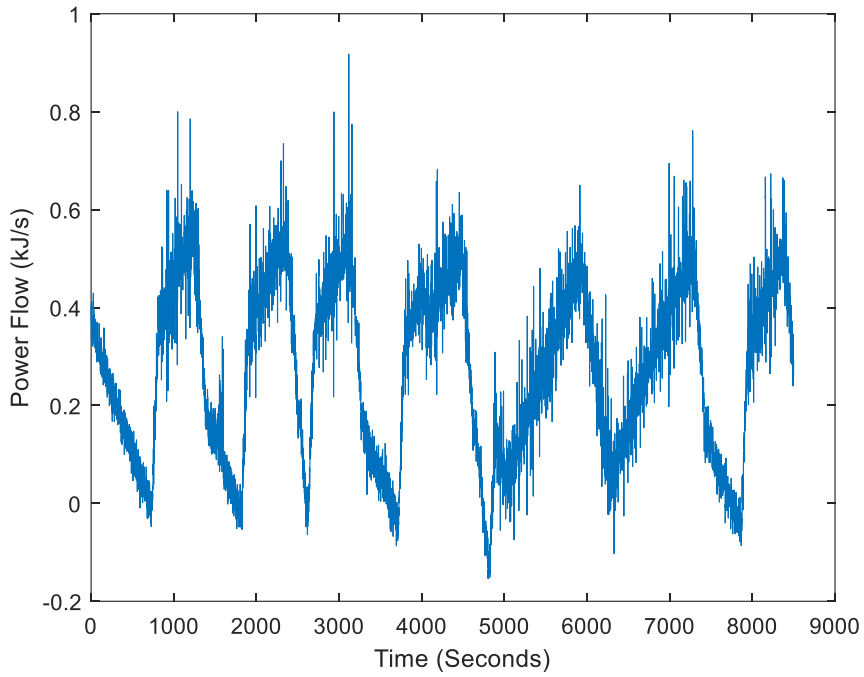


Figure 20. Masked power of flow downstream the reservoir.

3.2.3.1 Red Team Anomaly Analysis

The objective of this exercise is to demonstrate the data masking capabilities of the DIOD paradigm by verifying the preservation of inferential properties for the target AI/ML application (anomaly detection) and ensuring that the source of the data (i.e., the original process) cannot be gleaned. This section of the report was written independently of any preceding sections that relate to the process use-cases, as it was not revealed to the red team.

A red team was given a DIOD-masked data set tabulated as an 8500 × 7 matrix, consisting of seven responses labeled [y_1^D y_2^D ... y_7^D] of the variables in Section 3.2.1 and over 8,500 time-steps without information on the actual time intervals or the variables.

The task of the red team is to identify any abnormal behavior among the variables and identify the source of the data. In the data set, responses y_3^D and y_7^D display easily visible abrupt changes at various intervals as depicted in Figure 21 and Figure 22, indicating changes in operation and/or potential abnormal behavior, neither of which is as easily distinguishable in the other responses. The operational changes are drastic, occurring over a period of approximately 50 seconds, after which it achieves steady-state operation. This may be indicative of a ramp function input, or a swift response to a sudden change in setpoint, as expected by the proportional action of a proportional-integral (PI) controller. Furthermore, keeping the DIOD paradigm in mind, the observed responses may also be a linear or nonlinear combination of several original responses. However, the consistency of the peaks and dips indicate it is unlikely that some of the given responses were masked and may be from the proprietary system itself. The rest of this analysis assumes that y_3^D and y_7^D are not masked except for a scaling operation, while the others are masked via the DIOD paradigm.

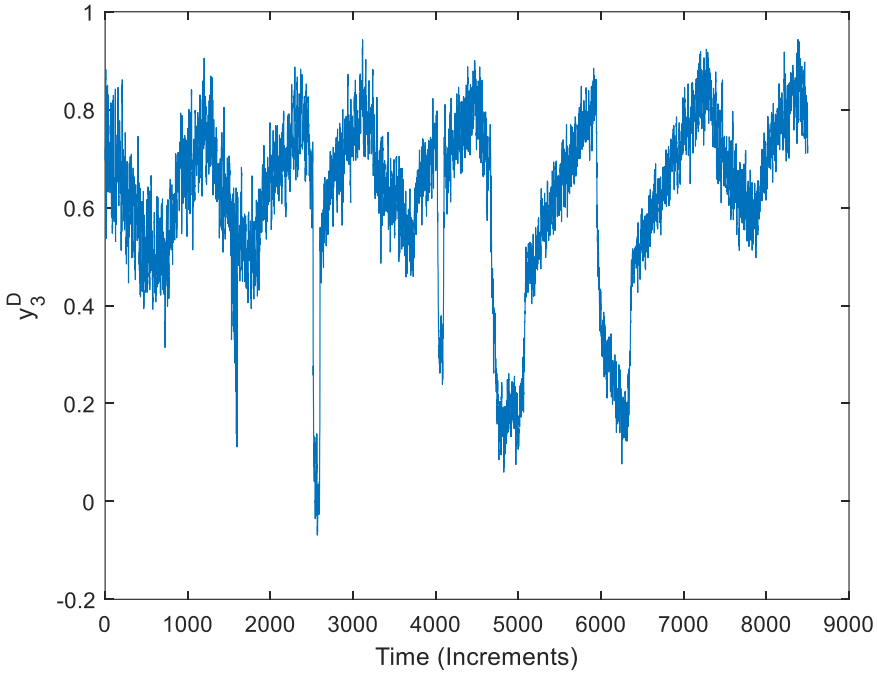


Figure 21. Response y_3^D showing anomalous regions.

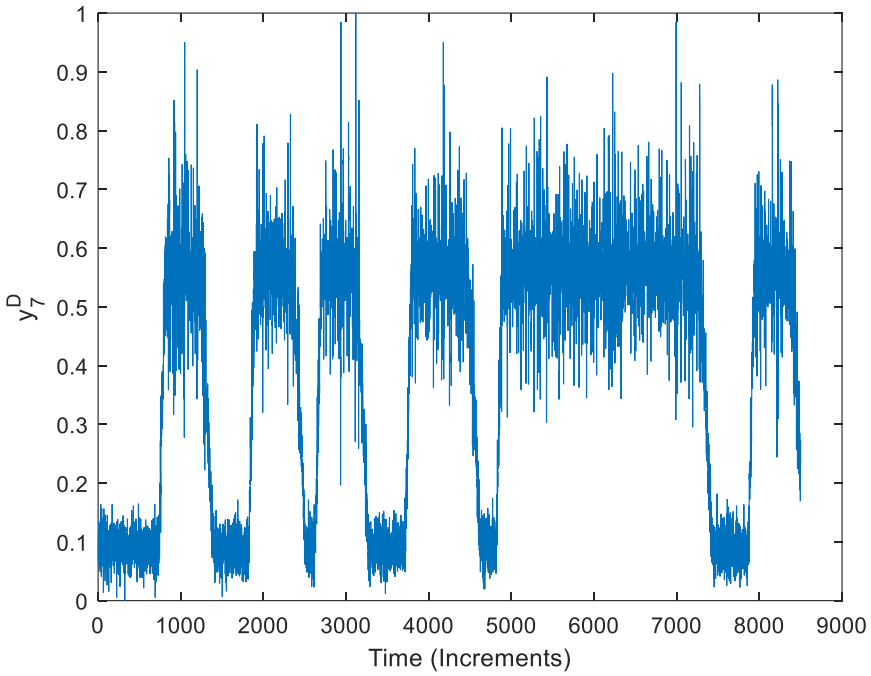


Figure 22. Response y_7^D.

Some of the changes in behavior in response y_3^D appear to be indicative of anomalous behavior due to its inconsistency with other responses. This is especially evident in the time intervals $t \in (1500, 1600) \cup (2500, 2600) \cup (4000, 4100) \cup (4650, 5100) \cup (5900, 6400)$. For instance, Figure 23 shows the quotient of y_3^D with respect to y_4^D. Here, four of the five intervals identified are clearly anomalous due to abrupt dips in the time series while the intervals outside respect the correlations over the operational periods identified using y_7^D earlier.

To distinguish between operational regime changes and anomalies further, the quotient of y_3^D with respect to y_7^D is plotted in Figure 24. It is expected that operational changes manifest themselves differently than anomalies, especially in the interval where a sizable anomaly is expected to occur, such as in the interval $(5900, 6400)$ shown in Figure 25. If the features of this anomaly persist over the other identified anomalous intervals, it may be considered further evidence in favor of the hypothesis that the identified intervals are anomalies. Furthermore, they may also be categorized as the same type of anomaly if the features are similar.

It is observed that the anomalous region is characterized by a less-noisy dip than the surrounding region, as in Figure 24. Such regions are identified in the intervals $(4000, 4100)$ and $(4650, 5100)$. While a significant dip with less relative noise is identified in the interval $(2500, 2600)$, it is unclear whether the type of anomaly is the same given the higher noise level, which may or may not be due to the change in operational regime during the interval. An earlier anomaly in the interval $(1500, 1600)$ was also identified. However, it is difficult to distinguish this anomaly from its surroundings as shown in Figure 26, and it is thus classified as a different type of anomaly or a potential artifact in the data.

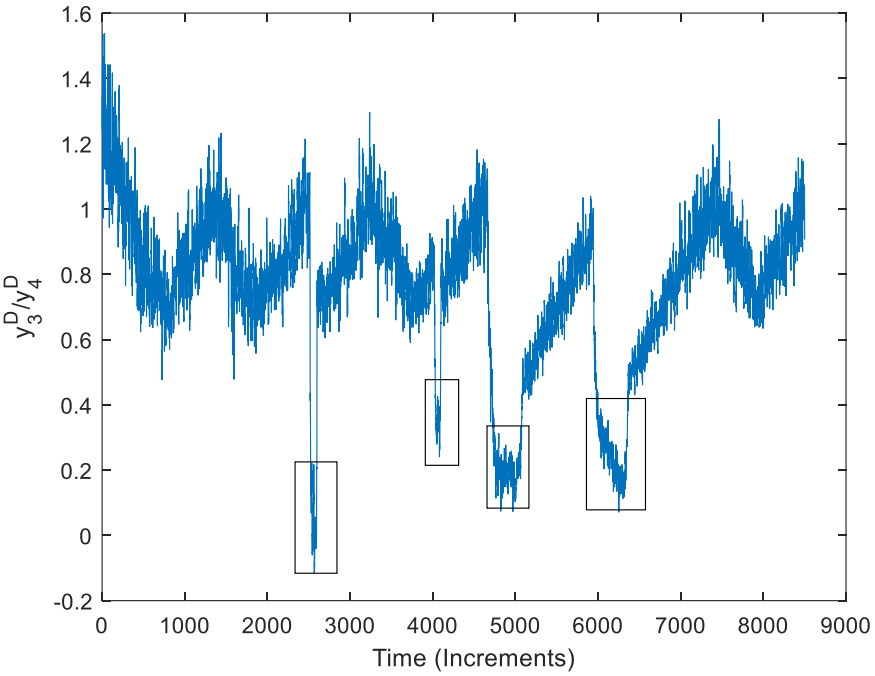


Figure 23. Response y_3^D / y_4^D depicting anomalous regions.

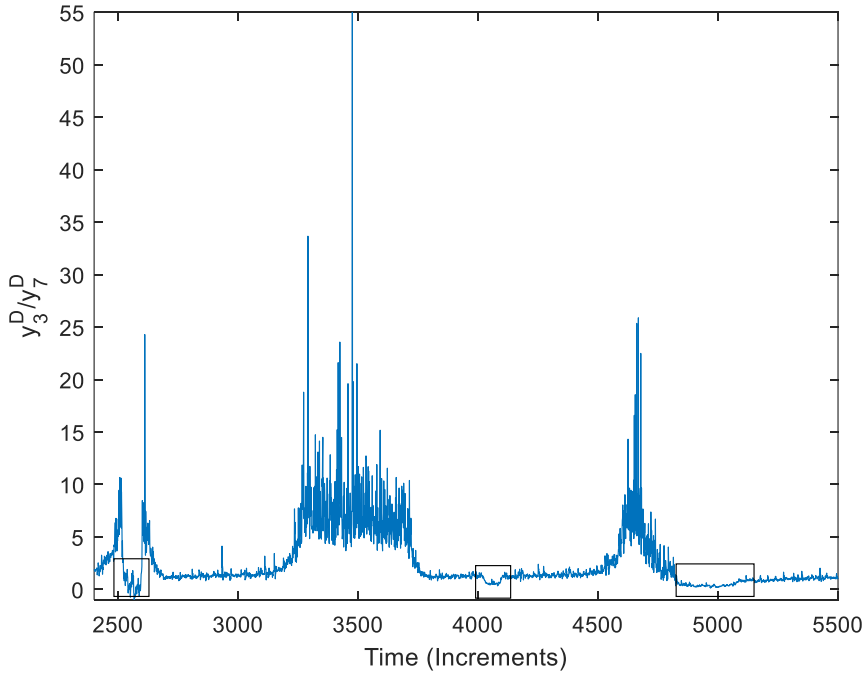


Figure 24. Response y_3^D / y_7^D depicting anomalous regions.

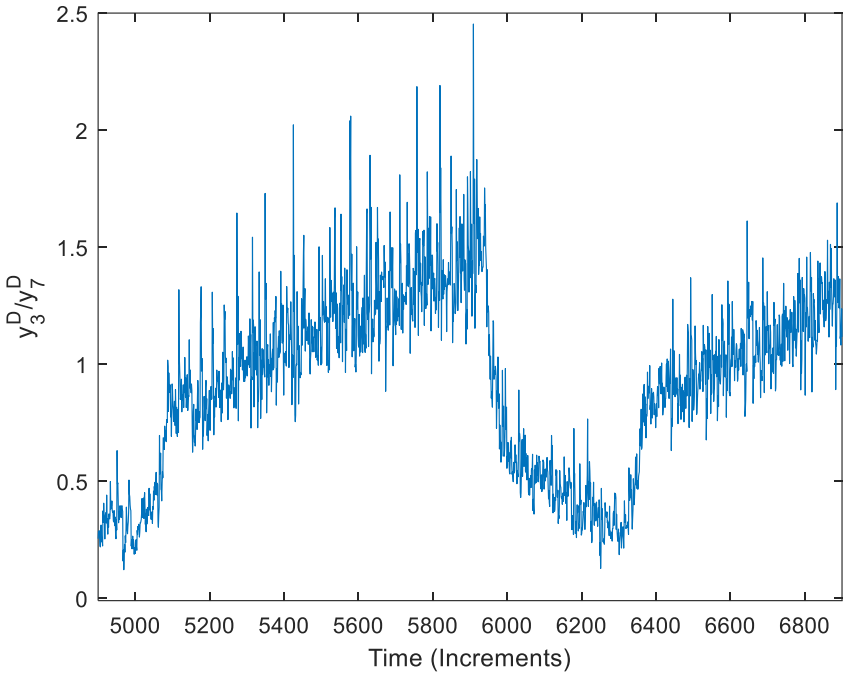


Figure 25. Response y_3^D / y_7^D depicting features of the anomalous region in the interval (5900,6400).

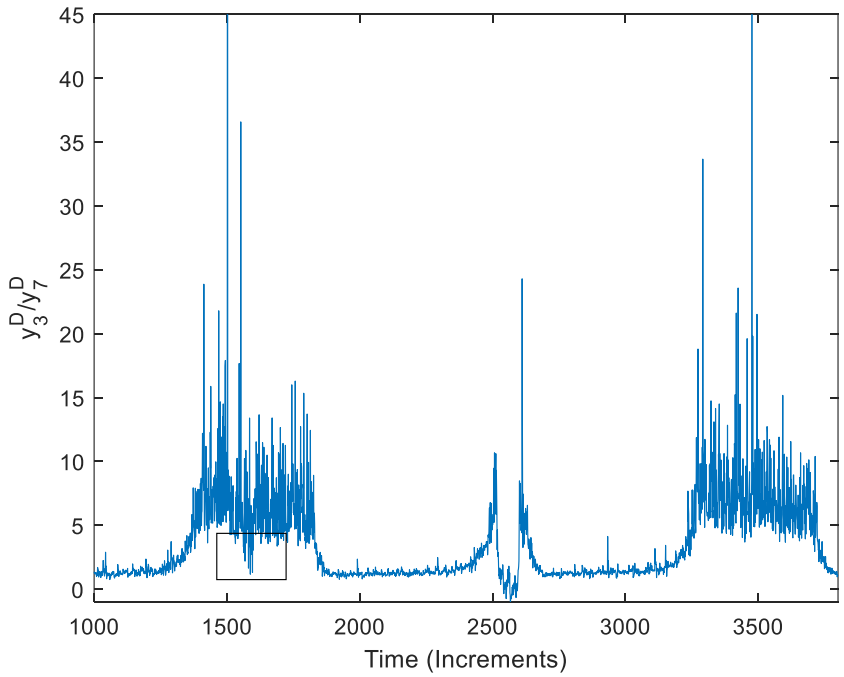


Figure 26. Response y_3^D / y_7^D depicting features of a potential anomalous region.

If the mutual information is preserved across each time-step with respect to anomaly detection, the newly formed responses would be expected to be a linear combination of the original sensors, implying that the two sets of data span the same subspace. An attempt to extract an orthogonal basis for this

subspace using singular value decomposition is next performed to validate some of the above conclusions. The five anomalous regions are clearly seen in the fifth left singular vector, as depicted in Figure 27.

In summary, the red team identified five suspicious regions using response y_3^D that are not representative of the surrounding regions. Of these anomalies, the last three share similar characteristics and may be classified as the same type. While the team is confident in its assessment that the second identified region is an anomaly, it is unclear as to whether it is a distinct type of anomaly, or the same as the last three since its features are convoluted with those of a change in operational regime. The subtleness of the potential anomaly in the first region and the lack of a clear distinction from its neighboring regions indicate that it may be an artifact in the operational data. The red team is least confident in the classification of this region.

After the conclusion of the experiment, it was confirmed that the red team correctly identified all anomalies but was unable to decipher any other information from the system.

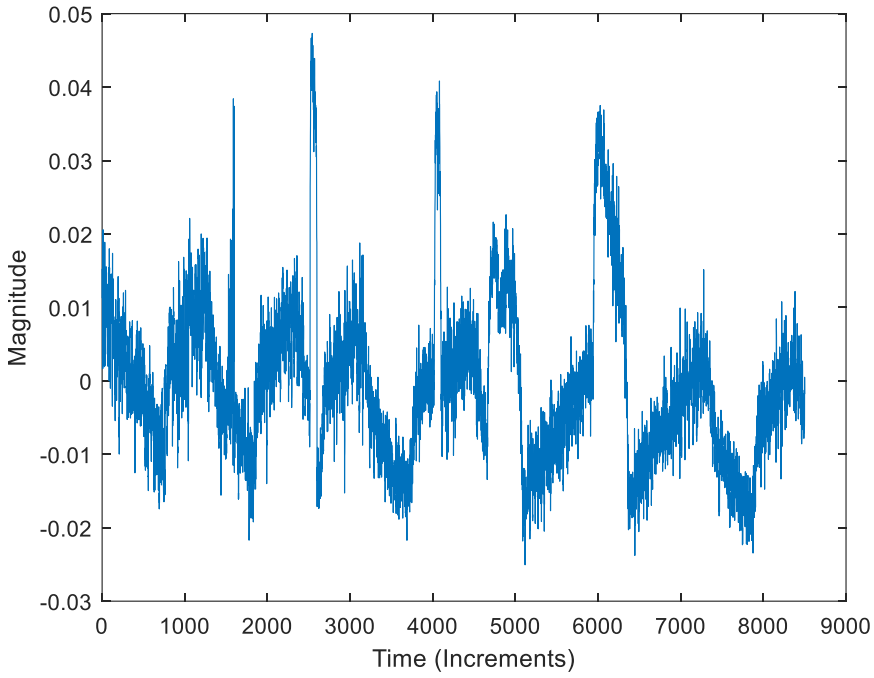


Figure 27. Fifth left singular vector depicting the anomalous regions identified earlier.

3.2.3.2 Red Team Source Analysis

Reverse-engineering efforts based on knowledge of the DIOD paradigm and domain knowledge on simulations of modern proprietary systems indicate that responses y_3^D and y_7^D are not masked and may originate from the sensors of the proprietary system due to the high level of noise. It is expected that an actuator has a significantly lower noise level due to the presence of filters and/or PI controllers, and may also exhibit saturation behavior, none of which were observed in y_3^D and y_7^D. As previously mentioned, operational regime changes were introduced through an abrupt change in setpoint or a very short ramp function, and the quick approach to steady state with a lack of overshoot indicates a PI controller with a relatively high proportional action (i.e., a first-order input-output response). Derivative action is not expected to perform well in noisy scenarios, while a higher integral action typically increases the time constant and may result in saturation unless the error is reset periodically. The red team was unable to arrive at a guess to the identity of the system beyond that the system may contain a first-order controller.

Upon communication of the results with the blue team, it was revealed that the red team was able to detect the presence of one of the unmasked variables owing to the consistency of the process in contrast with other variables. However, the proprietary system of simulated water pumped into a reservoir was not revealed. While the anomalies were correctly identified and grouped, the pump model followed a first-order input-output response, and the operational regimes were the result of abrupt changes (revealed as valve action). In summary, the red team was unsuccessful in its attempt to determine the identity of the system or reverse the masking process to obtain the proprietary data.

4. CONCLUSION

This report outlines the DIOD data-masking paradigm and its applicability to regression and unsupervised learning tasks. Specifically, data from physics- and process-based systems were simulated using a simple PK model and a realistic reservoir-pump system model. These data were masked to appear like data from a generic system. Inference procedures were employed on the original and masked data sets for both use cases. Comparing the results from inference models (in the physics-based system) and human evaluation (in the process-based system) validated the claim that the DIOD application did not alter the inference conclusions and preserved information content for a target AI/ML application.

In addition to preserving mutual information, the irreversibility of DIOD-masked data was demonstrated by a red and blue team exercise on the process-based use case. The red team was able to successfully identify the five anomalies and broadly categorize them into two types in an unsupervised manner. The red team successfully identified one of the two intentionally unmasked responses, but this did not lead to any clues to the process identity of the response. The team was also able to categorize the type of operational regime change in the proprietary system as an abrupt change in setpoint or a very fast ramp; however, the red team was unable to correctly identify the proprietary system or reconstruct the proprietary data.

5. REFERENCES

1. Sarada G., N. Abitha, G. Manikandan, and N. Sairam. 2015. "A few new approaches for data masking," 2015 International Conference on Circuits, Power and Computing Technologies [ICCPCT-2015], pp. 1–4. DOI: 10.1109/ICCPCT.2015.7159301.
2. Yi, X., R. Paulet, E. Bertino. 2014. "Homomorphic Encryption." *Homomorphic Encryption and Applications. SpringerBriefs in Computer Science*. Springer, Cham. DOI: 10.1007/978-3-319-12229-8_2.
3. Dwork, C. 2008. "Differential Privacy: A Survey of Results." *Agrawal, M., Du, D., Duan, Z., Li, A. (eds) Theory and Applications of Models of Computation.* TAMC 2008. Lecture Notes in Computer Science, vol. 4978. Springer, Berlin, Heidelberg. DOI: 10.1007/978-3-540-79228-4_1.
4. Sundaram, A., H.S. Abdel-Khalik, and A. Al Rashdan. 2022. "Deceptive Infusion of Data: A Novel Data Masking Paradigm for High-Valued Systems." Nuclear Science and Engineering, 196:8, 911–926. DOI: 10.1080/00295639.2022.2043542.
5. Li, H., X. Shang and W. Chen. 2010. "An accurate solution of point kinetics equations of one-group delayed neutrons and an extraneous neutron source for step reactivity insertion." Chinese Science Bulletin, 55, 4116–4119. DOI: 10.1007/s11434-010-4220-2.
6. Duderstadt, J. J., L.J. Hamilton, S. Moorthy, and C.C. Scott. 1977. "Nuclear Reactor Analysis by James J. Duderstadt and Louis J. Hamilton," *IEEE Transactions on Nuclear Science*, 24:4, 1983–1983. DOI: 10.1109/TNS.1977.4329141.
7. "Grid Information – Load," Electric Reliability Council of Texas (ERCOT). Accessed October 10, 2022. https://www.ercot.com/gridinfo/.
8. Aversano, G., et al. 2019. "Application of reduced-order models based on PCA and Kringing for the development of digital twins of reacting flow applications." *Computers & Chemical Engineering* 121, 422–441. DOI: 10.1016/j.compchemeng.2018.09.022.
9. Hocker, A. and V. Kartvelishvili. 1996. "SVD approach to Data Unfolding," *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 372:3, 469-481.DOI: 10.1016/0168-9002%2895%2901478-0.
10. Clainche, S. and J. Vega. 2020. "A Review on Reduced Order Modeling using DMD-Based Methods." IUTAM Symposium on Model Order Reduction of Coupled Systems, Stuttgart, Germany, May 22–25, 2018, 55–66. DOI: 10.1007/978-3-030-21013-7_4.
11. Golyandina, N., A. Korobeynikov, and A. Zhiglijavsky. 2018. "Singular Spectrum Analysis with R." Springer-Verlag Berlin Heidelberg. DOI: 10.1007/978-3-662-57380-8.
12. O'Toole, J. 2020. "Mutual Information kNN." (Matlab Code). [Source Code] Accessed November 9, 2022. https://github.com/otoolej/mutual_info_kNN.
13. Winkler, D. 2020. Modelica Association. Pumping System. [Source Code] Accessed November 9, 2022. https://github.com/modelica/ModelicaStandardLibrary/blob/master/Modelica/Fluid/Examples/PumpingSystem.mo.
14. Beutlich, T. 2020. Modelica Association. Branching Dynamic Pipes. [Source Code] Accessed November 9, 2022. https://github.com/modelica/ModelicaStandardLibrary/blob/master/Modelica/Fluid/Examples/BranchingDynamicPipes.mo.