AFRL-RV-PS-
TR-2023-0046

AUTONOMOUS ANOMALY DETECTION VIA PHYSICS-REGULARIZED MACHINE LEARNING

Renato Zanetti and Felipe Giraldo-Grueso

University of Texas at Austin
201 E 24th St
Austin, TX 78712

22 March 2023

Final Report

AIR FORCE RESEARCH LABORATORY
Space Vehicles Directorate
3550 Aberdeen Ave SE
AIR FORCE MATERIEL COMMAND
KIRTLAND AIR FORCE BASE, NM 87117-5776

DTIC COPY

NOTICE AND SIGNATURE PAGE

AFRL-RV-PS-TR-2023-0046 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

//SIGNED//

Michelle Simon
Program Manager

//SIGNED//

Jeff Ganley
Technical Advisor, Spacecraft Control
Technology Branch

//SIGNED//

John Beauchemin
Chief Engineer, Space Craft
Technology Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
22-03-2023	Final Report	1 Sep 2021 - 31 Dec 2022

4. TITLE AND SUBTITLE
Autonomous Anomaly Detection via Physics-Regularized Machine Learning

5a. CONTRACT NUMBER
FA9453-21-1-0045

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
61102F

6. AUTHOR(S)
Felipe Giraldo-Grueso and Renato Zanetti

5d. PROJECT NUMBER
3001

5e. TASK NUMBER
EF134389

5f. WORK UNIT NUMBER
V1Q5

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Texas at Austin
201 E 24th Street
Austin, TX 78712

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Air Force Research Laboratory
Space Vehicles Directorate
3550 Aberdeen Avenue SE
Kirtland AFB, NM 87117-5776

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RVS

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
AFRL-RV-PS-TR-2023-0046

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited (AFRL-2023-2704 dtd 15 Jun 2023)

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Manual inspection of telemetry data in the search for anomalies is a time-consuming threat detection technique. Most multi-signal systems send back extensive data that a single person cannot easily monitor in real time. Machine learning techniques that autonomously scan data and flag anomalies are attractive alternatives. The autonomous anomaly detection problem can be divided into two sub-problems: regression analysis and a classification process. In the regression analysis, a machine learning model is trained to reconstruct a given signal, and the classification process categorizes the reconstruction error as anomalous or nominal. This report examines the autonomous anomaly detection problem and proposes improvements to both the regression and classification sub-problems. With regard to the regression analysis, it was found that including the physics of the target signal in the machine learning model yielded a lower reconstruction error compared to a purely data-driven model. The classification approaches studied showed that cluster-based thresholding techniques accompanied by a pruning procedure can outperform non-parametric dynamic thresholds.

15. SUBJECT TERMS
anomaly detection, regularization, regression, classification, machine learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Michelle Simon
Unclassified	Unclassified	Unclassified	Unlimited	36	19b. TELEPHONE NUMBER *(include area code)*

(This Page Intentionally Left Blank)

TABLE OF CONTENTS

Section			Page
LIST OF FIGURES			**iii**
LIST OF TABLES			**iv**
1	**SUMMARY**		**1**
2	**INTRODUCTION**		**1**
3	**RELATED WORK**		**2**
4	**DATA PREPARATION**		**2**
	4.1	Regression	3
	4.2	Classification	4
		4.2.1 Real Anomalies.	4
		4.2.2 Synthetic Anomalies.	5
	4.3	Summary of Data	6
5	**LEARNING FRAMEWORK**		**7**
	5.1	ACE dataset	7
		5.1.1 Autoencoders.	7
		5.1.2 Loss functions.	7
		5.1.3 Bias-Variance Trade-off.	9
		5.1.4 Curriculum Regularization.	9
	5.2	SMAP/MSL dataset	10
		5.2.1 Long Short-Term Memory Networks.	10
		5.2.2 Gaussian Assumption.	10
		5.2.3 K-means clustering.	11
		5.2.4 Pruning procedure.	11
	5.3	ACEC dataset	12
6	**TRAINING AND TESTING**		**12**
	6.1	Training Hyperparameters	12
	6.2	Metrics	13
7	**RESULTS AND DISCUSSION**		**13**
	7.1	Regression	13
		7.1.1 Physics Regularization.	13
		7.1.2 Curriculum Regularization.	15
	7.2	Classification	16
		7.2.1 Real Anomalies.	16
		7.2.2 Synthetic Anomalies.	20

TABLE OF CONTENTS (continued)

Section		Page
8	CONCLUSION AND FUTURE WORK	24
9	RECONSTRUCTIONS APPENDIX	27

LIST OF FIGURES

Figure		Page
Figure 1.	Magnetic field measurements filtered from the ACE dataset	3
Figure 2.	Divergence of magnetic field measurements from the ACE dataset	4
Figure 3.	Training and testing time series for channel P-11 of the SMAP/MSL dataset	4
Figure 4.	Process followed to create the ACEC dataset	5
Figure 5.	Example of a periodic fault introduced synthetically in the ACE testing data	5
Figure 6.	Example of a drift fault introduced synthetically in the ACE testing data	6
Figure 7.	Example of random faults introduced synthetically in the ACE testing data	6
Figure 8.	Model used to reconstruct the magnetic field signal in the ACE dataset	8
Figure 9.	Bias-variance trade-off as a function of the regularization weight, as λ increases, the model decreases its complexity. Figure adapted from James et al. [10]	9
Figure 10.	Network architecture used for the SMAP/MSL dataset. Orange boxes represent the telemetry value at each time step. Figure adapted from Hundman et al. [8]	10
Figure 11.	Reconstruction error (Top) between the real and predicted telemetry value for channel T-12. Comparison (Bottom) between true, predicted and pruned anomalies for T-12 in the SMAP/MSL dataset	12
Figure 12.	Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\boldsymbol{\lambda_2} \in [0, 10]$	14
Figure 13.	Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\boldsymbol{\lambda_2} \in [10^1, 10^{10}]$	15
Figure 14.	Average number of epochs as a function of $\boldsymbol{\lambda_2} \in [0, 10]$	15
Figure 15.	$F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the gaussian assumption	17
Figure 16.	True, predicted, and pruned anomalies for channel T-9 (top) using the Gaussian assumption threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the Gaussian assumption threshold	17
Figure 17.	Error classification using K-means clustering for an arbitrary channel (E-11)	18
Figure 18.	$F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the K-means clustering algorithm	19
Figure 19.	True, predicted, and pruned anomalies for channel T-9 (top) using the K-means clustering threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the K-means clustering threshold	20
Figure 20.	$F_{0.5}$ score for the *Soil Moisture Active Passive* (SMAP) and *Mars Science Laboratory* (MSL) using the original paper's classification model [8] (OP), the Gaussian assumption classification model with pruning (GA) and the K-means clustering classification model with pruning (KM)	21
Figure 21.	Test set division to determine and evaluate the pruning threshold	22
Figure 22.	F_1 score obtained for the evaluation set on the ACEC data set using the data-driven model (DD), the physics-regularized model (PR) and curriculum-regularized model (CR) both with and without pruning	24

LIST OF TABLES

Table		Page
Table 1.	Comparison between the best-averaged physics informed model ($\boldsymbol{\lambda_2 = 2}$) and the averaged model trained by curriculum regularization	16
Table 2.	Comparison between the different thresholding techniques. The first score refers to SMAP and the second score to MSL	19
Table 3.	Comparison between the performance metrics obtained with the different training techniques on the ACEC test set	21
Table 4.	Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for the purely data-driven model	22
Table 5.	Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for the physics regularization model	23
Table 6.	Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for curriculum regularization model	23

1 SUMMARY

Manual inspection of telemetry data in the search for anomalies is a time-consuming threat detection technique. Most multi-signal systems send back extensive data that a single person cannot easily monitor in real time. Machine learning techniques that autonomously scan data and flag anomalies are attractive alternatives. The autonomous anomaly detection problem can be divided into two sub-problems: regression analysis and a classification process. In the regression analysis, a machine learning model is trained to reconstruct a given signal, and the classification process categorizes the reconstruction error as anomalous or nominal. This report examines the autonomous anomaly detection problem and proposes improvements to both the regression and classification sub-problems. With regard to the regression analysis, it was found that including the physics of the target signal in the machine learning model yielded a lower reconstruction error compared to a purely data-driven model. The classification approaches studied showed that cluster-based thresholding techniques accompanied by a pruning procedure can outperform non-parametric dynamic thresholds.

2 INTRODUCTION

Anomaly detection is a crucial process in multi-signal systems. Complex systems, such as spacecraft, have multiple channels with individual telemetry signals that operators cannot monitor in real time to identify anomalies and prevent faults and failures. As most systems have many channels that require constant observation, manual monitoring is nonviable, and fluctuations in the data that could lead to larger anomalies often remain undetected. Alternatively, artificial intelligence approaches can be used as autonomous anomaly detection models, specifically those that rely on machine learning. Multiple machine-learning models have already been used to monitor CT scans, prices, stocks, and telemetry data [4]. In recent years, machine learning has proven useful for anomaly detection and has been an active research field.

Traditional anomaly detection in satellites relies on manual inspection by a qualified person with sufficient knowledge, and alarms to signal anomalous behavior when a value exceeds a predetermined limit [11]. Satellites typically generate large amounts of telemetry data, which are commonly multivariate time series, making traditional anomaly detection inefficient [11]. Recently, artificial neural networks have been used to replace traditional anomaly detection methods in satellite telemetry data. More specifically, in 2018, a deep learning model based on Long Short-Term Memory (LSTM) layers was implemented to identify anomalies in labeled telemetry data obtained from the *Mars Science Laboratory* and *Soil Moisture Active Passive* satellites [8]. In 2019, Jin et al. [11] also trained a denoising autoencoder to detect anomalies in a satellite power subsystem.

When nominal data is available, the autonomous anomaly detection problem can be solved in the following two steps [4]. First, a machine learning model (usually an autoencoder or a multi-layer neural network) is trained to reconstruct or predict nominal data. This first step can be considered as a regression problem. Second, the reconstruction error is used to classify each point as either an anomaly or a nominal behavior [15]. For the classification process, nearest neighbors and clustering techniques have been used to set automatic anomaly thresholds

[11]. An efficient anomaly detection algorithm must reconstruct the nominal behavior as closely as possible so that the actual anomalies in the data return an unusually high reconstruction error, which classification algorithms can easily detect.

Most signal reconstruction algorithms are trained to minimize the reconstruction error without considering physical constraints in the training data. Physics Informed Neural Networks (PINNS) were first introduced in 2019 to allow the reigning physics of the problem to be included [16]. These are specific neural networks trained to solve supervised learning tasks while considering any given physical law described by general partial differential equations. If the physics governing the studied phenomenon are known, they can be introduced into the model to achieve better results. Instead of simply minimizing the difference between the real values and values predicted by the network, PINNS minimize the difference between the real and predicted values while conforming to the physics present in the problem as soft constraints [16].

This report studies the autonomous anomaly detection problem and proposes improvements to the regression and classification subproblems. For the regression portion, we study how including physics in the machine learning models used for signal reconstruction can benefit the signal reconstruction process. While different methods have been studied for the classification portion to determine the optimal anomaly threshold efficiently.

3 RELATED WORK

The regression subproblem in autonomous anomaly detection has been accomplished as a purely data-driven approach in previous work when applied to satellite data. The results published by Hundman et al. [8] and Jin et al. [11] performed regression analysis by training a machine learning model based on actual telemetry data from the satellites being studied. As mentioned before, Hundman et al. [8] trained a model based on Long Short-Term Memory (LSTM) layers to identify anomalies in real labeled telemetry data obtained from the *Mars Science Laboratory* and *Soil Moisture Active Passive* satellites. Jin et al. [11] trained a denoising autoencoder to detect anomalies in a satellite power subsystem.

The classification portion of this study builds on the results presented by Hundman et al. [8], as their labeled data are readily available for public use. Their classification work relies on nonparametric dynamic thresholding techniques to identify anomalies. In other words, their anomaly threshold is found that, if all values above are removed, it causes the most significant percent decrease in the mean and standard deviation of the reconstruction errors [8].

4 DATA PREPARATION

Two different datasets were used, each specifically tailored to demonstrate the proposed contributions to the regression and classification portions.

4.1 Regression

The data used for the signal reconstruction were obtained from the Advanced Composition Explorer launched on August 25, 1997, [17], which is referred to as the ACE dataset. More specifically, four-minute average magnetic field measurements taken by the onboard dual tri-axial magnetometer with respect to the spacecraft's geocentric solar ecliptic (GSE) coordinates were used. This dataset also contains position measurements in GSE coordinates. The data were filtered using an exponential moving average with a 500-point window and shifted such that all values were positive. Position measurements help to understand some of the physics that drive the magnetic field. Following Maxwell's laws and neglecting relativity contributions [6],

$$\nabla \cdot \mathbf{B} = 0 \tag{1}$$

That is,

$$\nabla \cdot \mathbf{B} = \frac{\partial B_x}{\partial x} + \frac{\partial B_y}{\partial y} + \frac{\partial B_z}{\partial z} = 0 \tag{2}$$

The divergence of the magnetic field must always be zero. This is known as Gauss's law of magnetism, which is expressed in a differential form and denies the existence of magnetic monopoles [9]. The divergence of the magnetic field can be calculated numerically with the experimental data using numerical differentiation, which can help determine whether the data follow Equation (2). In Figure 1, the filtered magnetic field measurements can be seen, and Figure 2 shows the divergence of the magnetic field measurements calculated by following Equation (2) and using numerical differentiation. Despite noise (which is likely to be introduced when performing numerical differentiation [13]), the magnetic field divergence approaches zero for all time steps, thus confirming the physics behind the magnetic field measurements. Therefore, this dataset aims to explore how including physics in a regression model's loss function can improve signal reconstruction.

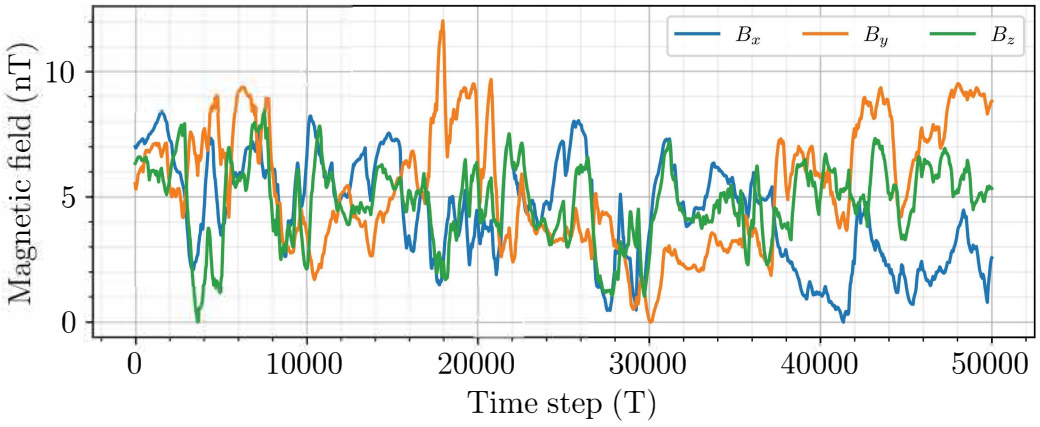

Figure 1: Magnetic field measurements filtered from the ACE dataset

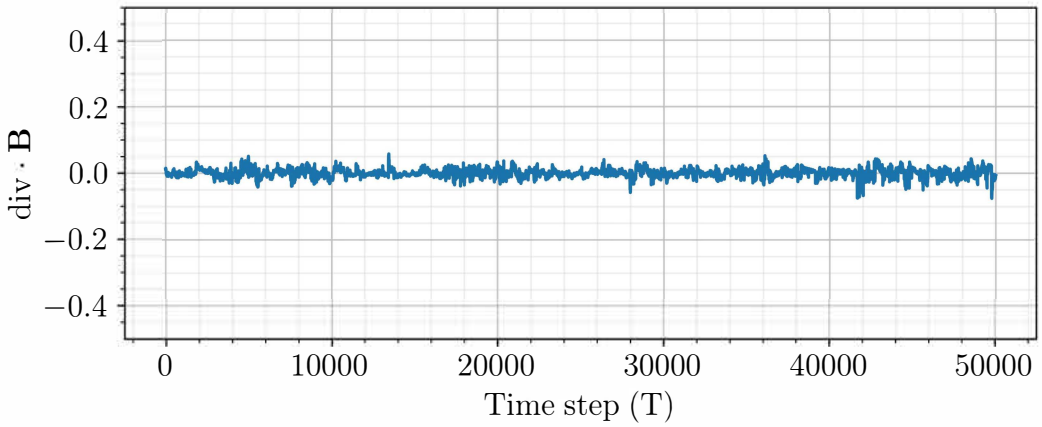

Figure 2: Divergence of magnetic field measurements from the ACE dataset

4.2 Classification

4.2.1 Real Anomalies.

The data used for the classification problem were obtained from the *Soil Moisture Active Passive* and the *Mars Science Laboratory* [8], hereafter referred to as the SMAP/MSL dataset. For each satellite, a different number of telemetry channels were included and contained a time series with one-hot encoded information for commands sent to each satellite module and the actual telemetry value of the channel. Because the data have already been scaled and preprocessed, no data filtering or scaling is necessary. For this dataset, the training samples include only nominal data, whereas the test samples have labeled anomalies that are useful for evaluating classification performance. In Figure 3, the training and testing time series for channel P-11 can be seen. It is important to note that the test dataset is labeled with the exact time step at which an anomaly starts and ends. Therefore, this dataset aims to explore different thresholding techniques that can improve the classification performance.

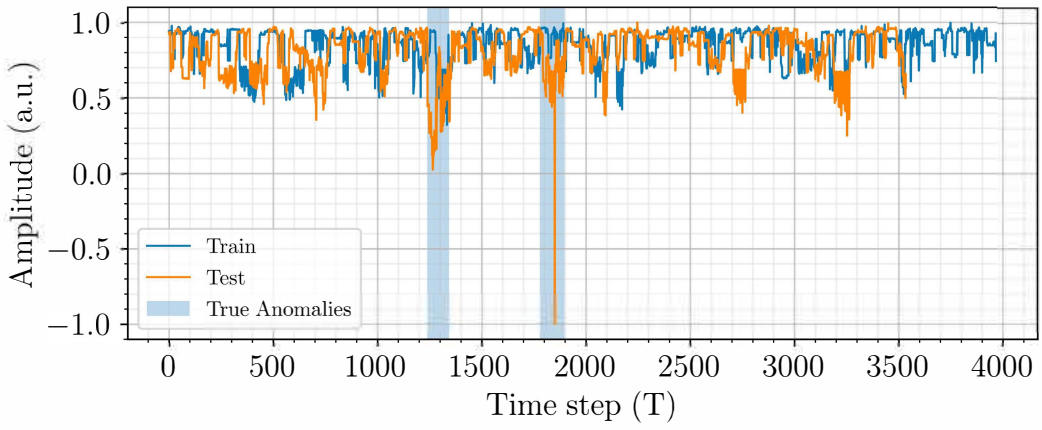

Figure 3: Training and testing time series for channel P-11 of the SMAP/MSL dataset

4.2.2 Synthetic Anomalies.

Different types of synthetic anomalies were introduced into the ACE test set to evaluate successful thresholding techniques on the ACE dataset. Specifically, 30 different test sets were created, containing random instances of the three types of anomalies in the magnetic field measurements. This new dataset will be referred to as the ACEC dataset. Figure 4 shows the process used to generate the ACEC dataset.

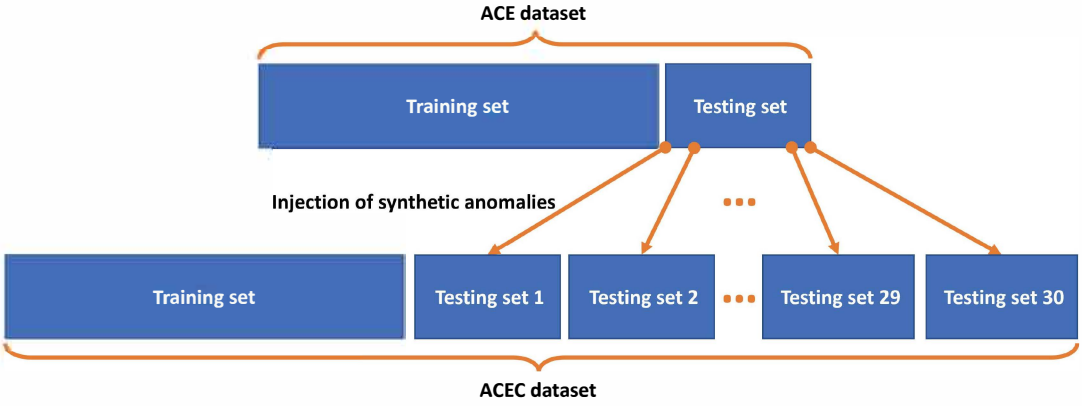

Figure 4: Process followed to create the ACEC dataset

The three types of anomalies are shown in Figures 5, 6 and 7. Figure 5 shows a periodic fault in the magnetic field, where the period and anomaly significance were randomly selected for all samples in the different test sets. Figure 6 shows a drift fault in magnetic field measurements. As with the periodic faults, the time at which the drift occurs and the anomaly significance were randomly selected for all samples in the test sets. Figure 7 shows the random faults in the magnetic-field measurements. For each test set, the time at which the random fault occurs, number of faults, and anomaly significance were kept random.

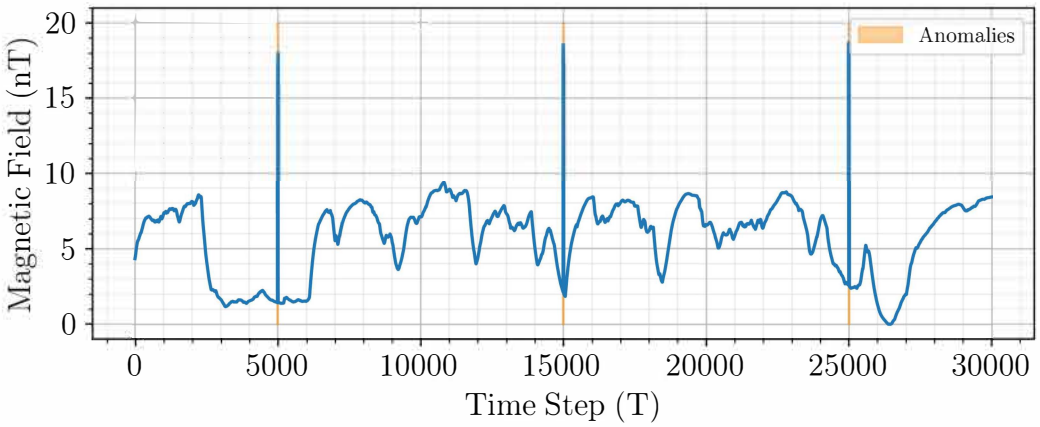

Figure 5: Example of a periodic fault introduced synthetically in the ACE testing data

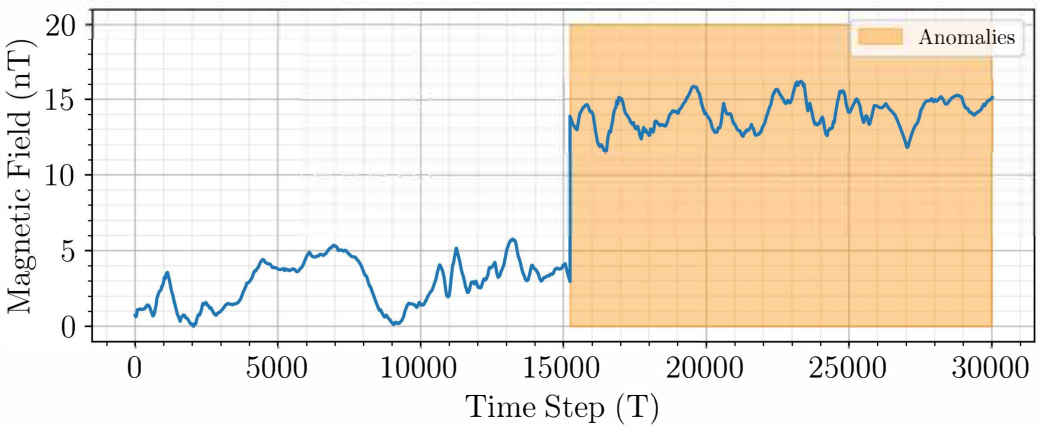

Figure 6: Example of a drift fault introduced synthetically in the ACE testing data

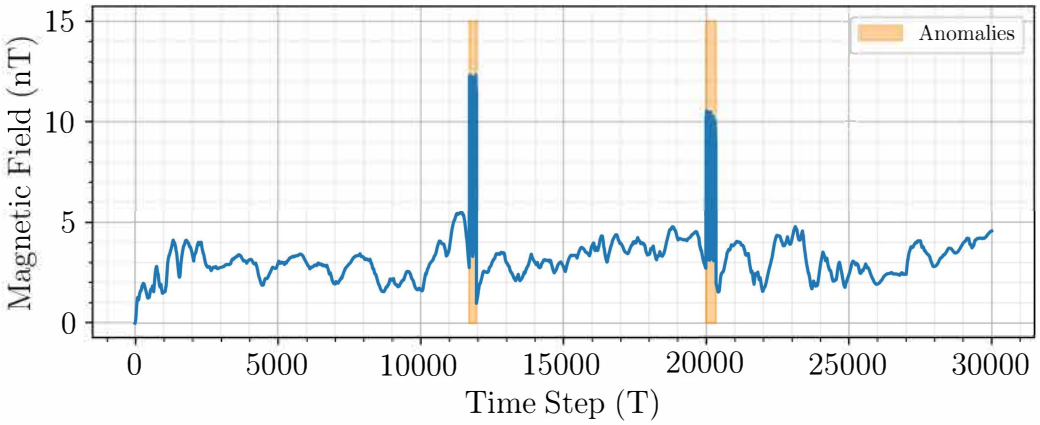

Figure 7: Example of random faults introduced synthetically in the ACE testing data

4.3 Summary of Data

To summarize, three different datasets were used, each with different purposes:

- **ACE dataset:** Consists of nominal magnetic field data and position measurements in GSE coordinates. Both the training and testing sets have nominal data since the purpose is to evaluate the regression performance of the trained models.
- **SMAP/MSL dataset:** Consists of telemetry data for each channel in the satellites discussed. The training set has only nominal data, whereas the testing set has labeled anomalies. The purpose of this dataset is to evaluate the classification performance using different thresholding techniques.
- **ACEC dataset:** Consists of nominal magnetic field data and position measurements in GSE coordinates for the training portion of the dataset. The test portion has synthetic

anomalies introduced randomly, as previously discussed. The purpose of this dataset is to test both the models trained with the ACE dataset and the thresholding techniques used for the SMAP/MSL dataset.

5 LEARNING FRAMEWORK

5.1 ACE dataset

In machine learning regression problems, a model is trained to predict a numerical value given a known input. The model is asked to predict or approximate a function f that maps the input to the desired output [5]. In this sense, signal reconstruction can be considered a regression problem. Several models can be used in the deep learning framework to solve this problem. A standard model for signal reconstruction problems is the autoencoder [4].

5.1.1 Autoencoders.

An autoencoder is a neural network trained to reconstruct its input and display it as an output (In some cases, the model is trained to copy the input to its output). Its architecture consists of two basic blocks: the encoder and the decoder. The encoder reduces the dimensionality of the input and maps it to the latent space, and the decoder handles the reconstruction of the signal from the latent space [5]. In the specific context of anomaly detection, the autoencoder is trained to reconstruct nominal signal data, such that if there is an anomaly in the signal, the autoencoder will not be able to reconstruct the anomaly.

5.1.2 Loss functions.

Different training loss functions can be used to train an autoencoder to reconstruct a signal. Most commonly, the loss functions used for signal reconstruction or enhancement are distance metrics between the reconstructed signal and the desired target [2]. With this in mind, a standard metric is the mean squared error:

$$\mathcal{L}_{\text{mse}} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 \tag{3}$$

Where N is the number of points, y refers to the desired target, and $\hat{y}$ is the predicted output. The use of these loss functions shows that traditional autoencoders are purely data-driven. If the physical properties of the signal that is being reconstructed are known, they can be implemented in the loss function as soft constraints to restrict the output [16]. These soft physical constraints can give the model more information to predict the correct output. With this, a new loss function can be used as follows:

$$\mathcal{L}_{\text{pi}} = \lambda_1 \mathcal{L}_{\text{mse}} + \lambda_2 f(\boldsymbol{x}, \hat{\boldsymbol{y}}) \tag{4}$$

Where f refers to the additional knowledge on the problem, which might be a function of both the input and the approximated output of the model, and λ_1, λ_2 are simply scaling factors.

The model used to reconstruct the magnetic field signal in the ACE dataset can be seen in Figure 8. This model consists of three different autoencoders whose task is to reconstruct the signal for each magnetic field component by inputting the position and the magnetic field component itself. The output of each autoencoder is concatenated to use a single loss function that includes all three reconstructed signals. Each dense layer contains a dropout of 0.2 to avoid overfitting and a hyperbolic tangent activation function, except the last one, which has a ReLU activation function. The loss function used to train the model can be seen below:

$$\mathcal{L}_{\text{pi}} = \frac{\lambda_1}{N}\sum_{i=1}^{N}\left(\mathbf{B}_i - \hat{\mathbf{B}}_i\right)^2 + \frac{\lambda_2}{N}\sum_{i=1}^{N}\left(\frac{\partial \hat{B}_{x_i}}{\partial x_i} + \frac{\partial \hat{B}_{y_i}}{\partial y_i} + \frac{\partial \hat{B}_{z_i}}{\partial z_i}\right)^2 \tag{5}$$

Where $\boldsymbol{B} = [B_x, B_y, B_z]^T$ refers to the true magnetic field and $\hat{\boldsymbol{B}} = \left[\hat{B}_x, \hat{B}_y, \hat{B}_z\right]^T$ is the reconstructed magnetic field. The first term is the data mean squared error reduced over the number of components and the second term refers to the $\mathcal{L}_2$ physics regularization. Having the position as part of the input to the model allows the calculation of the predicted magnetic field divergence through automatic differentiation. Automatic differentiation is a computational technique for efficiently and accurately evaluating derivatives of numeric functions without facing issues with limited precision due to approximation errors as in numerical differentiation [1].

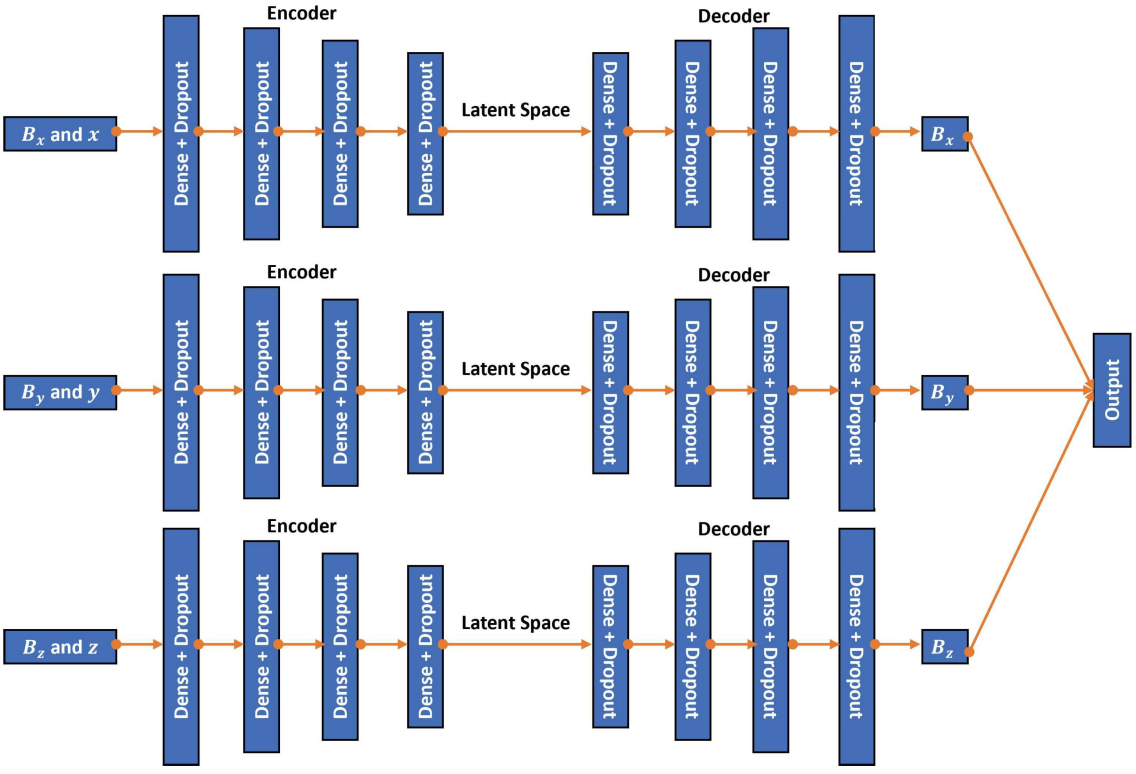

Figure 8: Model used to reconstruct the magnetic field signal in the ACE dataset

5.1.3 Bias-Variance Trade-off.

In statistical learning, the mean squared error of a prediction is defined as [10]:

$$\text{MSE}\,(\hat{y}) = \text{E}\left[(y - \hat{y})^2\right] = \text{Var}\,(\hat{y}) + \text{Bias}^2\,(\hat{y}) + \text{Irreducible Error} \tag{6}$$

As a general rule, high complexity models can be prone to overfitting the training data, increasing the prediction's variance while decreasing its bias. As seen in equation (6), the MSE depends on these two quantities' relative rate of change. Thus, as a model's complexity increases, the prediction bias decreases faster than the variance rises, decreasing the MSE. However, at a specific point, increasing the model's complexity has little impact on the bias but significantly increases the variance, increasing the MSE. To restrain the complexity of a model, regularization techniques are used, which prevent the model from becoming too complex [10]. If a regularization term is introduced in the model's loss function, as seen in equation (5), the MSE will generally follow the behavior described in figure 9. This is known as the bias-variance trade-off.

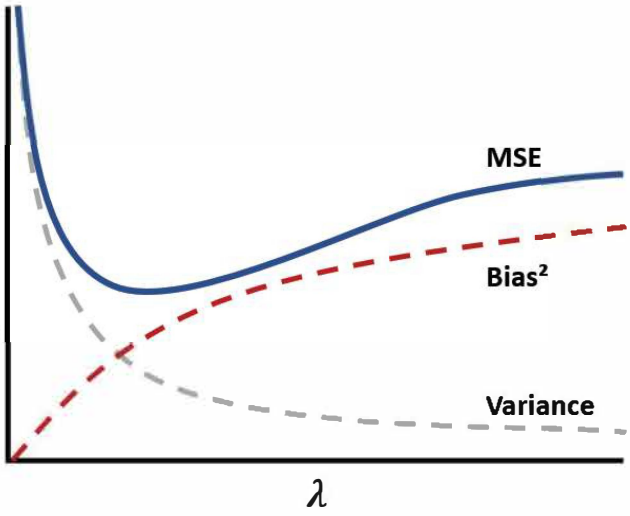

Figure 9: Bias-variance trade-off as a function of the regularization weight, as λ increases, the model decreases its complexity. Figure adapted from James et al. [10]

5.1.4 Curriculum Regularization.

The term curriculum regularization was first introduced in 2021 by Krishnapriyan et al. [14] when characterizing the possible failure modes in physics-informed neural networks. With this technique, the loss term in the neural network starts from a simple partial differential equation regularization and becomes more complex as the training advances. Curriculum regularization allows a warm start to the neural network training by finding a proper set of initial weights [14]. Based on this idea, this report implements a similar technique where the importance/weight of the physics term in the loss function starts to increase gradually as the model is being trained.

5.2 SMAP/MSL dataset

Once the signal reconstruction model has been implemented, identifying anomalies becomes a classification problem. The classification problem is responsible for classifying the reconstruction error into nominal behavior and potential anomalies. Therefore for this dataset, a model based on long short-term memory networks was used to obtain the reconstruction error.

5.2.1 Long Short-Term Memory Networks.

Recurrent networks use Long Short-Term Memory layers that can use feedback connections to take into account past representations of recent inputs in the form of activation neurons [7]. In Figure 10, the sequential network architecture used for the SMAP/MSL dataset can be seen. This is the same architecture presented by Hundman et al. [8], which has proven to be sufficiently good at reconstructing the signal. In this case, the input is not the entire time series. Instead, the input is a time window of the time series (specifically 250-time steps) [8], and the output is the prediction of the following ten-time steps. The output can then be compared to the actual telemetry data by averaging the prediction, which works as a moving average filter. Since the dataset includes contextual anomalies rather than point anomalies, the LSTM layers help identify these anomalies [8]. As this dataset's physics is unknown, the mean squared error loss was used to train a different model, following the same architecture shown in Figure 10, for each channel. It is important to note that the purpose of this dataset is to classify the error between the reconstructed output and the ground truth. Considering that the regression model returns the error between the reconstructed values and the real signal, an error threshold must be set to know when an anomaly has been detected. This threshold establishes the limit between the nominal behavior and the anomalies in the signal. Different techniques can be used to establish the anomaly threshold. In this work, two different alternatives have been studied.

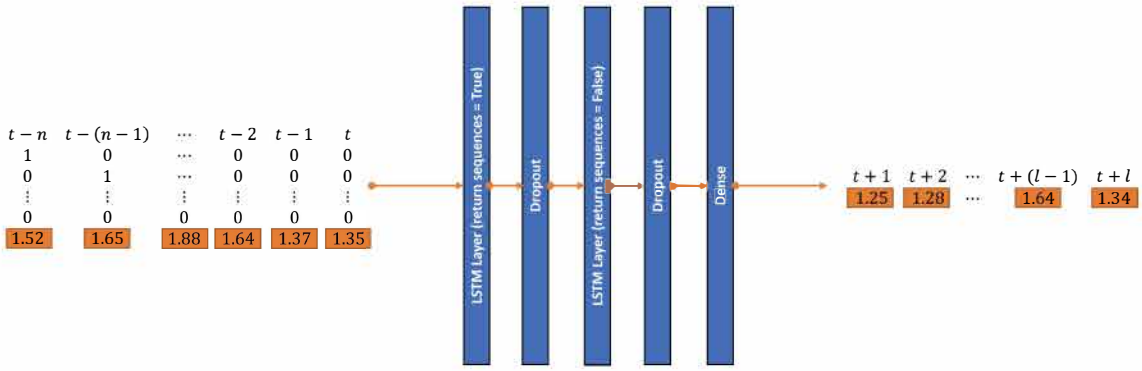

Figure 10: Network architecture used for the SMAP/MSL dataset. Orange boxes represent the telemetry value at each time step. Figure adapted from Hundman et al. [8]

5.2.2 Gaussian Assumption.

The Gaussian assumption method is based on the premise that the error between the predicted values and the ground truth follows a Gaussian distribution. Therefore, the anomaly threshold

(ξ) is set as [3]:

$$\xi = \mu(\text{test MAE}) + \alpha\sigma(\text{test MAE}) \tag{7}$$

Where μ is the mean, σ is the standard deviation, α is a scaling parameter, and MAE refers to mean absolute error. If the testing MAE is higher than the threshold at a given point, the model can flag this point as an anomaly. A high α will produce true positives but might return an increased number of false negatives, while a low α will result in an increased number of true positives and a large amount of false positives. Usually, the Gaussian assumption will not hold in practice [18], so other models can be used to find the necessary threshold.

5.2.3 K-means clustering.

A different method to find the anomaly threshold is the K-means clustering algorithm. K-means clustering is an algorithm that divides a dataset into K distinct, non-overlapping clusters [10]. In this case, the reconstruction error is clustered into two groups representing nominal behavior and anomalies. To understand how K-means clustering works, let C_n and C_a be the clusters representing nominal behavior and anomalies, respectively. Now let $\sigma(C_n), \sigma(C_a)$ denote the standard deviation of the reconstruction errors inside each cluster. Therefore, the best possible cluster configuration will be the one that minimizes:

$$W(C_a, C_n) = \sigma(C_a) + \sigma(C_n) \tag{8}$$

The anomaly threshold (ξ) is set as the lowest reconstruction error in the anomalies cluster [11].

5.2.4 Pruning procedure.

Once the anomaly threshold has been set, a pruning procedure in the predicted anomalies must be done. If the anomaly threshold is set too low, in addition to predicting real anomalies, it is possible to set a high rate of false positives. Therefore, the pruning procedure takes care of dealing with false positives. In this work, the pruning procedure identifies all the anomalies predicted, calculates the difference (ϕ) between their reconstruction error and the anomaly threshold, and calculates the ratio between ϕ_{max} and the rest. Then, a pruning threshold (γ) can be set to prune the anomalies if the earlier ratio is below the pruning threshold.

In Figure 11, the pruning procedure used can be seen. The predicted anomalies include four false positives since the error signal exceeds the threshold five times. The pruning procedure eliminates the false positives and gives only the true positive. It is essential to mention that the pruning procedure will not eliminate all the false positives every time. Regardless, different techniques such as grid search, Bayesian search, and coarse to fine random search can be used to find a pruning threshold that maximizes a desired metric. Finding a pruning threshold by maximizing a metric is only valid if the dataset is labeled so quantities such as true positives, false negatives, and false positives can be calculated.

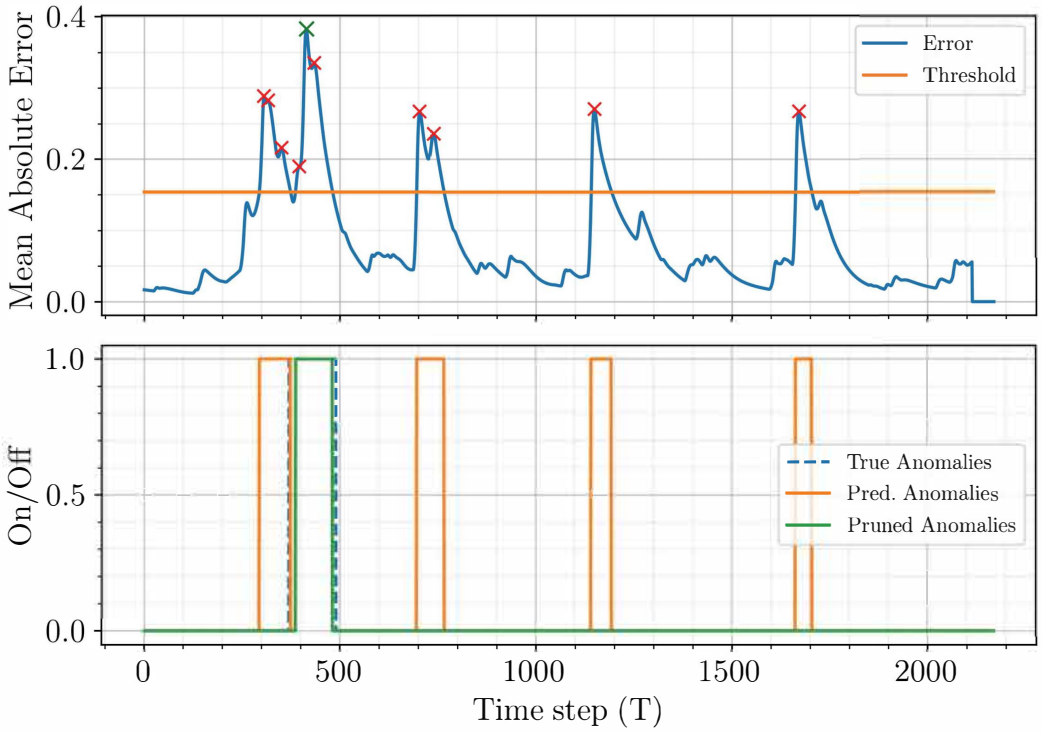

Figure 11: Reconstruction error (Top) between the real and predicted telemetry value for channel T-12. Comparison (Bottom) between true, predicted and pruned anomalies for T-12 in the SMAP/MSL dataset

5.3 ACEC dataset

In this case, the model presented in figure 8 was used for the reconstruction of the magnetic field. To classify the reconstruction error as nominal or anomalous, the techniques discussed in section 5.2 were used.

6 TRAINING AND TESTING

6.1 Training Hyperparameters

To train the ACE model (Figure 8), the dataset was divided into a training set containing 90,000 data points and a test set with 30,000 data points. The Adamax optimizer [12] with an initial learning rate of 0.01 and a decay rate of 0.0001 was used. The model was trained for 100 epochs with a batch size of 500, a validation split of 0.2, and an early stop callback monitoring validation loss with a patience of ten epochs.

The SMAP/MSL model (Figure 10) was trained independently for each channel in each dataset. The Adamax optimizer [12] with an initial learning rate of 0.001 was used. The model was trained for 35 epochs with a batch size of 30, a validation split of 0.1, and an early stop callback monitoring validation loss to avoid overfitting.

6.2 Metrics

To test the ACE model's ability to reconstruct the signal, the standard MSE metric described in Equation (3) was used. With this, the distance between the reconstructed and real signal on the test set was calculated for different model configurations.

To test different thresholding techniques, the following metrics were used [10]:

$$\text{Precision } (P) = \frac{TP}{TP + FP} \tag{9}$$

$$\text{Recall } (R) = \frac{TP}{TP + FN} \tag{10}$$

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{P \cdot R}{\beta^2 \cdot P + R} \tag{11}$$

Where TP refers to true positives, FP to false positives, FN to false negatives, and β is a scaling factor. A lower β value gives more weight to precision and less to recall, whereas a larger β gives less weight to precision and more weight to recall. With these three metrics, the model's performance can be obtained, which can be helpful for fine-tuning and threshold searches. Having introduced these concepts, it is essential to note that a pruning procedure can only improve the precision of the model, and in some cases, it might decrease recall. For the SMAP/MSL and ACEC datasets, the true positives, false positives, and false negatives were recorded as follows [8]:

- A true positive is recorded if any portion of the predicted anomalies overlaps any true anomaly. Only one true positive is recorded if multiple portions of the predicted anomalies fall within a true anomaly.
- All predicted anomalies that do not overlap with any true anomaly are recorded as false positives.
- If no portion of the predicted anomalies overlaps a true anomaly, a false negative is recorded.

7 RESULTS AND DISCUSSION

7.1 Regression

7.1.1 Physics Regularization.

For the signal reconstruction of the magnetic field, the ACE model was first trained by varying the loss function scaling factor λ_2 and keeping $\lambda_1 = 1$ constant. A low value of λ_2 refers to a data-driven model, while a high value of λ_2 describes a model with a high physics regularization. Figure 12 shows the average mean squared error between the magnetic field from the test set and the reconstructed signal as a function of the parameter $\lambda_2 \in [0, 10]$. The model was

trained ten different times, starting from different initial weights. As it can be seen, there is a tendency for the error to decrease as the weight of physics in the loss function increases. This demonstrates that the model benefits from the constraints induced by including a physics-based regularization term in the loss function. The lowest error is found at $\lambda_2 = 2.0$. As with standard regularization methods [10], the mean squared error reaches a point where it stops decreasing (in this case, at $\lambda_2 = 2.0$) and adopts an increasing trend as λ grows from this point on. Along with reducing the test MSE, including a physics-based regularization term in the training loss function reduces the variance of the results obtained. As the test MSE is closely related to the variance plus the bias squared [10], it can be seen from Figure 12 that increasing the physics weight lowers the variance, thus lowering the overall test MSE.

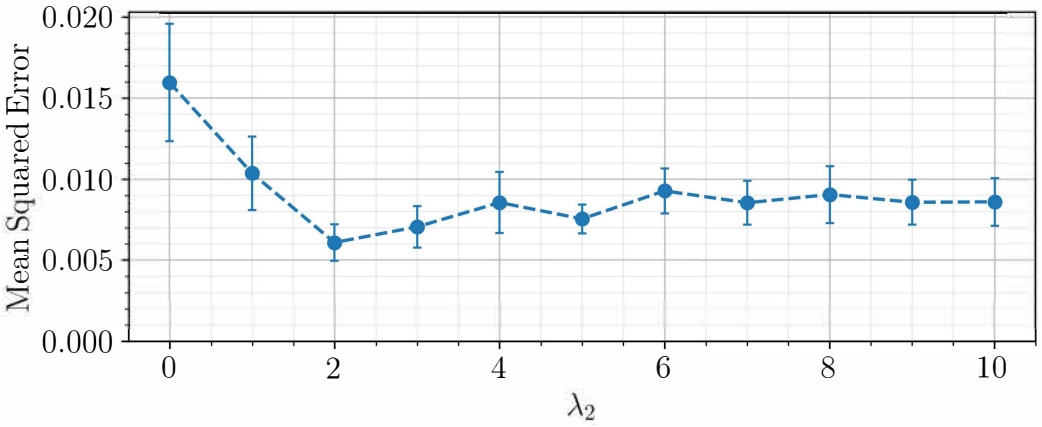

Figure 12: Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\lambda_2 \in [0, 10]$

Figure 13 shows the average mean squared error between the magnetic field from the test set and the reconstructed signal as a function of $\lambda_2 \in [10^1, 10^{10}]$. With this figure, it is easier to see the increasing trend that the test MSE adopts as λ_2 grows beyond its optimal point. The increase found for large values of λ shows that a purely physics-based model will not perform as well compared to a mixed model. It is important to note that there is an increase in variance for very high values of λ_2, which could be because the physics used in the regularization term is not enough to reconstruct the magnetic field.

Figure 14 shows the average total training epochs as a function of λ_2. Since the training scheme included an early stop callback monitoring validation loss with a patience of ten epochs, a higher number of total training epochs means that the model avoids overfitting the data at the early stages of training. For $\lambda_2 = 0$, which means that the model is purely data-driven, the training is stopped, on average, before it reaches 30 epochs as the model starts overfitting the data at around this point. As the weight of the physics-based regularization term increases, it can be seen from the figure that the average total training epochs adopts an increasing tendency. Thus, including the physics in the loss function helps the model avoid overfitting the training data as expected. It is important to note that training for more epochs will not guarantee a lower test MSE, as a high regularization might lead to an under-fitted model due to the bias-variance trade-off [10].

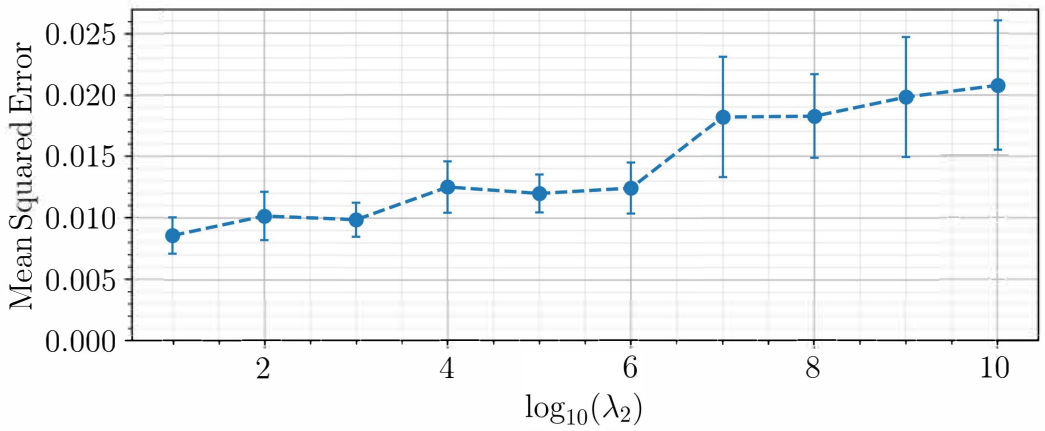

Figure 13: Average MSE between the reconstructed magnetic field and the original signal from the ACE test set as a function of $\lambda_2 \in [10^1, 10^{10}]$

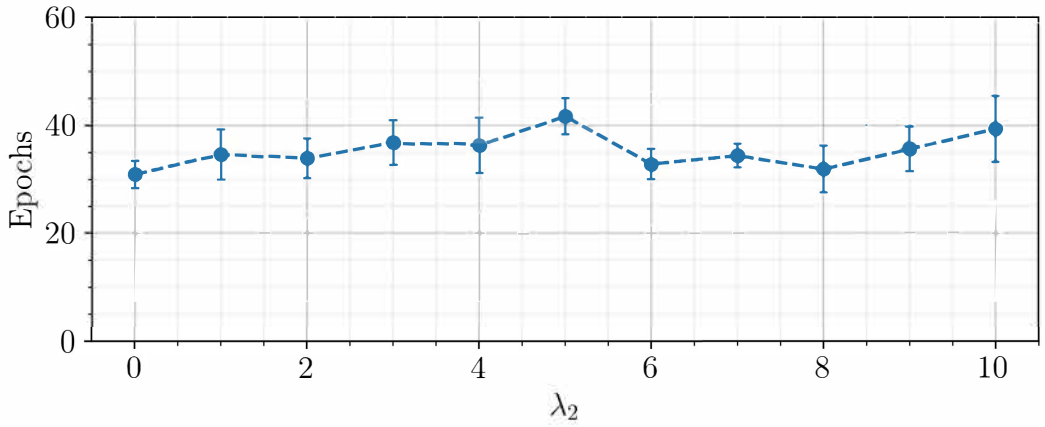

Figure 14: Average number of epochs as a function of $\lambda_2 \in [0, 10]$

Figure A-1 shows the real and averaged reconstructed signals by the lowest error model ($\lambda_2 = 2$ according to Figure 12) for the three components of the magnetic field in the ACE dataset. This reconstruction refers to data that the model has never seen in training. Therefore, it can be established that the model successfully reconstructs the signal as it closely follows the ground truth.

7.1.2 Curriculum Regularization.

As mentioned before, curriculum regularization was used to train the model while gradually increasing the weight of the physics term in the loss function (λ_2). Ten models were trained starting from different starting weights and varying the physics weighting term from zero to six. This range of λ_2 yielded better results, although it is clear that the range of λ_2 is case-dependent. The weight of the physics-based regularization term was varied in increments of one every fifteen epochs of training. Figure A-2 shows the real and averaged reconstructed

signals for the three magnetic components using the curriculum regularization scheme mentioned before. As it can be seen, the reconstructed signals in this figure follow more closely the real values when compared to Figure A-1.

Table 1 shows the average reconstruction mean squared error between the real and reconstructed signal for the two different training techniques. The results obtained show that the curriculum regularization training scheme achieves a lower reconstruction error when compared to the physics regularization used in the previous section. The improvement found shows that slowly increasing the weight of the regularization portion in the loss function can help achieve a lower reconstruction error which fits with previous results reported in literature [14]. Still, characteristics such as determining the model's sensitivity to the rate of increase of the regularization term or finding the optimal range of the regularization term are questions that need to be answered but lay beyond the scope of this work.

Table 1: Comparison between the best-averaged physics informed model ($\lambda_2 = 2$) and the averaged model trained by curriculum regularization

Training Technique	Mean Squared Error
Physics Regularization	0.00609 ± 0.00113
Curriculum Regularization	0.00324 ± 0.00059

Having shown that the inclusion of physics in the model's loss function can help reconstruct the desired signal, the classification procedure was studied and analyzed.

7.2 Classification

7.2.1 Real Anomalies.

As mentioned above, two different thresholding techniques were studied for the classification procedure. Once the SMAP/MSL model was trained for signal reconstruction, the Gaussian assumption technique was first used to classify the reconstruction error as anomalous or nominal. Using $\alpha = 1$, the anomaly threshold was found, as discussed before. Having determined the anomaly threshold for each channel, a line search was performed to find the optimal pruning threshold to maximize the $F_{0.5}$ score.

In Figure 15 the line search for the pruning threshold in channels T-9 and F-8 is shown. First, it can be seen that the pruning procedure for the F-8 channel improves the classification accuracy of the gaussian assumption classifier since the $F_{0.5}$ score increases as the pruning threshold grows. The optimal pruning threshold for this channel (defined as the lowest threshold that maximizes the $F_{0.5}$ score) lies close to 0.6, meaning that only the highly significant anomalies are being considered, and the less important anomalies (possibly false positives) are being pruned. The pruning procedure for the T-9 channel brings no improvement to the $F_{0.5}$ score. Since the pruning procedure takes care of eliminating false positives, seeing no improvement in the line search simply means that the initial anomaly threshold results in no false positives. The behavior discussed previously can be seen more clearly in Figure 16.

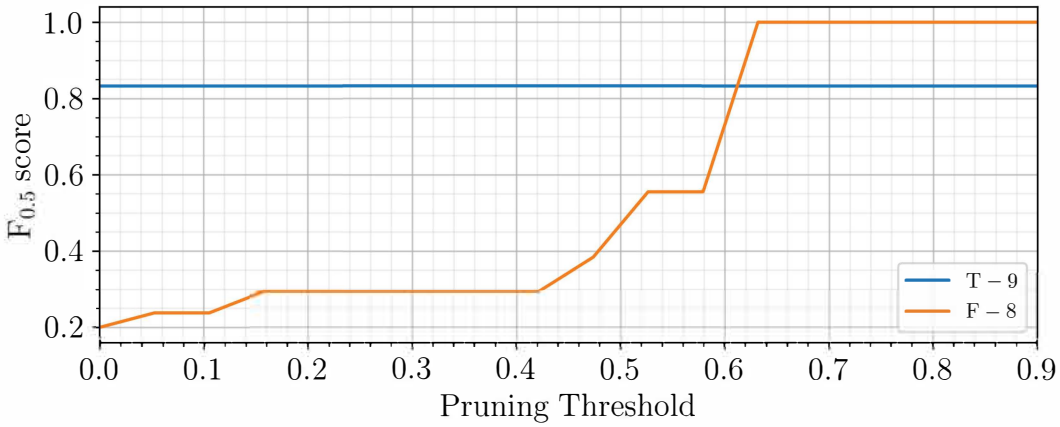

Figure 15: $F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the gaussian assumption

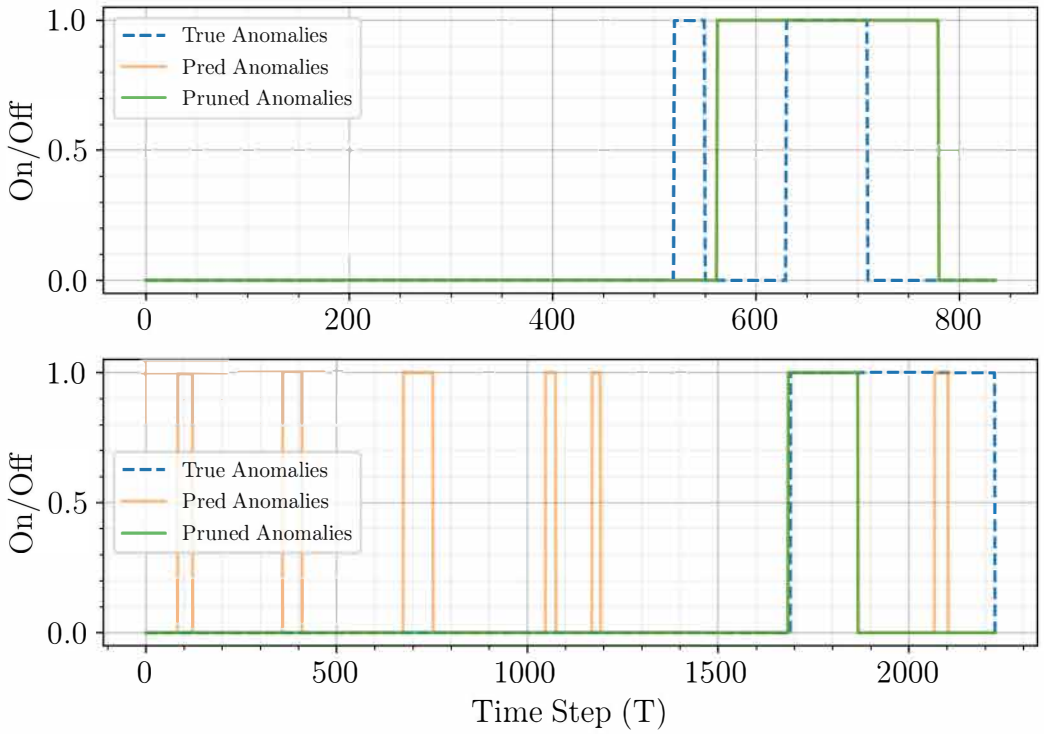

Figure 16: True, predicted, and pruned anomalies for channel T-9 (top) using the Gaussian assumption threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the Gaussian assumption threshold

Figure 16 shows the true, predicted and pruned anomalies for the channels T-9 and F-8. In channel T-9, the initial anomaly threshold sets no false positives, and the pruned anomalies are the same as the initially predicted anomalies. Setting no false positives means that the

anomaly threshold is set correctly to maximize precision. Regardless, the anomaly threshold set for channel T-9 fails to predict one true anomaly, suggesting that recall can still be improved using other thresholding techniques. Channel F-8 shows the importance of pruning after predicting anomalies. As shown, while the predicted anomalies include the true anomaly, multiple false positives are also present. The pruning procedure eliminates these false positives, given that the most significant reconstruction error is found at the true anomaly. If the true anomaly is not associated with the most significant reconstruction error, the pruning procedure may decrease recall, which shows the importance of the regression portion of the autonomous anomaly detection problem. This process was done for the rest of the channels in the SMAP/MSL dataset, where the global results can be seen in table 2.

Following the second technique, K-means clustering was used to find the anomaly threshold. Figure 17 shows the K-means clustering classification technique. The algorithm divides the reconstruction error into two clusters, one for nominal behavior and one for anomalies. The anomaly threshold can be found by selecting the lowest error in the anomalies cluster. Thus, a base anomaly threshold was obtained by performing this procedure for each channel in the SMAP/MSL dataset. Once again, having determined the base anomaly threshold for the K-means clustering classifier, a line search was performed to find the best pruning threshold for each channel.

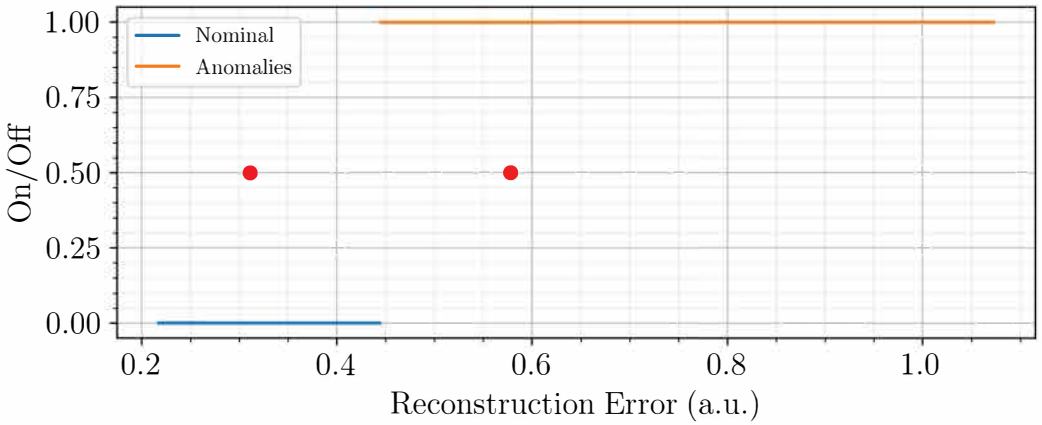

Figure 17: Error classification using K-means clustering for an arbitrary channel (E-11)

Figure 18 shows the line search for the pruning threshold regarding the K-means clustering classifier in channels T-9 and F-8. As with the Gaussian assumption model, the $F_{0.5}$ score for the F-8 channel increases as the pruning threshold grows. In this case, the optimal pruning threshold is higher when compared to the Gaussian assumption model. A higher pruning threshold means the initial anomaly threshold is set at a lower value. This is also true for channel T-9; as the anomaly threshold is set lower, the classifier can obtain a perfect score by predicting all true anomalies while still having no false positives. As with the Gaussian assumption, the K-means classifier predicts no false positives for channel T-9, meaning that the pruning procedure results in no improvement when considering the $F_{0.5}$ score.

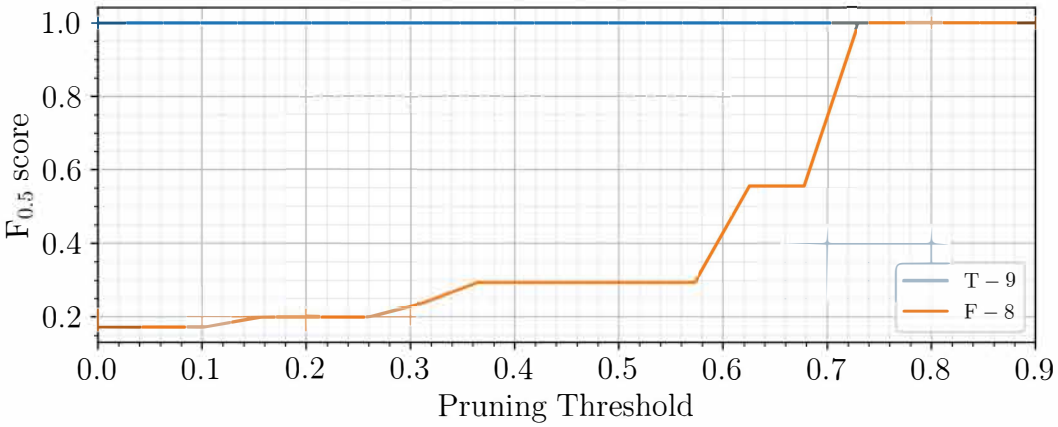

Figure 18: $F_{0.5}$ score as a function of the pruning threshold for the predicted anomalies found using the K-means clustering algorithm

Figure 19 shows the pruning procedure in the two channels discussed above. In the same way as the Gaussian assumption model, the initial anomaly threshold found with the K-means model for channel T-9 sets no false positives. The pruned anomalies are the same as the predicted anomalies. However, the anomaly threshold set by the K-means model can identify all true anomalies in the signal, meaning that for this specific case, the K-means thresholding tech-nique is more appropriate. For channel F-8, it can be seen that the predicted anomalies found with the K-means thresholding technique result in a higher number of false positives than with the Gaussian assumption technique. The pruning procedure removes these false positives and achieves a perfect score. Once again, this is only possible since the highest reconstruction error found by the regression portion is precisely at the true anomaly.

Table 2 summarizes the generalized $F_{0.5}$ scores for the model configurations presented above. To calculate each $F_{0.5}$ score in this table, the total number (including every channel) of true positives, false positives, and false negatives for each satellite was found, and Equation (11) was used to calculate the resulting score. This table shows that the K-means clustering tech-nique obtains a better $F_{0.5}$ score in both satellites being studied compared to the Gaussian assumption technique. These results show that, as mentioned before, reconstruction errors will not always follow a Gaussian distribution, leaving other thresholding techniques better suited for anomaly detection. Regardless, straightforward techniques such as line searches have been shown to achieve satisfactory results.

Table 2: Comparison between the different thresholding techniques. The first score refers to SMAP and the second score to MSL

Thresholding Technique	**$F_{0.5}$ score** (SMAP)	**$F_{0.5}$ score** (MSL)
Gaussian assumption	0.843	0.781
K-means clustering	0.869	0.938

Having encountered the best configurations for the classifiers, the results were compared with the initial results obtained by the original paper [8] from which the dataset was published. As

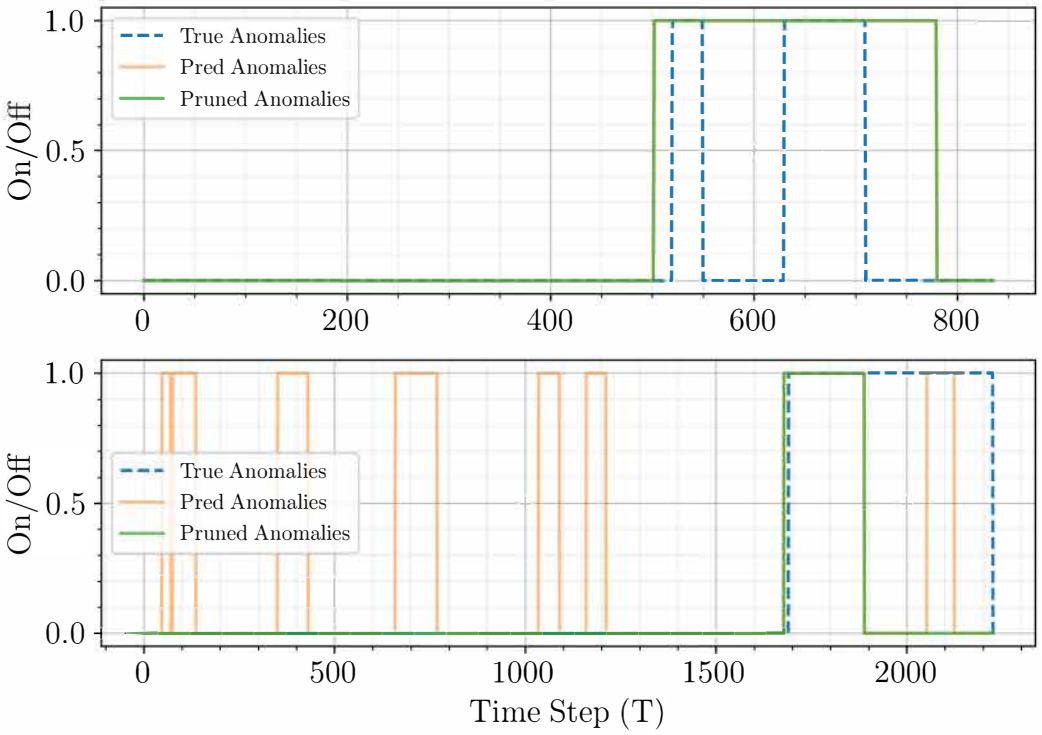

Figure 19: True, predicted, and pruned anomalies for channel T-9 (top) using the K-means clustering threshold. True, predicted, and pruned anomalies for channel F-8 (bottom) using the K-means clustering threshold

mentioned before, the SMAP/MSL model used in this paper for the signal reconstruction is the same as the one presented by Hundman et al. [8], which allows a direct comparison of the thresholding techniques. In Figure 20, the $F_{0.5}$ score for the original paper's classification model, which uses a non-parametric dynamic thresholding technique, the Gaussian assumption classification model, and the K-means clustering classification model is shown. This figure shows that the original paper outperforms the Gaussian approach. In contrast, the K-means clustering technique outperforms the original article in both satellites. With this, it can be seen that out of the three thresholding techniques considered, K-means clustering accompanied by the pruning procedure described before has the highest $F_{0.5}$ score.

7.2.2 Synthetic Anomalies.

Since using K-means clustering as a thresholding technique yielded the best results in the real anomalies, this technique was used to classify the synthetic anomalies in the ACEC dataset.

For this purpose, the ACE model was trained on nominal data using a purely data-driven training scheme, the physics regularization technique, and the curriculum regularization scheme discussed before. These three training techniques were chosen to compare the difference between including and not including physics in the classification context of anomaly detection. In this case, the two best physics-informed model were selected, reason why the model trained

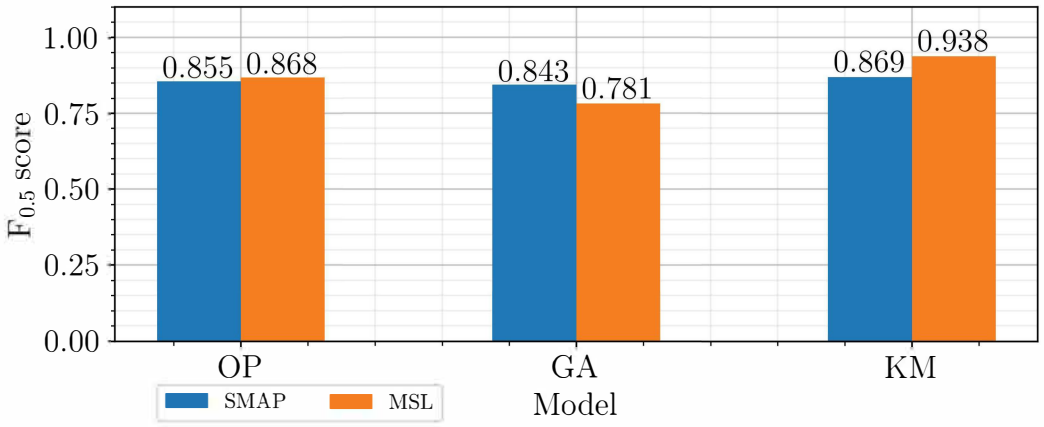

Figure 20: $F_{0.5}$ score for the *Soil Moisture Active Passive* (SMAP) and *Mars Science Laboratory* (MSL) using the original paper's classification model [8] (OP), the Gaussian assumption classification model with pruning (GA) and the K-means clustering classification model with pruning (KM)

with $\lambda_2 = 2$ and trained with curriculum regularization were compared. Having trained the models, a reconstruction error was obtained for each of the three models using the ACEC test set where the synthetic anomalies are present. With this, K-means clustering was used to classify the reconstruction error as nominal or anomalous, and the predicted anomalies were compared with the true anomalies in the test set.

Table 3 shows the precision, recall, and F_1 score for each training technique implemented for this data set. It is important to mention that up to this point, no pruning has been done on the predicted anomalies, meaning that the classification metrics shown in this table correspond to unsupervised clustering. The results include uncertainty since the regression model was trained ten times, starting from different random weights, thus resulting in ten different models being evaluated. Therefore, the results reported show the average precision, recall, and F_1 score.

Table 3: Comparison between the performance metrics obtained with the different train-ing techniques on the ACEC test set

Training Technique	Precision	Recall	F_1 Score
Data-Driven	0.548 ± 0.043	0.943 ± 0.009	0.684 ± 0.037
Physics Regularization	0.571 ± 0.053	0.935 ± 0.008	0.693 ± 0.042
Curriculum Regularization	0.570 ± 0.054	0.943 ± 0.002	0.696 ± 0.040

From the results presented, it can be seen that recall is high for the data-driven technique, the physics regularization technique, as well as the curriculum regularization scheme. Having a high recall shows that both models can predict most real anomalies in the ACEC test set. It is important to note that the physics regularization technique has a lower recall than both the data-driven model and the model trained using curriculum regularization. In the case of precision, all training techniques achieve a low value compared to the recall value. There is a

significant increase in precision when comparing the curriculum regularization scheme with the purely data-driven model. Increasing precision might indicate that including the physics in training can help the model learn the nominal behavior better as it is more capable of discerning between real anomalies and strange nominal behavior.

To evaluate the generalization of the pruning procedure for unseen anomalies, the ACEC test set was divided into two subsets. As seen in figure 21, the ACEC test set was divided into a determination set and an evaluation set, both of the same sizes. The determination set was used to establish a correct pruning threshold using the same technique as in section 7.2.1. In contrast, the evaluation set was used to assess the performance of the pruning threshold on new unseen data.

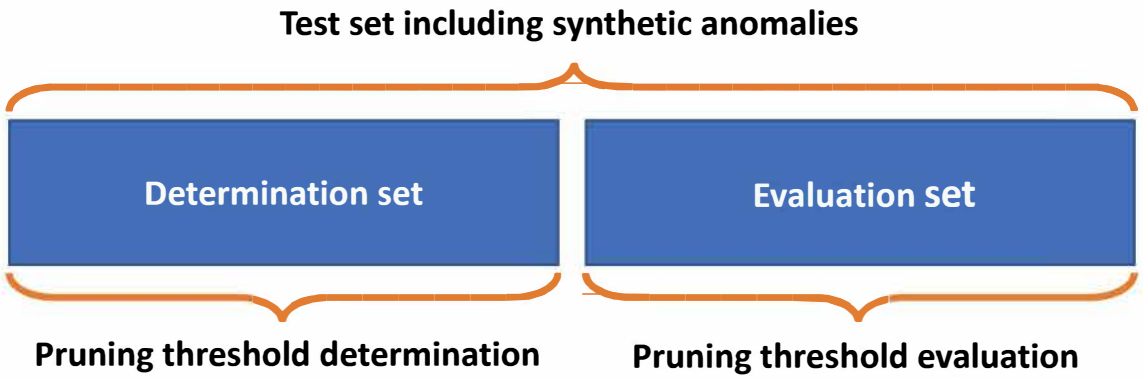

Figure 21: Test set division to determine and evaluate the pruning threshold

In table 4 the performance metrics for the data-driven model evaluated on the evaluation set can be seen. These performance metrics were calculated for the results obtained with and without the pruning procedure to judge the efficiency of the pruning threshold applied to new data.

Table 4: Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for the purely data-driven model

Pruning	Precision	Recall	F_1 Score
No	0.634 ± 0.041	0.939 ± 0.007	0.750 ± 0.031
Yes	0.776 ± 0.043	0.924 ± 0.009	0.837 ± 0.027

From table 4 several things can be seen. First, as expected, including the pruning procedure increases the precision of the predicted anomalies. More specifically, the pruning procedure increases precision by around 14%. This suggests that the pruning procedure described throughout this work successfully prunes false positives from unseen data. But, it is important to mention that, as discussed before, the pruning procedure might decrease recall. In this case, the recall decrease is around 1% when the pruning procedure is implemented. Regardless, the precision increase is significantly higher than the recall decrease, thus resulting in a higher F_1 score. These results show that the designed pruning procedure can be implemented in new data, allowing a flexible pruning threshold that, which instead of fixing a static threshold, determines the right anomaly power for a possible anomaly to be considered a true positive. It is important to

note that there is an increase in precision without using the pruning procedure compared to the results presented in table 3. The increase in precision in the unsupervised context is due to the fact that the results shown in table 4 refer only to the data in the evaluation set, while the results in 2 were calculated using the entirety of the ACEC test set. The same is true for tables 5 and 6.

In table 5 the performance metrics for the physics regularized model evaluated on the evaluation set can be seen.

Table 5: Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for the physics regularization model

Pruning	Precision	Recall	F_1 Score
No	0.680 ± 0.051	0.938 ± 0.004	0.777 ± 0.036
Yes	0.808 ± 0.036	0.929 ± 0.004	0.860 ± 0.023

As with the data-driven model, the pruning procedure increases precision in the physics regularized model. In this case, the increase in precision is about 13%. The decrease in recall is low, at around 1%. Thus the same behavior is seen for the physics regularized model, where with the pruning procedure, the precision of the model is increased significantly while only slightly decreasing recall. This results in a higher F_1 score for the model including the pruning procedure, showing that it generalizes well to unseen data.

Table 6 shows the performance metrics for the model trained with the curriculum regularization training scheme evaluated on the evaluation set.

Table 6: Comparison between the performance metrics obtained with and without the pruning procedure on the evaluation set for curriculum regularization model

Pruning	Precision	Recall	F_1 Score
No	0.706 ± 0.039	0.930 ± 0.002	0.797 ± 0.024
Yes	0.793 ± 0.029	0.925 ± 0.003	0.851 ± 0.017

In this case, it is important to note that the pruning procedure reduced recall only by 0.5%. This means that most true anomalies are associated with the highest reconstruction error, which shows, once again, that including the physics while training the reconstruction model helps to learn the nominal behavior better. As well as this, it is important to note that the recall of the model, including physics, is higher than that of the purely data-driven model (as seen in table 4). The pruning procedure helps increase precision without reducing recall. In this case, there is a 9% increase in precision. When compared to the data-driven model, the increase is less. Still, the absolute precision of the model trained with curriculum regularization is higher than the precision of the data-driven model. This also suggests that including the physics while training the reconstruction model on nominal data helps the model learn nominal patterns better. The increase in precision without decreasing recall using the pruning procedure results in a higher F_1 score for the model trained with curriculum regularization than the data-driven model.

In figure 22 a summary of the results obtained for the three models and the pruning procedure is shown. The results obtained with the physics regularized model and the model trained

using curriculum regularization are comparable as both achieve very similar F_1 scores. While the model trained with curriculum regularization achieves a better score than the physics regularized model when there is no pruning, the physics regularized model performs better when applying the pruning procedure. Regardless, the pruning procedure improves the scores for both models, showing that the technique applied can work on unseen data. The F_1 score obtained with the data-driven model is lower than the scores achieved with the physics regularized and curriculum regularized models.

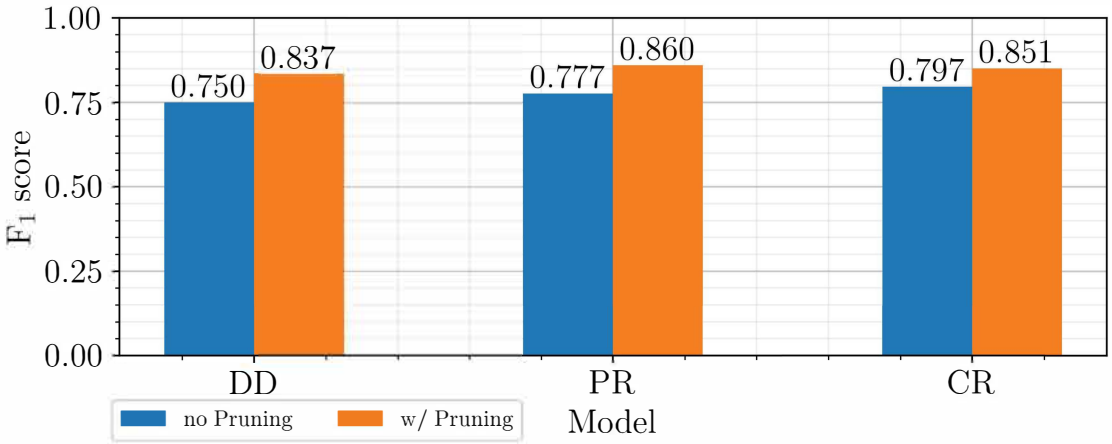

Figure 22: F_1 score obtained for the evaluation set on the ACEC data set using the data-driven model (DD), the physics-regularized model (PR) and curriculum-regularized model (CR) both with and without pruning

These results show that including physics while training the reconstruction model helps the model learn nominal behavior better, thus resulting in a better classification scheme as anomalies induce a high reconstruction error which can efficiently be classified. As well as this, implementing the pruning procedure on unseen data shows that it works competently in generalizing the process of pruning false positives.

8 CONCLUSION AND FUTURE WORK

The results presented above show improvements in the two sub-problems that make up the autonomous threat detection problem.

Regarding the regression sub-problem, it was shown that when using a data-driven machine learning algorithm to reconstruct a signal, including the known physics of the signal into the model's loss function can lower the reconstruction error by restricting the model from becoming too complex. Also, by varying the weight of the physics in the loss function, an optimal balance was established between the data's importance and the physics' importance in the model. Finding an optimal balance showed that a purely physics-based or entirely data-driven model will not perform as well as a model relying on physics and data. As well as this, by using a curriculum regularization training scheme, it was found that slowly increasing the weight of the

regularization term in the loss function can yield a lower reconstruction error.

For the classification sub-problem, the results obtained in the data set with real anomalies for the classification sub-problem showed that the thresholding process is case-based, as some thresholding approaches might work better for different applications. In the case of the two satellites being studied, the $F_{0.5}$ score obtained with the K-means clustering algorithm outperformed previous models relying on dynamic thresholding techniques. The two thresholding methods studied showed that a suitable regression would yield a high recall when classifying anomalies, and the pruning procedure can take advantage of the proper regression results to increase the precision of the model.

The results obtained in the data set with synthetic anomalies showed that including the known physics of the problem while training the reconstruction model can increase both precision and recall in the classification portion. Introducing physics in the training procedure allows the model to learn the nominal behavior better so that anomalies can yield a high reconstruction error, thus being easier to classify as anomalous behavior. As well as this, the division of the testing set in the synthetic anomalies data set for the determination and evaluation of the pruning process designed showed that it performs well on unseen data. The pruning process applied to unseen data increased precision while keeping recall considerably constant.

To finish, the potential for future work needs to be discussed. First, it would be interesting to find a different case where the physics model is known to characterize and validate the regularization approach applied throughout this report. As well as this, it would be interesting to experiment more with the curriculum regularization training scheme. As mentioned before, questions such as finding the model's sensitivity to the rate of increase and the range of the regularization term still need to be answered and can help improve the regression work presented here. For the classification portion, it would be important to test the pruning procedure on unseen real anomalies. For this, more examples of real data with labeled real anomalies would be needed, which can be challenging to find.

References

[1] Atilim G. Baydin, Barak A. Pearlmutter, Alexey Andreyevich Radul, and Jeffrey Mark Siskind. Automatic differentiation in machine learning: A survey. *Journal of Machine Learning Research*, 18(1):5595–5637, 2017.

[2] Sebastian Braun and Ivan Tashev. A consolidated view of loss functions for supervised deep learning-based speech enhancement. In *Proceedings of the 2021 International Conference on Telecommunications and Signal Processing*, Virtual Conference, July 2021.

[3] Jeffrey P. Buzen and Annie W. Shum. Masf - multivariate adaptive statistical filtering. In *Proceedings of the 1995 International Computer Measurement Group Conference*, Nashville, TN, December 1995.

[4] Raghavendra Chalapathy and Sanjay Chawl. Deep learning for anomaly detection: A survey. *CoRR*, abs/1901.03407, 2019.

[5] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[6] David J Griffiths. *Introduction to Electrodynamics*. Pearson, Boston, MA, 2013.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[8] Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Söderström. Detecting spacecraft anomalies using lstms and nonparametric dynamic thresholding. In *Proceedings of the 2018 International Conference on Knowledge Discovery & Data Mining*, London, United Kingdom, July 2018.

[9] John David Jackson. *Classical Electrodynamics*. Wiley, New York, NY, 1999.

[10] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning: with applications in R*. Springer, New York, NY, 2017.

[11] Weihua Jin, Bo Sun, Zhidong Li, Shijie Zhang, and Zhonggui Chen. Detecting anomalies of satellite power subsystem via stage-training denoising autoencoders. *Sensors*, 18:3216, 14.

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proceedings of the 2015 International Conference for Learning Representations,*, San Diego, CA, May 2015.

[13] Ian Knowles and Robert J. Renka. Methods for numerical differentiation of noisy data. In *Proceedings of the 2012 Variational and Topological Methods: Theory, Applications, Numerical Simulations, and Open Problems Conference*, Flagstaff, AZ, June 2012.

[14] Aditi S. Krishnapriyan, Amir Gholami, Shandian Zhe, Robert M. Kirby, and Michael W. Mahoney. Characterizing possible failure modes in physics-informed neural networks. In *Proceedings of the 2021 Advances in Neural Information Processing Systems Conference*, Virtual Conference, December 2021.

[15] Pankaj Malhotra, Anusha Ramakrishnan, Gaurangi Anand, Lovekesh Vig, Puneet Agarwal, and Gautam Shroff. Lstm-based encoder-decoder for multi-sensor anomaly detection. In *Proceedings of the 2016 International Conference in Machine Learning Anomaly Detection Workshop*, New York, NY, June 2016.

[16] M. Raissi, P. Perdikaris, and G. E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2018.

[17] E.C. Stone, A.M. Frandsen, R.A. Mewaldt, E.R. Christian, D. Margolies, J.F. Ormes, and F. Snow. The advanced composition explorer. *Space Science Reviews*, 86(1/4):1–22, 1998.

[18] Chengwei Wang, Krishnamurthy Viswanathan, Lakshminarayan Choudur, Vanish Talwar, Wade Satterfield, and Karsten Schwan. Statistical techniques for online anomaly detection in data centers. In *Proceedings of the 2011 IFIP/IEEE International Symposium on Integrated Network Management and Workshops*, Dublin, Ireland, May 2011.

9 RECONSTRUCTIONS APPENDIX

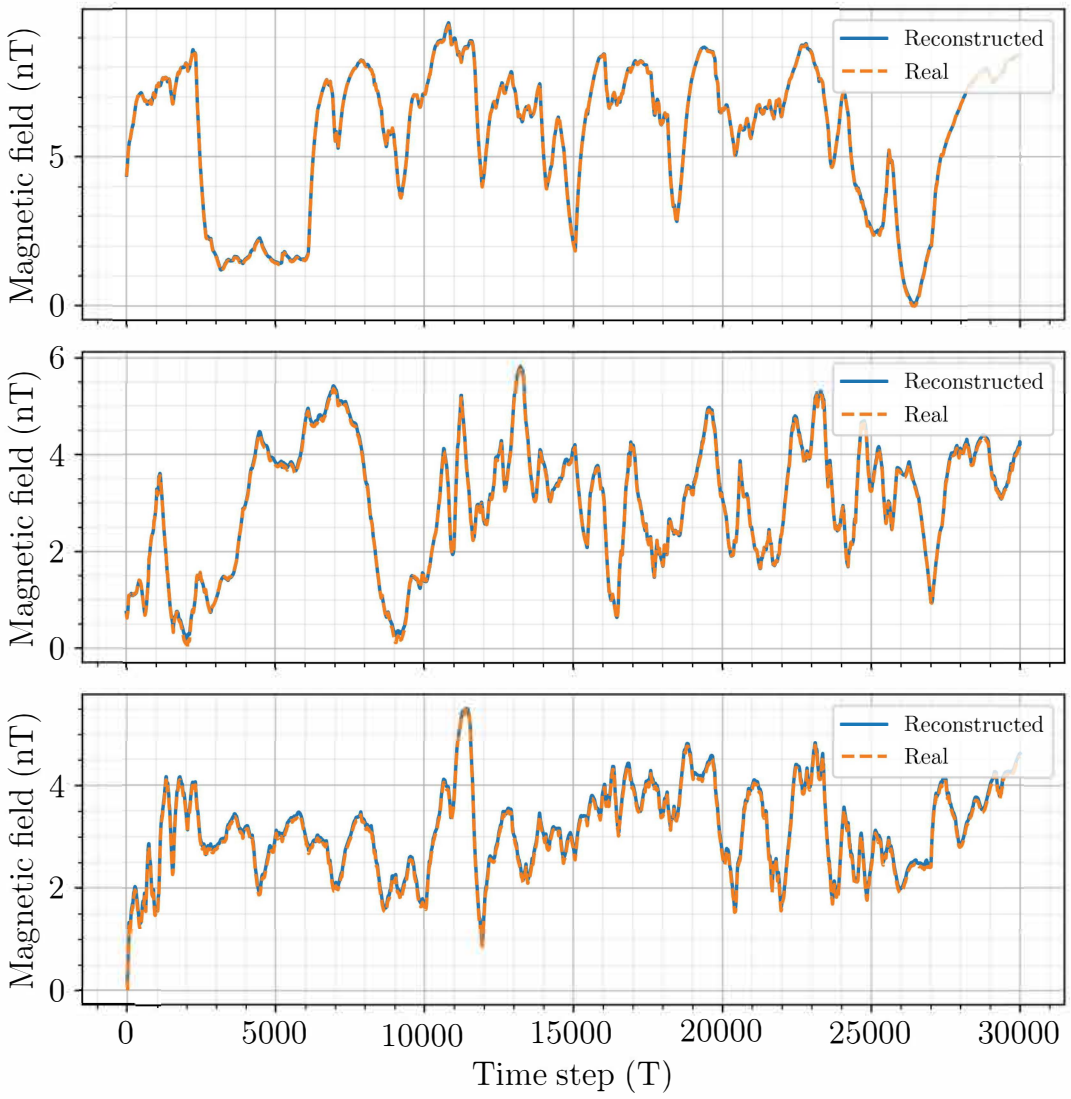

Figure A-1: Real and averaged reconstructed signals for the three components of the mag-netic field in the ACE dataset for the best physics regularization model. The top figure shows B_x, the middle figure shows B_y and the bottom shows B_z

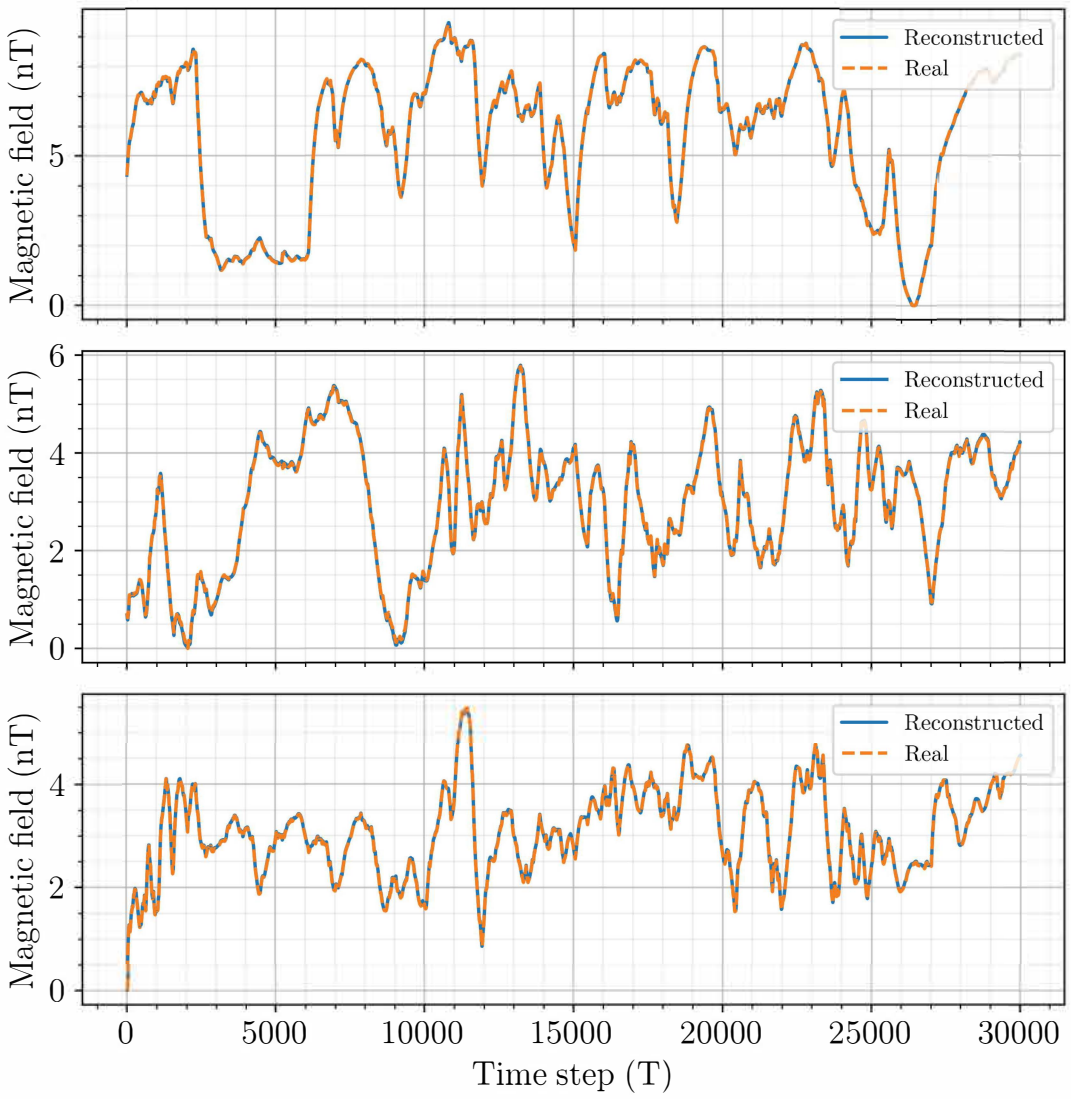

Figure A-2: Real and averaged reconstructed signals for the three components of the mag-netic field in the ACE dataset using curriculum regularization. The top figure shows B_x, the middle figure shows B_y and the bottom shows B_z

DISTRIBUTION LIST

DTIC/OCP
8725 John J. Kingman Rd, Suite 0944 1 cy
Ft Belvoir, VA 22060-6218

AFRL/RVIL
Kirtland AFB, NM 87117-5776 1 cy

Official Record Copy
AFRL/RVS/Michelle Simon 1 cy

(This Page Intentionally Left Blank)