18th International Conference on Space Operations, Montréal, Canada, 26 - 30 May 2025.

SpaceOps-2025, ID # 338

Explaining Satellite Anomalies–Causal Inference for Space Operations

Clemens Schefels[a*], Bilel Ben Salem[b], Andreas Gerhardus[b], Kathrin Helmsauer[a], Baptiste Lambert[b], Julia Niebling[b], Oana-Iuliana Popescu[bc], Martin Rabel[bc], Ferdinand Rewicki[b], Leonard Schlag[a]

[a]*German Aerospace Center (DLR), Space Operations and Astronaut Training, 82234 Weßling, Bavaria, Germany*
[b]*German Aerospace Center (DLR), Institute of Data Science, 07745 Jena, Thuringia, Germany*
[c]*Dresden University of Technology (TUD), Faculty of Computer Science, 01062 Dresden, Saxony, Germany*
[*]Corresponding Author: clemens.schefels@dlr.de

Abstract

The operation of satellites relies heavily on telemetry data, which has become increasingly complex due to the proliferation of parameters. Automatic anomaly detection and explanation are crucial for satellite operators to respond promptly to anomalies and ensure the reliability of their systems. This study aims to bring together classical machine learning methods and deep learning approaches in anomaly detection, with a focus on causal inference. For anomaly detection, we investigate the performance of the deep learning methods Graph-Augmented Normalising Flow (GANF) and Multi-Scale Temporal Variational Autoencoder (MST-VAE), as well as of the classical, density-based estimation Maximally Divergent Intervals (MDI) method. For causal inference, we apply two time series causal discovery algorithms, Peter and Clark Momentary Conditional Independence (PCMCI) and Joint Peter and Clark Momentary Conditional Independence (J-PCMCI), to identify causal relationships in the considered satellite telemetry data. Our methods are designed to provide explainable results and facilitate interpretation of the anomalies by satellite operators. We evaluate our approach using a use case study on satellite telemetry data collected during ground station contacts, incorporating telecommands given. This research contributes to the growing body of work on anomaly detection and causal inference in complex data sets, and advance our understanding of anomaly detection and causal inference.
Keywords: Anomaly detection, causal discovery, machine learning, satellite communications, correlation analysis, signal processing

Nomenclature

$\mathcal{T}$	time series,
p	probability density,
S	subsequence,
$\mathcal{D}$	deviation,
$\mathbf{A}$	adjacency matrix,
$\mathcal{X}$	training set,
$\mathcal{F}$	graph-augmented normalizing flow,
t	time step,
τ	hyper-parameter,
A, B	variables,
$\mathbf{C}$	set of variables

Acronyms/Abbreviations

Automated Telemetry Health Monitoring System (ATHMoS),
Continuous Integration (CI),
Directed Acyclic Graph (DAG),
German Aerospace Center / Deutsches Zentrum für Luft- und Raumfahrt e.V. (DLR),
German Space Operations Center (GSOC),
Graph-Augmented Normalising Flow (GANF),
Kullback–Leibler Divergence (KL-Divergence),
Maximally Divergent Intervals (MDI),
Multi-Scale Temporal Variational AutoEncoder (MST-VAE),
Peter and Clark Momentary Conditional Independence (PCMCI),
Joint Peter and Clark Momentary Conditional Independence (J-PCMCI),
Recurrent Neural Network (RNN)

1. Introduction

In satellite operations, telemetry data plays an important role to track the satellite's system and health status. Therefore, to gain more data and more detailed information, new satellites are equipped with plenty of parameters. For example, both of the earth observation satellites GRACE Follow-On send telemetry data of about 80,000 parameters to earth at each ground station contact. Such an immense number makes a manual inspection of single parameters impossible. For that reason, many space operations centers spend a lot of research in automatically detecting novel or anomalous behaviour in these time series data with the help of machine learning.

The German Space Operations Center (GSOC) at the German Aerospace Center (DLR) is using ATHMoS, the Automated Telemetry Health Monitoring System, to detect novel behaviour and to support its satellite operators. ATHMoS[1] is based on various classical machine learning methods like clustering that return highly explainable results to the operators. However, new anomaly detection methods, especially Deep Learning based approaches, have shown promising results in several research studies. The "black-box" character of these methods and the missing explanations of the results are often the reason why these methods are not yet included into the daily work routine of space operations centers. The operators need to understand the specific cause of the anomalies to take the appropriate actions.

The DLR project CausalAnomalies aims to bring together anomaly detection and causal inference methods to identify and explain anomalies in satellite telemetry data. In our use case study, we apply our methods to the satellite telemetry collected during ground station contacts in relation to the telecommands given. The usage of high dimensional data, telemetry combined with telecommands, differs from the anomaly detection employed by ATHMoS which only takes a single telemetry parameter into account. For anomaly detection, we investigate the performance of the deep learning methods Graph-Augmented Normalising Flow (GANF) and Multi-Scale Temporal Variational Autoencoder (MST-VAE), as well as of the classical, density-based estimation Maximally Divergent Intervals (MDI) method.

Causal inference is a framework that provides concepts and methods for data-driven reasoning about causal relationships. For our use-case, we explore the use of causal inference by applying two time series causal discovery algorithms to identify causal relationships in the considered satellite telemetry data. These algorithms perform a series of statistical tests of (conditional) independence in the data and use the results of these tests to deduce qualitative causal relationships. The visual representation of the learned relationships into a so-called causal graph facilitates the interpretation of the results for the satellite operators. The employed algorithms are implemented in the open-source Python package Tigramite that has been co-developed at the DLR-Institute of Data Science.

The overarching goal of this project is to enable an automated workflow where both steps, anomaly detection and causal inference, are integrated into a continuous integration (CI) pipeline, that offers good predictive performance without sacrificing explainability of the results.

This paper is structured as follows: in Section 2, we report on the utilization of diverse methodologies for detecting anomalies in datasets. Specifically, our approach employs one classical statistical model and two deep learning-based architectures for unsupervised anomaly detection. Furthermore, for causal inference, we leverage two time series causal discovery algorithms to uncover underlying relationships between variables. The next Section 3 is dedicated to our use case, to detect anomalies in ground station contact telemetry and understand their causes using causal inference techniques. Various parameters related to satellite-ground communication are analysed, including electric field strength, relay states, and temperature. Concluding with Section 4, this section provides a final summary of the paper and an overview of planned future activities.

2. Methods

In this section, we delve into the details of each unsupervised anomaly detection method, providing an in-depth examination of their underlying architectures and configurations used in our experiments. This section aims to equip readers with a comprehensive understanding of the strengths and limitations of MDI, GANF, and MST-VAE, enabling them to evaluate the proposed methods in the context of real-world anomaly detection tasks.

In addition to exploring the anomaly detection methods, we also examine how causal inference can be applied to uncover relationships within the ground station-contact dataset. We use the time series causal discovery algorithms PCMCI+ [2] and J-PCMCI+ [3], implemented within the Python package Tigramite.

2.1 Anomaly Detection

In this paper, we investigate one classical and two deep learning-based models for unsupervised anomaly detection, namely Maximally Divergence Intervals (MDI) [4] algorithm, the Graph-Augmented Normalizing Flow (GANF) [5], and the Multi-Scale Temporal Variational Autoencoder (MST-VAE) [6]. In the following section, we briefly introduce these methods and describe the configuration used in the experiments in Section 3.

2.1.1 Maximally Divergence Intervals (MDI)

MDI is a density-estimation based algorithm for unsupervised detection of sequential anomalies in spatio-temporal data. The method identifies spatio-temporal regions as anomalous that differ maximally from the remaining data wrt. their probability density. As we focus purely on temporal data in this work, we omit the spatial aspects and refer to [4] for the original definitions. Given a multivariate time series $\mathcal{T}$, MDI detects anomalous subsequences by comparing the probability density p_S of a subsequence $S_{a,b} \subseteq \mathcal{T}$ to the density p_Ω of the remaining part of the time series $\Omega(S) := \mathcal{T} \setminus S_{a,b}$ for all subsequences. The distributions are modelled using *Kernel Density Estimation (KDE)* or *Multivariate Gaussians*. To measure the degree of deviation $\mathcal{D}(p_S, p_\Omega)$ between p_S and p_Ω, divergence measures for probability densities, such as the *Kullback–Leibler Divergence (KL-Divergence)* are used. [4] also proposed an unbiased version of the KL-Divergence, which mitigates the bias of the KL-Divergence towards low-variance intervals. To mitigate the violation of the i.i.d. assumption as made by the density estimation methods, coming from the auto-correlation of time series data, time delay embedding [7] is used to incorporate context from previous timestamps into each observation. The most anomalous subsequence $\tilde{S}$ is found by solving the underlying optimization problem:

$$\tilde{S} := \arg\max_{S \subseteq \mathcal{T}} \mathcal{D}(p_S, p_{\Omega(S)})$$

MDI locates this most anomalous subsequence $\tilde{S}$ by scanning all subsequences $S \subseteq \mathcal{T}$ with a length in a predefined interval $[L_{min}, L_{max}]$ and estimates the divergence $\mathcal{D}(p_S, p_{\Omega(S)})$, which is then used as the anomaly score. The parameters L_{min} and L_{max} need to be defined in advance. MDI employs various techniques to estimate the probability distributions, including KDE and multivariate Gaussian models. For the Gaussian model, the exact KL-Divergence, as well as the unbiased version, has a closed-form solution, allowing for efficient computation. To detect multiple anomalies, MDI uses a non-maximum suppression method to select the k non-overlapping intervals with the highest divergence. This approach allows MDI to identify coherent anomalous regions rather than isolated anomalous points, making it particularly suitable for detecting anomalies driven by complex natural processes [4]. To accommodate the application to large-scale data, an interval proposal technique based on Hotelling's T^2 method [8] is employed, which selects interesting subsequences based on point-wise anomaly scores instead of performing full scans over the entire time series.

We use MDI with a multivariate Gaussian model for density estimation and the unbiased KL-Divergence for comparing probability densities and automatic time delay embedding. As we aiming at identifying anomalous ground station contacts, we analyse intervals between $L_{min} = 1324$ and $L_{max} = 2281$ data points, which refers to the length of the shortest and the second-longest contact. We did not consider the length of the longest contact (32223 data points), as it is clearly outlying. To improve computational performance, we use Hotelling's T^2 interval proposals. As MDI identifies intervals within a time series, we concatenate the multivariate time series of all analysed ground station contacts as a pre-processing step.

2.1.2 Graph-Augmented Normalizing Flow (GANF)

GANF is another density-based method for unsupervised time series anomaly detection method that leverages normalizing flows for density estimation. Normalizing flows are generative models $f : \mathbb{R}^d \to \mathbb{R}^d$ that applies a sequence of invertible and differentiable transformations to map complex data distributions onto a simple "base" distribution, such as an isotropic Gaussian, where the density evaluation is typically straightforward [5]. Beyond density estimation, GANF integrates a Bayesian Network to model causal relationships among multivariate time series $\mathcal{T} = (T_1, ...T_d)$.

Given a training set $\mathcal{X}\{\mathcal{T}i\}_{i=1}^{|\mathcal{D}|}$ of multiple time series, GANF aims to learn the adjacency matrix $\mathbf{A}$ of the Bayesian Network and, simultaneously, the graph-augmented normalizing flow $\mathcal{F} : (\mathcal{T}, \mathbf{A}) \to \mathcal{Z}$, where $\mathcal{Z}$ is a random variable with a "simple" (base) distribution [5]. Once $\mathcal{F}$ is learned, the estimated density $p(\mathcal{T})$ can be evaluated to identify anomalies in low-density regions of the base distribution. GANF's dependency encoder consists of a recurrent neural network that summarizes the time series up to a given time step t and a graph convolution layer that captures dependency representations. These are then used to condition the normalizing flow f. Since anomalies are rare events by definition, they are assumed to have low densities. Thus, the estimated density serves as an anomaly score [5].

2.1.3 Multi-Scale Temporal Variational Autoencoder (MST-VAE)

Variational Autoencoders (VAEs) are a class of generative models that learn a structured latent representation of input data. They consist of an encoder network that maps the input data to a lower-dimensional latent space and a decoder network that reconstructs the input from this latent representation. In the context of anomaly detection, VAEs are trained on normal data, and instances exhibiting high reconstruction error are identified as anomalies.

In this work, we explore a convolution-based VAE architecture inspired by the Multi-Scale Temporal Variational Autoencoder (MST-VAE) [6]. The encoder consists of two parallel blocks made of convolutional layers, each designed to capture different temporal patterns: one block employs a smaller kernel size in the first Conv1D layer to focus on short-term dependencies, while the other utilizes a larger kernel size to capture long-term dependencies. Subsequent layers in both blocks perform dimensionality reduction, all utilizing a kernel size of 2. The outputs from these parallel branches are concatenated and passed through a final convolutional layer to match the dimensionality of the latent space. The decoder mirrors the encoder's architecture in a symmetric fashion.

The key advantage of the MST-inspired architecture is its ability to effectively capture both intra-series temporal dependencies and inter-series correlations while maintaining a lower computational footprint compared to Recurrent Neural Network (RNN)-based VAEs, which offer similar capabilities [6].

2.2 Causal Inference

In addition to anomaly detection, we also apply techniques from causal inference to the analysis of the ground station-contact dataset. Causal inference, see for example [9–12], is a research field at the intersection of statistics, computer science, and machine learning that develops theory and methods for data-driven reasoning about cause-and-effect relationships. Such methods are of particular interest in applications where targeted experimentation is not possible or undesirable.

More specifically, we here apply methods for what is called *causal discovery*, see for example [10, 12]. Causal discovery refers to the task of learning qualitative causal relationships between a collection of variables from data. This qualitative information is typically represented by a graph that, given its causal semantics, is often referred to as a *causal graph*. Depending on the assumptions that one is willing to make about the data-generating process, the causal graph can belong to different types of graphical models. In this paper, to simplify the explorative analysis presented here, we impose the following two assumptions: First, the assumption of no cyclic causal relationships, which says that if variable A causally influences variable B, then B does not causally influence A. Second, the assumption of no unobserved confounders—also known as assumption of no hidden common causes or as assumption of *causal sufficiency* [10]—, which says that if variables A and B are part of the dataset and there is third variable C that causally influences both A and B, then C is part of the dataset too. Given these two assumptions, the causal graph is a directed acyclic graph (DAG) whose vertices—also known as nodes—correspond to variables and whose directed edges signify direct causal influences. * For example, if the causal graph is $A \to B \to C$, then A, then A has a direct causal influence on B and B has a direct causal influence on C (and A has an indirect causal influence on C via B).

Both the assumption of no cyclic causal relationships and of causal sufficiency can be relaxed in causal discovery, see for example [13] and [10, 14] respectively. Due to the distinction between causal relationships on the one hand (what one wants to learn) and statistical relationships on the other hand (what one can "see", that is, test for in the data), at least some assumptions are fundamentally necessary to enable causal discovery from observational data. The precise form and extend of such enabling assumptions depend on the specific approach and method to causal discovery that one chooses to apply. Since a discussion of the different approaches and increasingly large number of methods is out of scope here, we instead, for example, refer the reader to the review papers [15, 16].

For the work presented here, we use the time series causal discovery algorithms PCMCI$^+$ [2] and J-PCMCI$^+$ [3] These algorithms are implemented within the Python package *Tigramite*[†] and fall into the *constraint-based* approach to causal discovery, see for example [10, 15]. The basic idea of this approach is to infer the causal graph, or, more precisely, a set of possible causal graphs, from marginal and conditional independencies[‡] in the data. To this end, constraint-based causal discovery algorithms employ the so-called *causal Markov* and *causal faithfulness* assumptions [10], or variations thereof. These assumptions establish a one-to-one correspondence between independencies in the data on the one hand and the graphical notion of d-separation [17–19] applied to the causal graph on the other hand. Thus, for now assuming to have access to perfect knowledge about independencies in the data, the algorithms can use this knowledge to put constraints on the causal graph. For example, these constraints can be that certain variables A and B are not connected by an edge whereas two other certain variables C and D are connected by the edge $C \to D$. The constraints do not, however, always determine a unique causal graph but rather a set of possible causal graphs. Specifically, some edges may remain "unoriented" in the sense that, while the algorithm is able to conclude that certain variables A and B are connected by an edge, the algorithm cannot decide whether this edge is $A \leftarrow B$ or $A \to B$. The constraint-based approach to causal discovery has originally been developed for data without temporal information, an

*To simplify the discussion, we often do not distinguish between the vertices and the variables they represent.

[†]`https://github.com/jakobrunge/tigramite/tree/master/tigramite` (accessed: March, 28th, 2025 at 12:18pm UTC)

[‡]To simplify the discussion, we from here drop the "marginal and conditional" and instead simply write "independencies".

important representative being the PC algorithm [20].

The PCMCI algorithm [21] adapts the PC algorithm to (discrete time) time series data in a non-trivial way with the aim to keep a high detection power in cases where relatively few data are available [21]. It assumes the time series data to be stationary and targets to also learn the *time lags* of causal influences. That is, the goal is not only to learn that, say, variable A causally influences variable B, but also that this influence takes τ time steps to manifest itself. The latter is graphically represented by an edge $A_{t-\tau} \to B_t$, where the subscripts are time indices and, due to stationary, t is an arbitrary reference time point. Thus, giving rise to what is called *full time graph* in [12], the causal graph is "resolved in time" in the sense that for each variable, for example, A, it contains multiple vertices, namely $\ldots, A_{t-1}, \ldots, A_t, A_{t+1}, \ldots$. Supposing that all direct causal influences have a finite time lag, and again resorting to stationarity, it is, however, sufficient to look at the finite part of the full-time graph within the time window $[t-p, t]$, where p is the maximum time lag in the full-time graph. Correspondingly, see [21], the PCMCI algorithm has a hyper-parameter $\tau_{\max}$ that needs to be chosen such that $\tau_{\max} \geq p$. Then, the algorithm attempts to learn a causal graph in which for each variable, for example, A, there are $\tau_{\max} + 1$ many vertices, namely $A_{t-\tau_{\max}}, \ldots, A_t$. In addition to stationarity, the algorithm assumes the causal Markov and causal faithfulness and causal sufficiency assumptions (see above) as well as the absence of causal influences at time scales smaller than the time resolution, that is, the absence of edges of the type $A_{t-\tau} \to B_t$ with $\tau = 0$. Noting that causal influences cannot go back in time, so $A_{t-\tau} \to B_t$ implies $\tau \geq 0$, the latter guarantees the absence of cyclic causal relationships. However, the assumptions allow that variables mutually influence each other. For example, the full-time graph can contain both $A_{t-\tau_{AB}} \to B_t$ and $B_{t-\tau_{BA}} \to A_t$. For more details on the PCMCI algorithm see the original paper [21] or, for example, the review paper [16].

The PCMCI$^+$ algorithm [2] generalizes PCMCI by allowing for causal influences at time scales smaller than the time resolution, which correspond to edges with lag $\tau = 0$. However, this type of causal influences remains restricted by the assumption of no cyclic causal relationships. For example, if there is the edge $A_t \to B_t$, then the assumption excludes that there is also $B_t \to A_t$. See [22] for an example case study that applies PCMCI$^+$.

Both PCMCI and PCMCI$^+$ take as input a single multivariate time series or, as applied in [23], multiple multivariate time series that are assumed to be realizations of the same process. Using so-called *context variables*, the J-PCMCI$^+$ algorithm generalizes the latter setting by allowing for a certain type of variability between the multiple multivariate time series [3]. For a precise statement of which form of variability is allowed, see [3].

The above explanation of constraint-based causal discovery assumed that the algorithms have access to perfect knowledge about independencies in the data. In practice, however, this is not the case. Rather, independencies need to be tested for by statistical means and the causal discovery algorithms need to operate with the results of these tests. In the work presented here, we utilize two types of independence tests that are implemented in Tigramite and can be combined with PCMCI$^+$ and J-PCMCI$^+$. First, the ParCorr test[§] considers testing the independence $A \perp\!\!\!\perp B \,|\, \mathbf{C}$, where $\mathbf{C}$ is a set of variables (if $\mathbf{C}$ is empty, then the independence is marginal, else it is conditional). If $\mathbf{C}$ is empty, the test works by testing for non-zero correlation between A and B. Else, that is if $\mathbf{C}$ is non-empty, the test works by first regressing A and B on $\mathbf{C}$ (using ordinary least squares) and then testing for non-zero correlation of the residual. Thus, ParCorr is appropriate for data that follows a multivariate normal distribution. As such, the test is (at least in theory) not appropriate for discrete variables. In order to also handle multivariate variables A and/or B, we also employ the *ParCorrMult* test.[¶] This test works by regressing all components of A and B on $\mathbf{C}$ (using ordinary least squares), then calculating the matrix of correlations between the residuals and finally testing whether the maximum of the correlations is non-zero.

Within one application, PCMCI and its variants typically perform multiple independence tests [2, 3, 21]. Each such test internally computes two numbers: The value of the test statistic and the corresponding p-value. PCMCI and its variants then compare the p-value to a significance value α, which is a hyperparameter, in order to judge whether or not the independence is supposed to be true. When the algorithms test independencies $A \perp\!\!\!\perp B \,|\, \mathbf{C}$ for the same pair (A, B) but different sets $\mathbf{C}$, they then keep track of the maximal p-value across all such tests. In addition, they also keep track of the value of the test statistic for the test with maximal p-value, which below we refer to as the *val-value of the pair of variables*. These numbers, that is the maximal p-value and corresponding value of the test statistic per pair (A, B) of variables, are available as an output of the algorithms.

[§] `https://github.com/jakobrunge/tigramite/blob/master/tigramite/independence_tests/parcorr.py` (accessed: April, 4th, 2025 at 16:09pm UTC).

[¶] `https://github.com/jakobrunge/tigramite/blob/master/tigramite/independence_tests/parcorr_mult.py` (accessed: April, 4th, 2025 at 16:15pm UTC).

3. Use Case

The reliable operation of satellite communication systems depends on a thorough understanding of telemetry data during ground station contacts. Effective monitoring and analysis of this data are essential for ensuring communication integrity and diagnosing potential issues. In this project, we analyse satellite telemetry with a particular focus on ground station contact events, aiming to identify anomalous interactions and explain their causes using causal inference techniques.

Satellite telemetry provides a wealth of information about the performance and health of communication systems during interactions with ground stations. By examining various parameters, we can gain insight into signal quality, equipment status, and potential operational anomalies. Our analysis encompasses several key telemetry parameters related to satellite-ground communication. These include the electric field strength for both low-rate and high-rate receivers, as well as the status parameter that indicates whether a downlink has been enabled. Understanding these parameters is crucial, as they provide fundamental insights into the satellite's ability to establish and maintain effective communication links.

Additionally, we examine the relay states for high-rate and low-rate transmitters, which determine whether the appropriate transmission pathways are active during ground station contacts. The electricity supply for the satellite BUS is also a critical factor, as it ensures the proper functioning of all subsystems involved in the communication process. Another important aspect of our analysis is the sub-carrier locks, which indicate whether the satellite's receiver has successfully synchronized with the ground station signal. This synchronization is essential for stable data transmission and overall communication reliability.

Beyond signal acquisition and transmission pathways, we also analyse the receiver carrier loop stress, which corresponds to the Doppler offset frequency, and the transmitter carrier loop stress, which reflects frequency deviations during transmission. These parameters are crucial for understanding the dynamic effects of satellite movement on signal integrity. Other critical telemetry parameters include the electricity consumption of the transmitters, which provides insights into power efficiency and potential anomalies in energy usage. The temperature of the transmitter during both low- and high-rate contacts is another factor of interest, as overheating or unusual temperature fluctuations can indicate potential hardware issues.

To differentiate between low-rate and high-rate contacts, we consider the downrate parameter, which serves as a distinguishing factor between different modes of communication. Furthermore, the ratio of BUS electricity consumption versus payload electricity consumption is analysed to assess how power resources are distributed between essential satellite functions and its operational payload. Understanding these aspects helps us evaluate the overall health of the satellite and optimize communication strategies.

The primary objective of this project is to detect anomalies in ground station contact telemetry and interpret their underlying causes. By leveraging anomaly detection techniques, we aim to identify deviations from expected communication behaviour that may indicate potential system failures or performance degradation. Additionally, causal inference methods are employed to establish relationships between telemetry parameters and detected anomalies, facilitating a deeper understanding of potential failure modes and operational inefficiencies. By identifying and understanding anomalies in ground station contacts, we can develop proactive measures to mitigate potential issues before they impact communication capabilities. Additionally, our findings can inform future satellite design and operational strategies, leading to improved robustness and resilience in satellite-ground communication systems. This research ultimately contributes to the advancement of space communication technology, ensuring reliable and efficient data transmission in various mission scenarios.

3.1 Data Processing

To facilitate accurate anomaly detection and causal analysis, we perform a series of pre-processing steps on the raw telemetry data. First, we load the complete set of telemetry records and use the relay state parameters to define the start and end times of both high-rate and low-rate ground station contacts. These relay states act as reliable indicators for when communication links are active, providing natural boundaries for segmenting relevant data. Around each identified contact event, we extract the corresponding telemetry parameters with an additional buffer period included both before the start and after the end of the contact. This buffer ensures that we capture transitional behaviours and potential precursors to anomalies.

The segmented telemetry data is then organized into two distinct Python dictionaries, one for high-rate and one for low-rate contacts, each containing only the telemetry records associated with these contact windows. To standardize the data for further analysis, we interpolate the telemetry parameters onto a uniform 0.5-second time grid. Numerical parameters are interpolated using standard numerical interpolation techniques to maintain continuity and resolution,

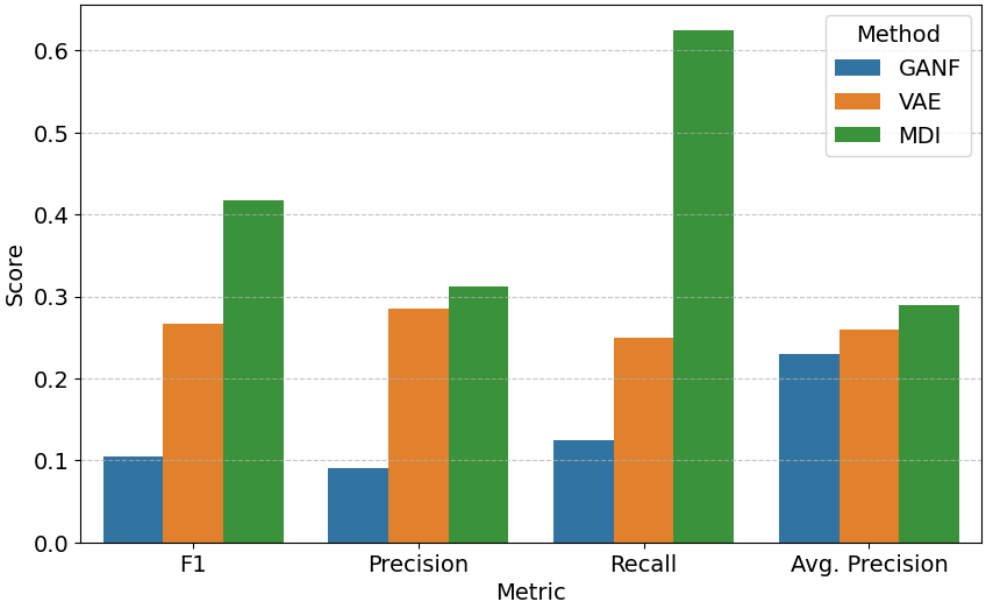

Fig. 1. F1 score, precision, recall and average precision of the three anomaly detection models.

while symbolic or categorical parameters (e.g., status flags) are interpolated using a nearest-value approximation to preserve semantic meaning. The result is two structured time series datasets that represent telemetry during contact events, forming a robust foundation for our subsequent work.

3.2 Anomaly Detecion

To detect anomalous ground station contacts, we apply MDI, GANF and MST-VAE to the pre-processed data as described in Section 3.1. Since MDI takes a single multivariate time series as input, we concatenate the individual ground station contacts for MDI. For GANF and MST-VAE however, we derive features for the individual ground station contacts using the Catch22 feature extraction technique, as described in Lubba et al. [24]. Catch22 reduces each transmission to a set of 22 summary features, which encapsulate key characteristics of the time series regardless of its length. This transformation substantially reduces the data's dimensionality, thereby decreasing both model size and complexity, making learning easier on limited data. Additionally, GANF and MST-VAE require training, which was done on the 2016 data (1055 contacts), ensuring a fair comparison of model predictions across the complete 2018 dataset (1447 contacts).

Since we want to identify anomalous ground station contacts, we flag a contact as anomalous if any subsequence of it has been detected by at least one of the three anomaly detection methods. If a sequence was flagged as anomalous, that spans multiple neighbouring contacts, we flag all of them as anomalous. For MDI, we focus on the top-10 detected sequences, resulting in 16 ground station contacts flagged as anomalous. GANF found 11 ground station contacts to be anomalous while MST-VAE flagged 7. Two ground station contacts have been detected by two of the three anomaly detection methods jointly. Hence, we got 32 candidates for anomalous ground station contacts out of the 1447 ground station contacts in the 2018 dataset. These 32 candidates have been investigated by a GSOC system engineer. This yielded a true positive rate of 25% across the 32 anomaly candidates.

While predictions were made on the entire dataset, the F1 score, precision, recall, and average precision, shown in Figure 1, are calculated using only the 32 inspected anomaly candidates to avoid making assumptions about the unlabelled data. Similarly, to previous studies [25, 26], the classical anomaly detection model, MDI, outperforms deep learning models. All three methods successfully identified at least one true anomaly, while no true anomaly was correctly predicted by more than one model, suggesting an ensemble approach. However, this carries the risk of increasing number of false positive results.

Among the 32 candidates visually inspected, one candidate labelled as normal by the GSOC expert was the only one associated with a known error report written by engineers in 2018. This anomaly was detected by both MST-VAE and GANF. More broadly, this highlights the challenge of accurately and consistently labelling anomalies in satellite data, emphasizing the value of explainability from a causal perspective as a potential tool to assist engineers in their

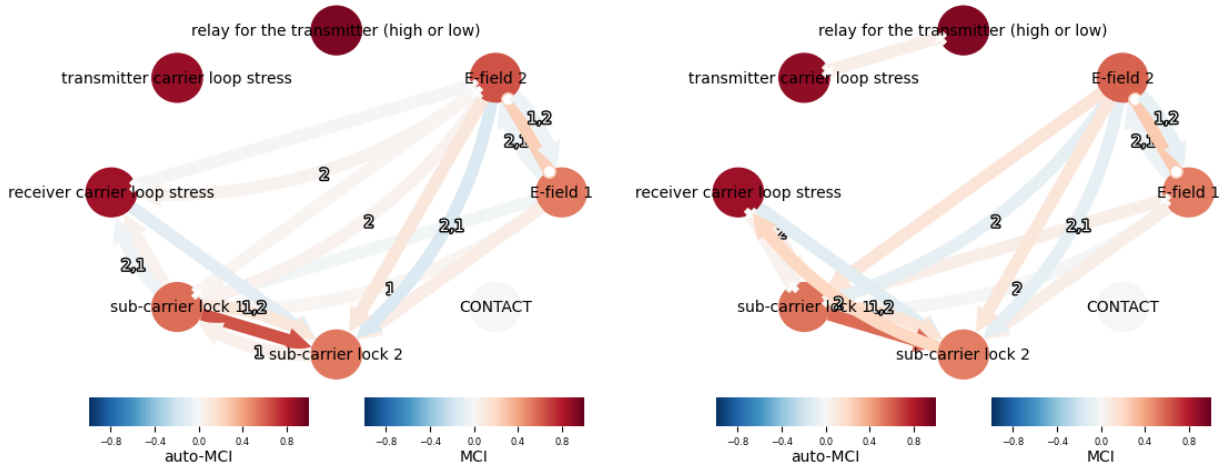

Fig. 2. **Left**: Causal graph learned by J-PCMCI$^+$ on 20 randomly selected normal contacts, using ParCorrMult for independence testing and with hyperparameters $\tau_{\max} = 2$ and $\alpha = 0.01$. **Right**: Result for 20 others randomly selected normal contacts with otherwise unchanged settings. Plots created with Tigramite.

diagnostic process.

3.3 Causal Inference

In this section, we present our causal discovery analysis. To this end, we first introduce the contact data used in the analysis and explain our selection process. In particular, we consider two types of cases: (i) normal contacts, where a successful connection between the satellite and the ground station has been established; as well as (ii) a specific non-normal contact, where the connection has been successfully established too but an anomaly has been reported and confirmed by experts. Second, we present and discuss the results of applying the J-PCMCI+ causal discovery algorithm to a number of normal contacts. Third and finally, we explore whether we can distinguish the non-normal contact from normal contacts by means of causal discovery with the PCMCI$^+$ algorithm. For our analysis, we employ version 5.2.6.7 of Tigramite.

Data selection: In our analysis, a contact is defined as the time interval during which data is downloaded from the satellite to the ground station. According to expert knowledge, a successful contact is characterized by the activation of at least one of the two sub-carriers. Based on this criterion, 907 out of the 1896 recorded contacts are successful, and for the following causal discovery analysis we restrict to these 907 successful contacts. Among the successful contacts, one particular contact has been identified as anomalous by engineers.

Causal discovery with J-PCMCI$^+$: As explained in Section 2.2, the J-PCMCI$^+$ causal discovery algorithm [3] can operate on *multiple* multivariate time series. We use this feature to combine several contacts, each constituting a multivariate time series, to a joint dataset. In this way, we increase the overall size of the dataset and, thus, expect more reliable results as compared to running causal discovery on single contacts. At the same time, to account for a certain variability across the individual contacts, we add an artificial multivariate variable *CONTACT* to the data. This variable serves as a one-hot-identifier (hence it is multivariate) for the individual contacts and is what [3] refer to as a *space-dummy variable*. In order to handle this multivariate variable in independence testing, we here employ the ParCorrMult conditional independence test. Figure 2 shows causal graphs obtained in this way, with further details given in the figure caption.

In these graphs, the coloured vertices and edges respectively represent the variables and causal influences between these. More specifically, uncurved edges represent causal influences at time scales smaller than the time resolution ($\tau = 0$). If such an edge is of the form $\circ\!\!-\!\!\circ$, as for example between *E-Field 1* and *E-Field 2* in the upper graph, then the algorithm was unable to decide between the two possible directions ($\rightarrow$ vs $\leftarrow$) of the edge. Moreover, if such an edge is of the form $\times\!\!-\!\!\times$, as for example between *receiver carrier loop stress* and *sub-carrier lock 1* in the lower graph, then the algorithm obtained conflicting information on whether the edge should be $\rightarrow$ or $\leftarrow$. The curved edges jointly represent

all lagged causal influences ($\tau > 0$), with the numbers indicating the lags. For example, in the upper graph *E-field 2* is found to have a causal influence on *receiver carrier loop stress* with the single lag $\tau = 2$ and a causal influence on *sub-carrier lock 2* at both lags $\tau = 1$ and $\tau = 2$ $\tau = 1$ and $\tau = 2$. The edges are coloured according to the val-value of the corresponding pair of variables, c.f. the last paragraph of Section 2.2 (for curved edges that represent more than one lag, the maximum across the val-values is taken). Since the ParCorrMult independence test was used, the val-values fall into the range $[-1, 1]$. The corresponding colour scale is given in the lower-right of the figure.

By visually examining both causal graphs in Figure 2, we observe that the edge colours are generally not very intense, thus indicating relatively weak causal relationships among the variables. The space-dummy variable *CONTACT* is isolated in both graphs, which indicates that the causal relationships remain constant across the individual contacts. Moreover, the variables *transmitter carrier loop stress* and *relay for the transmitter (high or low)* are isolated from the other five variables in both graphs, with a relatively faint edge of the type $\times\!\!-\!\!\times$ between them in the lower graph only. The other five variables are connected by various edges. Among these, many edges are consistent between the two graphs. Examples of this type are the edges between *E-field 1* and *E-field 2*, the edges from *E-field 2* to *sub-carrier lock 2*, as well as the directed edges from *receiver carrier loop stress* and *sub-carrier lock 1* to *sub-carrier lock 2*. However, there are also some inconsistencies between the two graphs. For example, in the upper graph only there is a lagged ($\tau = 2$) edge from *E-field 2* to *receiver carrier loop stress* and in the lower graph only there is a lagged ($\tau = 2$) edge from *sub-carrier lock 2* to *receiver carrier loop stress*. In addition, in both graphs there are edges of the type $\times\!\!\rightarrow\!\!\times$, which might indicate violations of the underlying assumptions (for example, of the assumption of no cyclic relationships) or errors in the results of the independence tests.

In summary, based on our visual judgement the two causal graphs in Figure 2 appear to be relatively similar to each other, thus building some trust in the results. Nevertheless, further analyses are necessary before drawing confident conclusions. In particular, one should vary the various hyperparameters of the analysis—such as $\tau_{\max}$ and α as well as the number of contacts that are combined to a joint dataset—and evaluate the stability of the results. Further, we note that since the variables *sub-carrier lock 1* and *sub-carrier lock 2* take binary values, it would be more suitable to use a conditional independence test for mixed-type data (with both continuous and discrete variables) rather than ParCorrMult. While Tigramite does implement such tests, for example RegressionCI[‖], we were not able to successfully combine this test with other functionalities of Tigramite to run a J-PCMCI$^+$ analysis with the multivariate space-dummy variable. In future work, one should resolve this issue and also obtain results with conditional independence test for mixed-type data.

Causal discovery with PCMCI$^+$ on the anomalous vs normal contacts: As the last part of our analysis, we briefly explore whether we are able to distinguish the anomalous contact from the normal contacts by means of causal discovery. To this end, we apply the PCMCI$^+$ algorithm [2] on the single anomalous contact with the same hyperparamters as above ()$\tau_{\max} = 2$ and $\alpha = 0.01$ and using the ParCorr conditional independence test[**] The upper left part of Figure 3 shows the resulting causal graph, and indeed this graph is quite different from the ones in Figure 2. However, in the upper right and lower left and right parts of Figure 3 we show the causal graphs obtained by applying PCMCI$^+$ with the same settings on three different randomly selected single normal contacts. Also, these three causal graphs are quite different from the graphs in Figure 2 and, in addition, from each other. These observations indicate that the results of causal discovery with PCMCI$^+$ on single contacts are not stable, and hence we cannot use these results to distinguish the anomalous from the normal contacts. However, further analyses could be beneficial.

4. Summary and Future Work

This paper investigates anomaly detection and causal inference methods for satellite telemetry data. We aim to develop explainable results that facilitate interpretation by satellite operators. For anomaly detection, we propose a combination of unsupervised methods, including Maximally Divergent Intervals (MDI), Graph-Augmented Normalizing Flow (GANF), and Multi-Scale Temporal Variational Autoencoder (MST-VAE). The performance of these methods is evaluated on a use case study involving satellite telemetry data collected during ground station contacts in relation to the telecommands given.

To make the results explainable to the satellite operators, we also apply causal inference techniques, specifically time series causal discovery algorithms, to identify causal relationships in the satellite telemetry data. We use the time series causal discovery algorithms PCMCI$^+$ and J-PCMCI$^+$, implemented within the Python package Tigramite. These

[‖] `https://github.com/jakobrunge/tigramite/blob/master/tigramite/independence_tests/regressionCI.py` (accessed: March, 28th, 2025 at 12:18pm UTC).

[**] Since in this application the dataset consists of a single contact, we employ PCMCI$^+$ rather than J-PCMCI$^+$. Moreover, since here there is no multivariate variable *CONTACT*, we employ ParCorr rather than ParCorrMult.

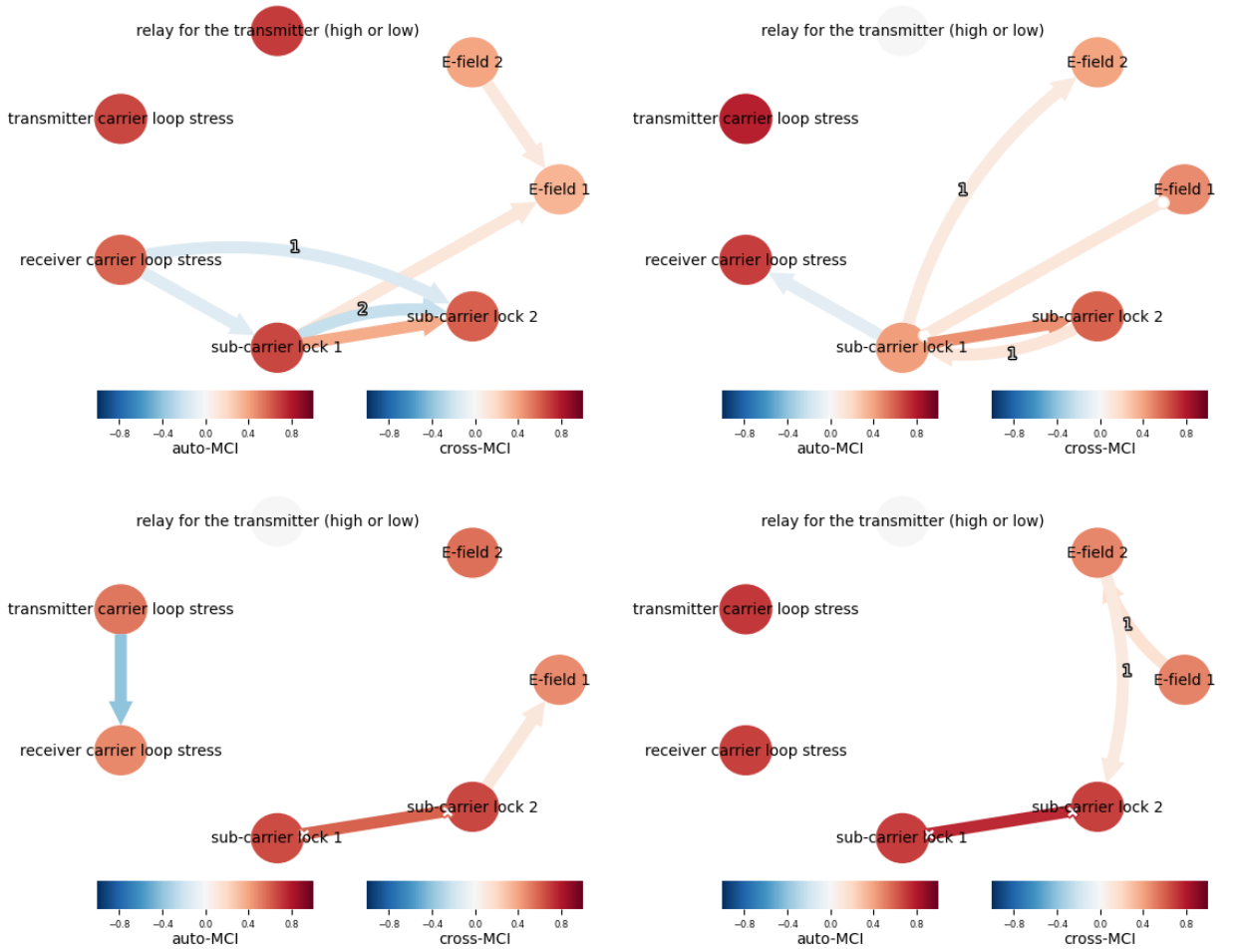

Fig. 3. **(Upper left)**: Causal graph obtained by applying PCMCI$^+$, with the settings as explained in the text to the anomalous contact. **(Other parts)**: Results for three different randomly selected single normal contacts with otherwise unchanged settings. Plots created with Tigramite.

algorithms are used to uncover relationships within the dataset and provide explanations for anomalies detected by the anomaly detection methods.

Overall, the paper contributes to the growing body of work on anomaly detection and causal inference in complex data sets, with a focus on developing explainable results that facilitate interpretation by satellite operators.

In a follow-on project, we plan to integrate the results of the CausalAnomalies project into our novelty detection pipeline. Specifically, we aim to enhance ATHMoS to accommodate multi-parameter capabilities, allowing for more nuanced and robust anomaly detection. Additionally, we intend to align the used causal inference framework with ATHMoS, enabling the incorporation of temporal and spatial relationships between variables in the analysis of anomalies. By integrating these components, we expect to provide more explainable results and facilitate interpretation of the anomalies to the satellite operators.

Acknowledgements

The authors are thankful to Steffen Zimmermann, Andreas Spörl, Arvind Kumar Balan, and all members of the DLR MBT-Team for their valuable support and fruitful discussions.

References

[1] O'Meara, C., Schlag, L., Faltenbacher, L., and Wickler, M., "ATHMoS: Automated Telemetry Health Monitoring System at GSOC using Outlier Detection and Supervised Machine Learning," *Proceedings of the 14th International Conference on Space*

Operations (SpaceOps 2016), 2016.

[2] Runge, J., "Discovering contemporaneous and lagged causal relations in autocorrelated nonlinear time series datasets," *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, Proceedings of Machine Learning Research, Vol. 124, edited by J. Peters and D. Sontag, PMLR, 2020, pp. 1388–1397.

[3] Günther, W., Ninad, U., and Runge, J., "Causal Discovery for time series from multiple datasets with latent contexts," *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, Proceedings of Machine Learning Research, Vol. 216, edited by R. J. Evans and I. Shpitser, PMLR, 2023, pp. 766–776. URL https://proceedings.mlr.press/v216/gunther23a.html.

[4] Barz, B., Rodner, E., Garcia, Y. G., and Denzler, J., "Detecting Regions of Maximal Divergence for Spatio-Temporal Anomaly Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 41, No. 5, 2019, pp. 1088–1101. doi:10.1109/TPAMI.2018.2823766.

[5] Enyan, D., and Jie, C., "Graph-Augmented Normalizing Flows for Anomaly Detection of Multiple Time Series," *arXiv*, 2022. URL https://arxiv.org/abs/2202.07857.

[6] Pham, T.-A., Lee, J.-H., and Park, C.-S., "MST-VAE: Multi-Scale Temporal Variational Autoencoder for Anomaly Detection in Multivariate Time Series," *Applied Sciences*, Vol. 12, No. 19, 2022, pp. 1–14. doi:10.3390/app121910078.

[7] Packard, N. H., Crutchfield, J. P., Farmer, J. D., and Shaw, R. S., "Geometry from a Time Series," *Phys. Rev. Lett.*, Vol. 45, 1980, pp. 712–716. doi:10.1103/PhysRevLett.45.712, URL https://link.aps.org/doi/10.1103/PhysRevLett.45.712.

[8] MacGregor, J., "Statistical Process Control of Multivariate Processes," *IFAC Proceedings Volumes*, Vol. 27, No. 2, 1994, pp. 427–437. URL https://www.sciencedirect.com/science/article/pii/S1474667017481882, iFAC Symposium on Advanced Control of Chemical Processes, Kyoto, Japan, 25-27 May 1994.

[9] Pearl, J., *Causality: Models, Reasoning, and Inference*, Cambridge University Press, New York, NY, USA, 2000.

[10] Spirtes, P., Glymour, C., and Scheines, R., *Causation, Prediction, and Search*, 2nd ed., MIT Press, Cambridge, MA, USA, 2000. URL http://books.google.com/books?hl=en{&}lr={&}id=vV-U09kCdRwC{&}oi=fnd{&}pg=PR11{&}dq=Causation,+Prediction,+and+Search{&}ots=DVVUvqBKic{&}sig=-JCft4QPEjAQgOQ2es{_}L9{_}AFI0Uhttp://books.google.com/books?hl=en{&}lr={&}id=vV-U09kCdRwC{&}oi=f.

[11] Imbens, G. W., and Rubin, D. B., *Causal Inference in Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press, Cambridge, United Kingdom, 2015.

[12] Peters, J., Janzing, D., and Schölkopf, B., *Elements of Causal Inference: Foundations and Learning Algorithms*, MIT Press, Cambridge, MA, USA, 2017.

[13] M. Mooij, J., and Claassen, T., "Constraint-Based Causal Discovery using Partial Ancestral Graphs in the presence of Cycles," *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, Proceedings of Machine Learning Research, Vol. 124, edited by J. Peters and D. Sontag, PMLR, 2020, pp. 1159–1168. URL https://proceedings.mlr.press/v124/m-mooij20a.html.

[14] Spirtes, P., Meek, C., and Richardson, T., "Causal Inference in the Presence of Latent Variables and Selection Bias," *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, edited by P. Besnard and S. Hanks, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995, p. 499–506.

[15] Glymour, C., Zhang, K., and Spirtes, P., "Review of Causal Discovery Methods Based on Graphical Models," *Frontiers in Genetics*, Vol. 10, 2019. doi:10.3389/fgene.2019.00524, URL https://www.frontiersin.org/journals/genetics/articles/10.3389/fgene.2019.00524.

[16] Camps-Valls, G., Gerhardus, A., Ninad, U., Varando, G., Martius, G., Balaguer-Ballester, E., Vinuesa, R., Diaz, E., Zanna, L., and Runge, J., "Discovering causal relations and equations from data," *Physics Reports*, Vol. 1044, 2023, pp. 1–68. doi:https://doi.org/10.1016/j.physrep.2023.10.005, URL https://www.sciencedirect.com/science/article/pii/S0370157323003411.

[17] Pearl, J., *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1988.

[18] Verma, T., and Pearl, J., "Causal Networks: Semantics and Expressiveness," *Uncertainty in Artificial Intelligence*, Machine Intelligence and Pattern Recognition, Vol. 9, edited by R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer, North-Holland, 1990, pp. 69–76. URL https://www.sciencedirect.com/science/article/pii/B9780444886507500111.

[19] Geiger, D., Verma, T., and Pearl, J., "Identifying independence in Bayesian networks," *Networks*, Vol. 20, No. 5, 1990, pp. 507–534.

[20] Spirtes, P., and Glymour, C., "An Algorithm for Fast Recovery of Sparse Causal Graphs," *Social science computer review*, Vol. 9, No. 1, 1991, pp. 62–72.

[21] Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., and Sejdinovic, D., "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science advances*, Vol. 5, No. 11, 2019, p. eaau4996.

[22] Runge, J., Gerhardus, A., Varando, G., Eyring, V., and Camps-Valls, G., "Causal inference for time series," *Nature Reviews Earth & Environmen*, Vol. 4, 2023, pp. 487–505. URL https://doi.org/10.1038/s43017-023-00431-y.

[23] S., S. G., Beucler, T., Tam, F. I.-H., Gomez, M. S., Runge, J., and Gerhardus, A., "Selecting robust features for machine-learning applications using multidata causal discovery," *Environmental Data Science*, Vol. 2, 2023, p. e27. doi:10.1017/eds.2023.21.

[24] Lubba, C. H., Sethi, S. S., Knaute, P., Schultz, S. R., Fulcher, B. D., and Jones, N. S., "catch22: CAnonical Time-series CHaracteristics," *Data Mining and Knowledge Discovery*, Vol. 33, No. 6, 2019, pp. 1821–1852. doi:10.1007/s10618-019-00647-x, URL https://doi.org/10.1007/s10618-019-00647-x.

[25] Rewicki, F., Denzler, J., and Niebling, J., "Is It Worth It? Comparing Six Deep and Classical Methods for Unsupervised Anomaly Detection in Time Series," *Applied Sciences*, Vol. 13, No. 3, 2023. doi:10.3390/app13031778, URL https://www.mdpi.com/2076-3417/13/3/1778.

[26] Liu, Q., and Paparrizos, J., "The Elephant in the Room: Towards A Reliable Time-Series Anomaly Detection Benchmark," *NeurIPS 2024*, 2024.