

# Causal discovery in a complex industrial system: A time series benchmark

**Søren Wengel Mogensen**

*Department of Automatic Control, Lund University, Lund, Sweden*

SOREN.WENGEL\_MOGENSEN@CONTROL.LTH.SE

**Karin Rathsman**

*The European Spallation Source, Lund, Sweden*

KARIN.RATHSMAN@ESS.EU

**Per Nilsson**

*The European Spallation Source, Lund, Sweden*

PER.NILSSON2@ESS.EU

**Editors:** Francesco Locatello and Vanessa Didelez

## Abstract

Causal discovery outputs a causal structure, represented by a graph, from observed data. For time series data, there is a variety of methods, however, it is difficult to evaluate these on real data as realistic use cases very rarely come with a known causal graph to which output can be compared. In this paper, we present a dataset from an industrial subsystem at the European Spallation Source along with its causal graph which has been constructed from expert knowledge. This provides a testbed for causal discovery from time series observations of complex systems, and we believe this can help inform the development of causal discovery methodology.

**Keywords:** Causal discovery, time series, causal graphs, benchmark data, European Spallation Source

## 1. Introduction

Datasets from engineered systems are of great value to the causal inference community for several reasons. First, system experts will often know the causal structure, or at least parts of it. Second, operator inputs to the system may be used to emulate interventions. This means that datasets from such systems are useful benchmarks, both for causal discovery and for causal effect estimation. In this paper, we present a benchmark dataset for causal discovery from time series data. This dataset was collected from a complex industrial system at the European Spallation Source, a neutron source facility in Lund, Sweden. It has several interesting properties in relation to causal discovery. Most importantly, a ground-truth causal graph is known, and it can therefore be used as a causal discovery benchmark.

In the remainder of this section, we give a short overview of causal discovery methods (Subsection 1.1) and discuss the use of engineered systems as benchmarks for causal discovery and causal inference (Subsection 1.2). Section 2 describes the data, its collection, the underlying physical system as well as its ground-truth causal graph. Section 3 provides visualizations and a simple analysis of the dataset. Section 4 highlights various properties of the dataset and of the underlying system that are important for causal discovery, and Section 5 provides a discussion of how the dataset can be used as a causal discovery benchmark.

The dataset can be accessed online (Rathsman et al., 2024), and R-code can be found at a companion website.<sup>1</sup> This website describes how the dataset can be preprocessed to facilitate causal analysis. It also contains additional analyses as well as help to get started with the dataset. Finally, a smaller version of the dataset as well as a preprocessed version of the dataset are available.

---

1. <https://soerenwengel.github.io/essdata>

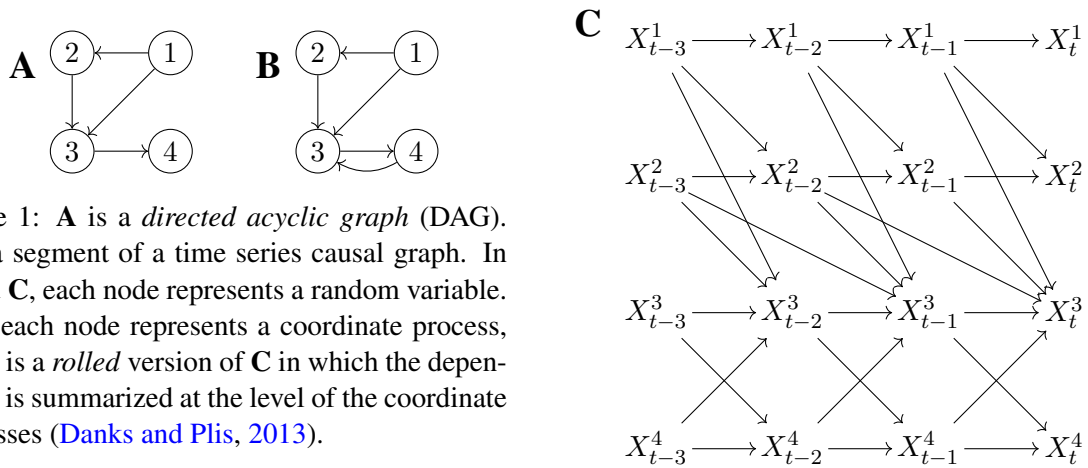


Figure 1: **A** is a *directed acyclic graph* (DAG). **C** is a segment of a time series causal graph. In **A** and **C**, each node represents a random variable. In **B**, each node represents a coordinate process, and **B** is a *rolled* version of **C** in which the dependence is summarized at the level of the coordinate processes (Danks and Plis, 2013).

### 1.1. Causal Discovery

*Causal discovery* methods learn causal graphs from observed data. It is common to assume the causal graph to be an unknown *directed acyclic graph* such that each node represents a random variable. Spirtes and Zhang (2018) give an overview of causal discovery in this context. In the case of *partial observation*, not every node/random variable is observable. In short, *constraint-based* methods use data to test observable constraints (most commonly conditional independence) that different causal graphs enforce and output graphs based on the observed constraints. *Score-based* methods instead define a certain criterion for the fit between causal structure and observed data, e.g., a penalized version of the maximum likelihood. Score-based algorithms then search for a graph to optimize this criterion. We will now briefly describe one approach to defining a causal model. This will allow us to define what we mean by a causal graph, also in the time series setting.

Assume first that we have a finite collection of random variables,  $X_1, X_2, \dots, X_n$ . A *structural causal model* (SCM, Pearl (2009); Peters et al. (2017)) assumes that for each  $i = 1, 2, \dots, n$

$$X_i = f_i(X_{\text{pa}_i}, \varepsilon_i)$$

where  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  are independent random variables and  $X_{\text{pa}_i} \subseteq \{X_1, X_2, \dots, X_n\}$  is the set of *parent variables* of the variable  $X_i$ . We construct the corresponding *causal graph* on nodes  $X_1, X_2, \dots, X_n$  by including an edge  $X_j \rightarrow X_i$  if and only if  $X_j \in X_{\text{pa}_i}$ . *Acyclicity* is often assumed, i.e., that the causal graph has no directed cycles,  $X_{i_1} \rightarrow X_{i_2} \rightarrow \dots \rightarrow X_{i_{m-1}} \rightarrow X_{i_1}$ . An SCM encodes both the observational and interventional distributions of the causal model (Pearl, 2009; Peters et al., 2017). It also implies a set of *Markov properties* such that conditional independence can be read off from the causal graph. This is used in many *constraint-based* causal discovery algorithms.

The SCMs can be extended to the time series setting straightforwardly, see Peters et al. (2013). We now have a multivariate time series,  $X_t = (X_t^1, \dots, X_t^n)$ , and we will assume

$$X_t^i = f_t^i(X_{\text{pa}_t^i}, \varepsilon_t^i)$$

where  $X_{\text{pa}_t^i}$  is the parent set of  $X_t^i$ . In this case, causal dependence, as represented by directed edges, is required to point forward in time, i.e.,  $X_{\text{pa}_t^i} \subseteq \{X_s^j\}_{s < t, j \in V}$ ,  $V = \{1, 2, \dots, n\}$ . One may, of course, impose regularity conditions, e.g., assuming that if  $X_{\text{pa}_t^i} = \{X_{t_1}^{i_1}, X_{t_2}^{i_2}, \dots, X_{t_m}^{i_m}\}$ , then  $X_{\text{pa}_{t-1}^i} = \{X_{t_1-1}^{i_1}, X_{t_2-1}^{i_2}, \dots, X_{t_m-1}^{i_m}\}$ . We can again define a causal graph,  $\mathcal{D}_X$ , with nodes  $\{X_t^i\}_{t \in \mathbb{Z}, i \in V}$  such that  $X_s^i \rightarrow X_t^j$  if and only if  $X_s^i \in X_{\text{pa}_t^j}$ . This graph is infinite as there is a node for each random variable in the time series. We may use a different graphical representation, a *rolled graph* (Danks and Plis, 2013), such that each node represents a *coordinate process*,  $X^i = \{\dots, X_{-2}^i, X_{-1}^i, X_0^i, X_1^i, X_2^i, \dots\}$ . In this representation, we have  $i \rightarrow j$  if and only if there exists  $s$  and  $t$  such that  $X_s^i \rightarrow X_t^j$  in  $\mathcal{D}_X$  as defined above. Figure 1C shows an example of a segment of a causal graph,  $\mathcal{D}_X$ , when  $n = 4$ . Figure 1B shows the corresponding rolled graph (assuming that the edges in the entire  $\mathcal{D}_X$  are described by those in the segment shown in the figure). We will think of the ground-truth causal graph (Subsection 2.4) as a rolled graph of a time series SCM.

### 1.1.1. CAUSAL DISCOVERY FROM TIME SERIES

When we observe data from a stochastic process, we will most often know the time index of each observation in which case there is a known temporal ordering of the observed values as reflected in the definition of a time series SCM. This is, of course, advantageous in causal discovery as it restricts the possible causal structures significantly. On the other hand, causal discovery from stochastic processes comes with other challenges, some unique to the time series setting as described in the following.

In recent years, many methods for causal discovery from time series data have appeared. These methods can be classified according to the assumptions they make. First, there is a natural divide between methods assuming that the underlying causal model evolves in continuous time and methods assuming a causal system evolving in discrete time. Second, methods tend to restrict the model class (e.g., point processes, discrete-time stochastic processes, stochastic differential equations, or (semi)parametric subclasses thereof) they are considering, and the available methods therefore depend on the type of data we wish to analyze. Third, some methods assume *causal sufficiency*, i.e., full observation of the causal system, while others do not. Fourth, methods may assume stationarity or nonstationarity of the observed processes. In the following, we describe some examples of methods with various combinations of these characteristics.

Causal discovery in stochastic processes naturally builds on prior work in the DAG-based causal discovery literature. It is possible to adapt, e.g., the classical FCI-algorithm to the time series setting and further exploit the temporal structure of the time series (Chu and Glymour, 2008; Entner and Hoyer, 2010; Malinsky and Spirtes, 2018). In stochastic processes, there is also the question of sampling frequency, and *subsampling*, i.e., using a sampling frequency which is lower than the ‘causal frequency’ may pose difficulties for causal discovery which means that specialized methods are required (Danks and Plis, 2013; Gong et al., 2015; Hyttinen et al., 2016). Gong et al. (2017) consider causal discovery under *temporal aggregation* in which the observed data consists of local averages. Other methods include non-Gaussian structural vector autoregressive models (Hyvärinen et al., 2010). It is common to assume stationarity of the observed time series, however, methods that exploit nonstationarity also exist (Hyvärinen and Morioka, 2016).

Constraint-based algorithms for time series data may either test conditional independence, or use stochastic process-analogues such as *local independence* in continuous-time processes and *Granger*

*causality* in discrete-time processes. [Mogensen et al. \(2018\)](#); [Mogensen \(2020\)](#); [Absar and Zhang \(2021\)](#) describe causal discovery from continuous-time processes using local independence constraints and [Eichler \(2013\)](#) describes causal discovery in discrete-time stochastic processes using Granger causality.

Data sampled from complex dynamical systems is found in many fields of science. In this paper, we consider data from an industrial system. This is mostly due to the fact that such a system may come with substantial expert knowledge of its functioning and therefore provide a useful benchmark. In engineered systems that are less well-understood, causal discovery can augment system understanding. In addition, causal discovery from complex time series has a large potential for other applications, for instance, health registry-based research, Earth system sciences ([Runge et al., 2019](#)), and economics ([Hall-Hoffarth, 2022](#)) as these are examples of fields that generate data from interacting stochastic processes.

## 1.2. Engineered Systems as Benchmarks

Benchmark data is useful for testing and comparing methods. Good benchmarks should come with a ground truth against which we can compare outputs and they should also resemble the data that we are actually interested in. Obtaining benchmark data with these properties may be difficult. One possibility is the use of *synthetic* data, i.e., using data from computer simulations. A number of synthetic datasets for causal discovery benchmarking are available ([Smith et al., 2011](#); [Runge et al., 2019](#); [Lawrence et al., 2020](#)). [Reisach et al. \(2021\)](#) show that it may not be easy to simulate realistic data, or that simulated data may show artifacts. This is a concern as causal discovery methods may exploit these artifacts that are products of simulation algorithms and not characteristics of actual causal structures. For this reason, it seems prudent to use real data for causal discovery benchmarking. Of course, in this case the issue is that for real-world data the underlying causal structure may not be known.

We believe that engineered systems can provide useful data in this context. First, their high-level behavior is well-understood as they serve a known purpose. Second, experts know how they are constructed which means that the causal structure is fully, or at least partially, known. Third, interventions are common, well-defined, and feasible, for instance, they occur when system inputs are changed. In the next section, we describe the dataset that we present. This data comes with an extensive understanding of the underlying system which makes it interesting as a causal discovery benchmark.

## 2. Data

The dataset we provide consists of measurements from a complex system. We explain the context and relevance of this system (Subsection 2.1). We then describe the system itself in detail (Subsection 2.2) and the data collection (2.3). Subsection 2.4 introduces the ground-truth causal graph. Appendix A describes metadata and explains how the csv-files containing the data are structured.

### 2.1. The European Spallation Source

The European Spallation Source ERIC (ESS) is a neutron source research facility in Lund, Sweden, along with a data center located in Copenhagen, Denmark. The ESS is under construction at the time of writing, however, some systems are already operational. Neutron sources produce beams of

neutrons and distribute them to experimental stations. An experimental technique known as *neutron scattering* is then used to study matter through its dispersal of free neutrons, providing a valuable tool for research in, for example, physics, biology, and materials science. Once construction of the ESS is completed, a superconducting linear accelerator will accelerate protons to hit a target which ejects neutrons in a process known as *nuclear spallation*. Upon its completion, scientists will be able to apply for access to the facilities to conduct experiments using various experimental stations. The projected construction cost of the ESS is in the billions of euros, and reliability of the ESS systems is, of course, paramount to minimize system downtime and maximize research output.

*Cryogenics*, that is, cooling to very low temperatures as well as material science at very low temperatures, is essential to the operation of the ESS facility as several components should operate at temperatures close to absolute zero. A *cryoplant* is a system which cools and distributes a coolant. In the following, we describe the accelerator cryoplant, *ACCP*, which is the largest of three cryoplants at ESS.

## 2.2. The ACCP System

The data we provide is sampled from the ACCP system of the ESS which is an industrial refrigeration system. The purpose of the ACCP is to cool the linear accelerator cavities. In a high-level description, coolant (helium) is cooled down to 4.5 Kelvin and distributed to a liquid helium bath that cools the cavities. The coolant is then returned to be re-cooled, and the overall structure of the system is therefore circular in this sense. In addition, there are several interconnected loops. Figure 2 gives a *functional* overview of the system, see Subsection 2.2.1 for details on how to interpret this diagram. More details on the ACCP and its construction are provided in Garoby et al. (2017).

### 2.2.1. FUNCTIONAL DIAGRAM

Figure 2 is provided to illustrate how the ACCP works. In this diagram, each node represents a subsystem and lines indicate a flow of coolant in the direction of the arrow. The section of the ACCP between the two dashed lines in Figure 2 is the *cold box*. Starting from the *warm end* (see Figure 2), the ACCP consists of three compressors (SP, LP, and HP), an oil removal and gas management unit (OG), six heat exchangers (HX1, HX2, HX3, HX4, HX5, HX6), six turbines (T1, T2, T3, T4, T5, T6), three adsorbers (A1, A2, A3), a thermal shield (TS), three compressors inside the cold box (C1, C2, C3), a helium subcooler (S), a dewar (D), and a test vessel (TV). From the cold end, the coolant proceeds to a liquid helium bath (HS) where it cools the cavities of the linear accelerator before returning to the ACCP. No measurements are provided from the liquid helium bath, and therefore there is an unobserved subsystem below the ACCP in the diagram. The thermal shield (TS) can also be thought of as partially unobserved as only the pressure, temperature, and flow of the coolant supplied to/from the thermal shield are observed. No measurements from the TS itself are available.

There are three additional subsystems (TG, LS, and CG) that are not shown in the functional diagram. They support more than one subsystem. The turbine general (TG) supports the six turbines, LS the low and sub-atmospheric pressure compressors (LP and SP), and the compressor general (CG) supports the cold compressors (C1, C2 and C3).

A central feature of the ACCP is the circulation of helium. For this reason, it is also natural that some measurements are made between the subsystems in the diagram in Figure 2. The interfaces SPWE, LPWE, MPWE, and HPWE (SP/LP/MP/MP line, warm end) represent measurements made

at the warm end dashed line in the diagram, and such measurements are also available in the data. These interfaces will be referred to as subsystems in the remainder of the paper, despite the fact that they are links, not actual subsystems.

While the diagram in Figure 2 is useful, it only describes the coolant flow in the ACCP. Subsection 2.4 describes what is believed to be the causal structure of the system. One should note that the subsystems represented are slightly different between the diagram in Figure 2 and the causal graph in Figure 3 (see also Subsection 2.4). The dataset and the remainder of this paper use the subsystems represented in the causal graph in Figure 3.

### 2.3. Data Collection

The ACCP first started operation in the summer of 2020 and has been in intermittent operation since. Operational data has been collected and stored in a central database. The state of each subsystem is described by a number of *process variables* (PVs) that are sampled and recorded repeatedly over time (Table 1). A PV may, for instance, be a measured temperature, pressure, or flow. We chose three time periods (Period 1: 2022-12-30 17:00 to 2023-01-02 08:00, Period 2: 2023-01-05 19:00 to 2023-01-09 09:00, Period 3: 2023-01-13 16:00 to 2023-01-16 06:00, all local time) and measurements from these three periods constitute the dataset. Over time, the ACCP has been running in different ‘modes’ as operators provide various input values to the system, e.g., set points that are ‘desired’ values for PVs. Periods 1, 2, and 3 were chosen such that these input values are constant within each period, but different across periods. This is related to the notion of *environments* that may represent different interventional settings (Peters et al., 2017).

Obviously, a complete description of the system state cannot be collected, however, the PVs that are provided with the dataset are thought to be the most important variables for describing the system state (some subsystems are unobserved, though, see Subsections 2.2.1 and 2.4). The sampling frequency was chosen by system designers to monitor the system in real time, and the system is expected to evolve quite slowly compared to the sampling frequency. That is, the sampling frequency is expected to be much higher than the frequency with which relevant changes might occur. In industrial systems, it may occur that observations are only recorded with some delay, and that this delay is different for each PV. However, no such discrepancies are expected in this dataset.

No data cleaning was done and the dataset is therefore raw data as measured during system operation. We believe that data cleaning is an integral, and nontrivial, aspect of causal discovery, and therefore we present the raw data such that users may experiment with different approaches. One should note that sensor data may be compromised in several ways. Sensors may malfunction or the data acquisition may be noisy for other reasons. Therefore, observed PV values may be outside of the range of possible values. A PV may also be ‘frozen’ at a certain value or make sudden jumps due to sensor malfunctioning. While the dataset contains raw data, the companion website provides some guidance on data cleaning.

### 2.4. Ground-truth Causal Graph

In this section, we present the *causal graph* of the system (Figure 3). The causal graph was constructed by system experts from their knowledge of the system. The causal structure is represented by a graph at the subsystem level, i.e., every node in the causal graph represents a subsystem and therefore a collection of PVs (in Figure 1B, each node represented a single coordinate process). In this sense, the graph is ‘rolled’ as every node represents a (possibly multivariate) stochastic process,

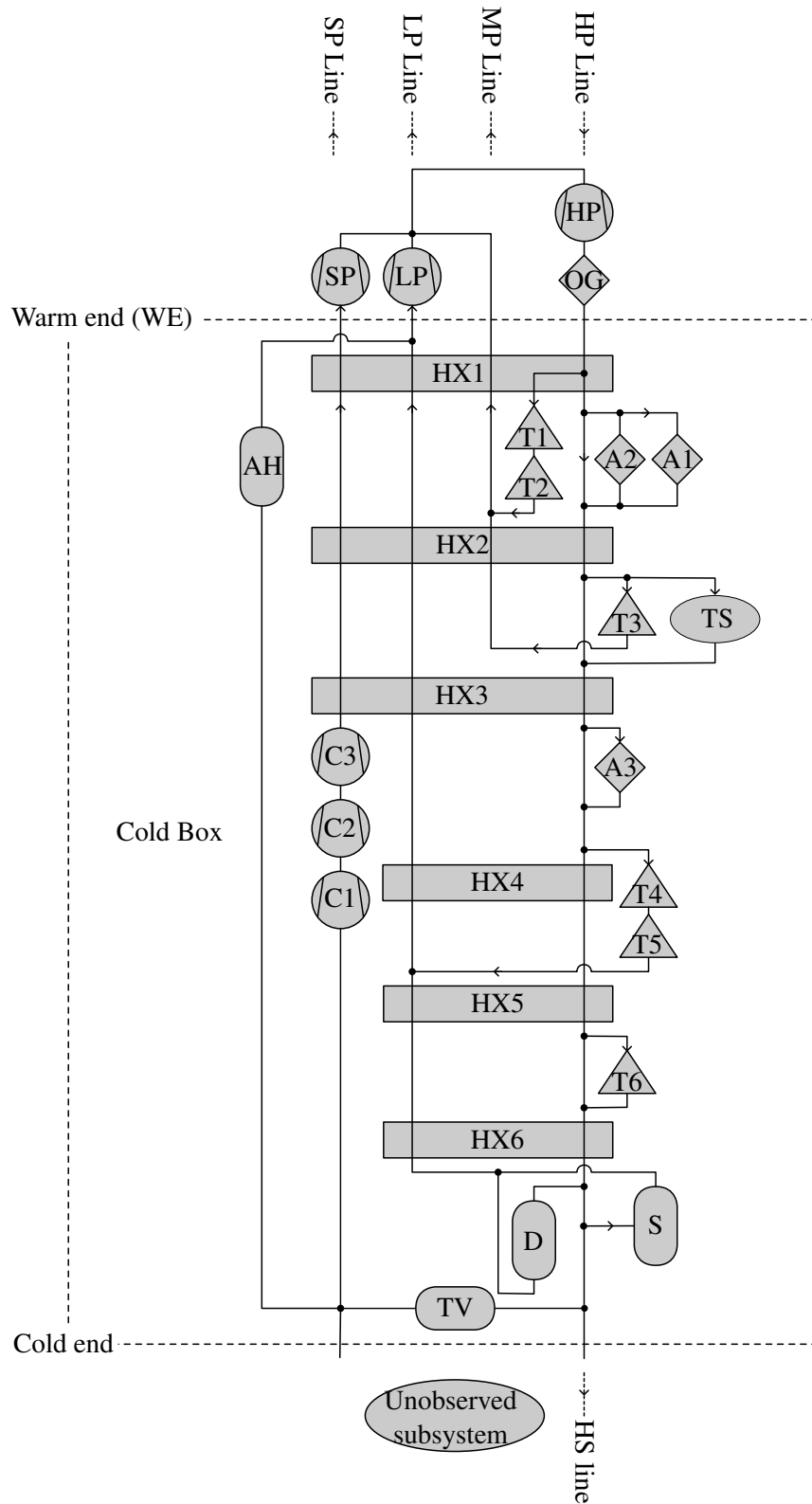


Figure 2: Diagram of the ACCP.

Index	Description	No. of PVs	Index	Description	No. of PVs
T1	Turbine 1	5	C1	cold Compressor 1	5
T2	Turbine 2	5	C2	cold Compressor 2	6
T3	Turbine 3	8	C3	cold Compressor 3	6
T4	Turbine 4	5	CG	Compressor General	11
T5	Turbine 5	6	SP	Sub-atm. Prs. compressor	17
T6	Turbine 6	8	SPWE	SP line Warm End	2
TG	Turbine General	1	LP	Low Pressure compressor	16
A1	Adsorber 1 80 K	3	LPWE	LP line Warm End	2
A2	Adsorber 2 80 K	4	LS	LP+SP compressor	8
A3	Adsorber 3 20 K	4	MPWE	MP line Warm End	3
HX1	Heat eXchanger 1	2	HP	High Pressure compressor	22
HX2	Heat eXchanger 2	9	HPWE	HP line Warm End	5
HX3	Heat eXchanger 3	2	OG	Oil and Gas removal	15
HX4	Heat eXchanger 4	3	AH	Ambient Heater	5
HX5	Heat eXchanger 5	2	D	Dewar	9
HX6	Heat eXchanger 6	5	S	Subcooler	7
TV	Test Vessel	8	TS	Thermal Shield*	8
HS	Helium Supply*	6			

Table 1: List of subsystems and number of process variables (PVs) in each subsystem. \* indicates that the subsystem is partially unobserved.

not a single random variable. In some causal graphs of time series, each node represents a single random variable, thereby also showing the lag structure of the causal system, i.e., how far back a dependence goes. In our case, the underlying system is developing very fast, and an explicit lag structure would therefore be an artifact of data, not a characteristic of the causal system. Moreover, the processes are not sampled at the same points in time which means that there is no inherent lag structure in the data.

There is a total of 35 subsystems, each of which is a collection of physical components (see Table 1). As explained below only a subset of them is used in the causal graph. Two of the subsystems are (partially) unobserved (HS and TS), however, the unobserved systems are quite sparsely connected to the observed systems (Figure 3). Subsystems SPWE, LPWE, MPWE, HPWE, and the heat exchangers (HX) are not represented in the causal graph. The HXs are split between the turbine systems, the SPWE is measured along the edge  $C3 \rightarrow SP$ , LPWE is measured between T5 and LP, MPWE is measured between T2/T3 and LS, and HPWE is measured along the edge from OG to T1. The dewar (D) is included in the T6 subsystem based on background knowledge and therefore not shown in the causal graph. The TG subsystem interacts with all the turbines (T1, T2, T3, T4, T5, T6). However, only a quite weak dependence is expected in stable operations and therefore TG is also not shown in the causal diagram.

Each edge in the graph denotes a direct causal link, in most cases an edge also represents a direct physical connection, e.g., the flow of coolant from one subsystem to another. The graph is directed (no bidirected edges) as the observed system is self-contained, safe for the partially unobserved subsystems as indicated in Figures 2 and 3, and system experts believe any confounding to be negligible (we also see from the structure in the causal graph, Figure 3, that the unobserved



subsystems are only sparsely connected to the observed subsystems). Each edge in the causal graph is annotated with ‘edge strength’ based on system knowledge (weak or strong link). A system expert classified each edge,  $S_1 \rightarrow S_2$ , as strong/weak based on a qualitative understanding of the influence of the parent subsystem,  $S_1$ , on the child subsystem,  $S_2$ . In some cases, a weak edge may also correspond to a causal connection that only manifests when certain conditions are met. In Figure 3, thick edges correspond to strong causal connections, and thin edges correspond to weak causal connections. An adjacency matrix,  $A$ , is available along with the dataset. Each row and column of this matrix corresponds to a subsystem. For subsystems  $i$  and  $j$ ,  $A_{ji} = 0$  indicates that there is no edge from  $i$  to  $j$ ,  $A_{ji} = 1$  indicates that there is a weak edge from  $i$  to  $j$ , and  $A_{ji} = 2$  indicates that there is a strong edge from  $i$  to  $j$ .

As the causal graph was constructed from background knowledge, it is not possible to ensure its ‘correctness’. It is not even obvious that a single ‘correct’ causal graph can be specified for a problem of this complexity. The division into subsystems could also be done in different ways. However, the causal graph which we present as the ground truth was constructed by engineers that have worked extensively with this system. The system is complex, however, it is also well-documented and as been running intermittently since 2020. Moreover, the design used in the ACCP has been used for more than 20 years. As the system circulates coolant many connections and interdependencies are obvious from the system layout.

### 3. Analysis

In this section, we visualize the data and compute some simple statistics. More analysis is available online.<sup>2</sup> The dataset consists of 233 processes, all of which are measured values, e.g., temperatures, pressures, and flows. We denote an observed value by  $x_t^p$  where  $t$  is a time index in  $[0, 1]$  (after normalization of the time interval) and  $p \in \{1, \dots, 233\}$  is a process index. That is,  $x_t^p$  is the measured value of process  $p$  in subsystem  $s$  at time  $t$ . We use  $X^p$  to denote a stochastic process corresponding to the observations  $x_t^p$ , and we say that  $X^p$  is a *process variable* (PV). We let  $\mathcal{T}_p = \{t_1^p, t_2^p, \dots, t_{m_p}^p\}$  denote the set of time points,  $t$ , such that we have observed process  $p$  at time  $t$ . The data is not sampled *regularly*, i.e.,  $\mathcal{T}_p$  and  $\mathcal{T}_{p'}$  need not be equal when  $p \neq p'$ . Moreover, there is not a fixed sampling frequency for each PV, i.e.,  $t_{l+1}^p - t_l^p \neq t_{l'+1}^p - t_{l'}^p$  may occur for  $l \neq l'$ , see Figure 4. This figure also highlights the fact that some PVs seem to change their values in a discrete manner, possibly due to the precision of measurement.

The PVs exhibit different behaviors which is to be expected from the fact that they represent different types of measurements, e.g., temperatures, pressures, and flows (measurement units are provided in the metadata). Figure 5 shows 62 PVs over the first hour of Period 1. We see that some PVs seem to be pure noise, some show a strong trend, and some oscillate (data in the figure is aggregated for each second, and then subsampled to one observation per 10 seconds). One process shows a sudden drop. Some PVs, e.g., the first three in the first row in Figure 5, look like noise during the first hour, but show clear trends over the 55-hour span in Figure 7 in Appendix B.

The sampling frequency is known to be quite high, so it may make sense to reduce the data. It is, for example, possible to subsample the time series (for  $p$  consider only observations  $\tilde{\mathcal{T}}_p \subsetneq \mathcal{T}_p$ ) or aggregate them over intervals, e.g., compute average PV values for each second of observation to achieve more regular time series. In the following correlation and Granger causality analysis, we

2. <https://soerenwengel.github.io/essdata>

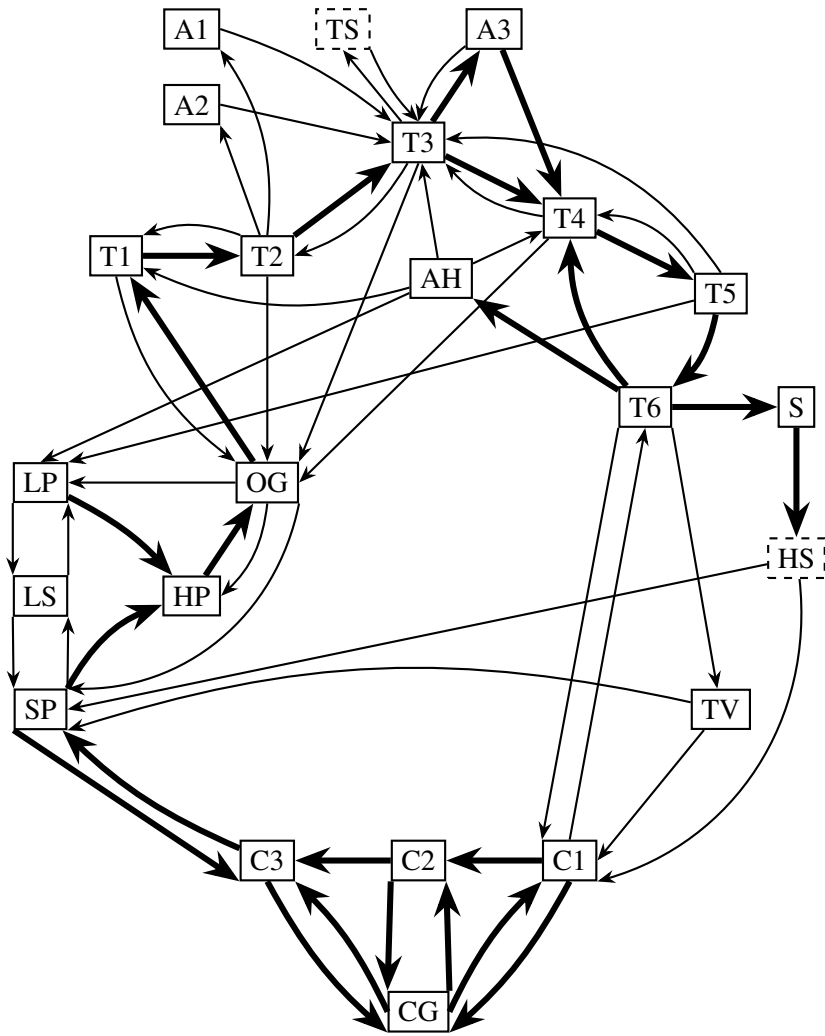


Figure 3: Causal graph. A dashed rectangle indicates that the subsystem is unobserved. Thick edges indicate a strong connection.

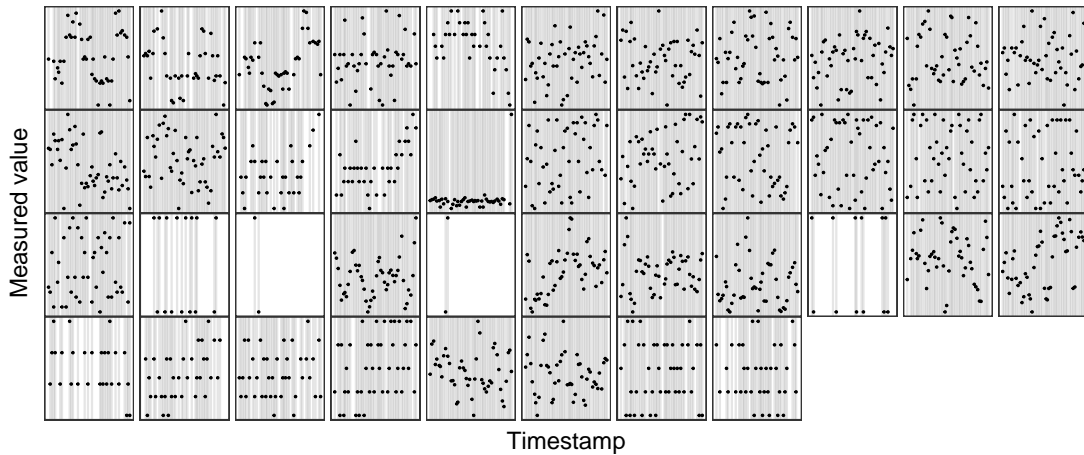


Figure 4: Raw data from 10 seconds (41 PVs). The vertical axis is different across subplots. Each point indicates an observation (timestamp and measured value). A vertical line has been added through each point to highlight the irregular nature of the sampling.

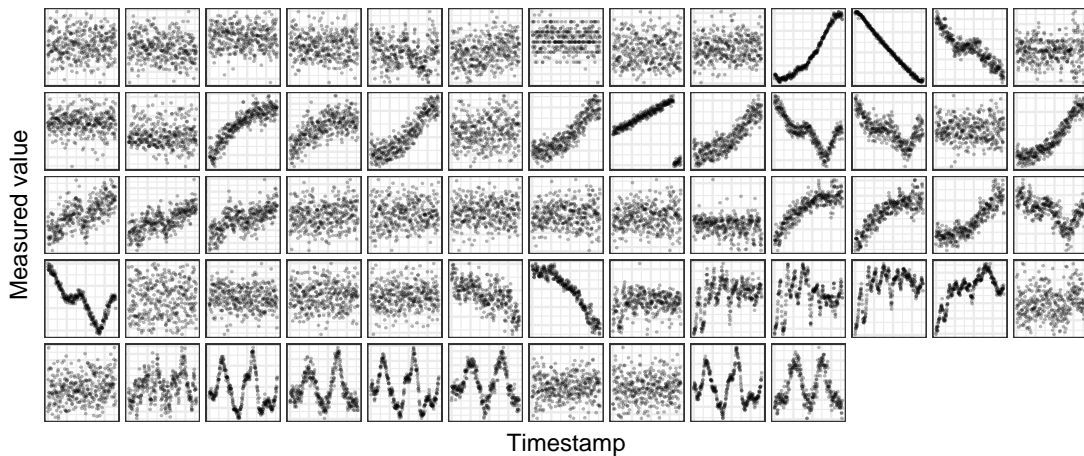


Figure 5: Aggregated and subsampled data from 62 PVs during one hour of operation. The vertical scale is different across subplots.

have done the latter. However, data reduction may complicate causal discovery ([Danks and Plis, 2013](#); [Gong et al., 2017](#)).

Figure 8 in Appendix B presents a correlation plot (Pearson correlation) of a subset of the PVs in Period 1. One should note that strong correlations may be *spurious* as they ignore the autocorrelation. On the other hand, we do observe that some PVs corresponding to the same subsystem are highly correlated which seems reasonable as they describe different aspects of the same system component. Similarly, Figure 9 in Appendix B shows  $p$ -values from tests of Granger noncausality. For an ordered pair of subsystems,  $(i, j)$ , this tests if the past of  $j$  is predictive of the present of  $i$  when conditioning on the past of  $i$ . This simple approach also recovers some of the structure from the causal graph.

## 4. Learning from ACCP Data

We imagine that the ACCP dataset will be useful for benchmarking causal discovery methodology. In this section, we discuss ways to use the data and the challenges that various properties of the data pose for causal discovery. The ACCP is a quite complicated system, and the ACCP dataset requires some preprocessing before applying causal discovery methods. Examples of how this preprocessing can be done are available online,<sup>3</sup> and a preprocessed dataset is also available to facilitate a quick look at the data.

The most straightforward use of the dataset, and its causal graph, is to compare output graphs from causal discovery with the causal graph. One variation is to assume partial observation such that only some subsystems are observed. Mogensen and Hansen (2020) use *directed mixed graphs* as graphical representations of partially observed systems. In this case, such graphs could be the learning target and they can easily be computed from the causal graph given a partition of the subsystems into observed/unobserved. It is also possible to input different types of information to a structure learning algorithm, e.g., the subsystem grouping may be taken as prior knowledge. Alternatively, one can input partial knowledge of the causal graph and use data to refine this knowledge.

The sampling frequency is quite high and this means that subsampling is possible, maybe even advantageous. Comparisons can be made for different subsampling frequencies. Other types of data preprocessing may be important before the application of a causal discovery algorithm, and this can also be tested.

The dataset is divided into three time periods. In each time period, the ACCP system ran without operators making any changes to input parameters. On the other hand, input parameters are different across the three time periods (the values of these parameters are not provided with the dataset). It may be possible for causal discovery algorithms to leverage this information (Peters et al., 2017).

### 4.1. Challenges

The dataset represents a number of challenges that are common when learning causal structure from real-world time series data. In the following, we discuss the properties that may pose issues when analyzing the data.

**Partial Observation** In causal modeling, the issue of *partial observation* (i.e., whether the entire system is observed) is central. In the context of stochastic process models, partial observation (or hidden confounding) often refers to the possible existence of unobserved coordinate processes, or unobserved PVs in our terminology. As described in Subsection 2.4, the system is essentially a closed system, and only two of its subsystems are (partially) unobserved. From the causal graph and the connections of the unobserved subsystems, we see that the assumption of no hidden confounding seems reasonable as only a couple of subsystems are unobserved, and these subsystems are only sparsely connected to the observed subsystems.

However, there is the potential of another type of ‘partial observation’ as the behavior of the physical ACCP system cannot be captured completely by any dataset of a reasonable size. Instead it must be described by certain well-chosen measurements of temperatures, pressures, etc. The measurements from the ACCP do clearly not, in themselves, represent the ‘entire system’. However, the measured PVs were chosen by the system manufacturer with the purpose of monitoring and understanding system behavior.

---

3. <https://soerenwengel.github.io/essdata>

**Hierarchical Structure** The data has a natural hierarchical structure. There is a large number of PVs and these are divided into 35 subsystems. The number of subsystems is manageable for a human to understand, and each subsystem represents a physical and well-defined component of the system. This means that the causal interpretation on this level is quite appealing. On the other hand, data is available at the lower PV-level, and this means that the causal structure should be inferred from a large number of underlying processes.

**Sampling and modeling frequencies** The underlying physical processes can be conceptualized as continuous-time stochastic processes, however, the sampling is done in discrete time. The sampling frequency is fairly high, typically 14 Hz, and it may be beneficial to choose a lower frequency for modeling. In addition, the sampling is *irregular* as described above as PVs are not sampled at the same time points. This complicates the application of time series methods that assume regular sampling.

**Cyclic dependence** The data-generating mechanism has cycles, and in fact the ACCP is dominated by a large cyclic component which roughly corresponds to the flow of the coolant.

**Different timescales** Different parts of the system may operate at different timescales, i.e., some causal effects may manifest faster than others, complicating causal learning.

## 5. Discussion

Benchmark data is important to test causal discovery methodology, and real-world data avoids the pitfalls of synthetic data. In the dataset presented in this paper, the ground-truth causal graph was constructed from expert knowledge of the system, and so was the division of observed processes into subsystems. The measurements describe what is believed to be the most important aspects of the physical systems, however, they are, of course, not exhaustive. For these reasons, one has to keep in mind that other graphs may be equally correct in some sense. This necessarily means that if algorithms fare similarly on this dataset, one should be careful not to overinterpret small differences in their performances. Therefore, the benchmark that this paper describes is perhaps best thought of as an opportunity for researchers to apply their methods to a real and complex dataset in which the gist of the causal structure is known.

## Acknowledgments

Søren Wengel Mogensen was supported by a DFF-International Postdoctoral Grant (0164-00023B) from Independent Research Fund Denmark. Søren Wengel Mogensen is a member of the ELLIIT Strategic Research Area at Lund University. The authors thank the reviewers and the area chair for their comments and suggestions.

## References

- Saima Absar and Lu Zhang. Discovering time-invariant causal structure from temporal data. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, 2021.
- Tianjiao Chu and Clark Glymour. Search for additive nonlinear time series causal models. *Journal of Machine Learning Research*, 9(32):967–991, 2008.

- David Danks and Sergey Plis. Learning causal structure from undersampled time series. In *NIPS 2013 Workshop on Causality*, 2013.
- Michael Eichler. Causal inference with multiple time series: Principles and problems. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371 (1997):20110613, 2013.
- Doris Entner and Patrik O Hoyer. On causal discovery from time series data using fci. In *Proceedings of The Fifth European Workshop on Probabilistic Graphical Models (PGM)*, 2010.
- Roland Garoby, A Vergara, H Danared, I Alonso, E Bargallo, B Cheymol, Christine Darve, Mohammad Eshraqi, H Hassanzadegan, A Jansson, et al. The European Spallation Source design. *Physica Scripta*, 93(1):014001, 2017.
- Mingming Gong, Kun Zhang, Bernhard Schoelkopf, Dacheng Tao, and Philipp Geiger. Discovering temporal causal relations from subsampled data. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. PMLR volume 37, 2015.
- Mingming Gong, Kun Zhang, Bernhard Schölkopf, Clark Glymour, and Dacheng Tao. Causal discovery from temporally aggregated time series. In *Proceedings of the 33rd Conference on Uncertainty in Artificial Intelligence (UAI)*, 2017.
- Emmet Hall-Hoffarth. Causal discovery of macroeconomic state-space models. *arXiv preprint arXiv:2204.02374*, 2022.
- Antti Hyttinen, Sergey Plis, Matti Järvisalo, Frederick Eberhardt, and David Danks. Causal discovery from subsampled time series data by constraint optimization. In *Proceedings of the Eighth International Conference on Probabilistic Graphical Models (PGM)*, 2016.
- Aapo Hyvarinen and Hiroshi Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2016.
- Aapo Hyvärinen, Kun Zhang, Shohei Shimizu, and Patrik O. Hoyer. Estimation of a structural vector autoregression model using non-Gaussianity. *Journal of Machine Learning Research*, 11 (56):1709–1731, 2010.
- Andrew R. Lawrence, Marcus Kaiser, Rui Sampaio, and Maksim Sipos. Data generating process to evaluate causal discovery techniques for time series data. *Causal Discovery & Causality-Inspired Machine Learning Workshop at Neural Information Processing Systems*, 2020.
- Daniel Malinsky and Peter Spirtes. Causal structure learning from multivariate time series in settings with unmeasured confounding. In *Proceedings of 2018 ACM SIGKDD Workshop on Causal Discovery*. PMLR volume 92, 2018.
- Søren Wengel Mogensen. Causal screening in dynamical systems. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.
- Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1):539–559, 2020.

- Søren Wengel Mogensen, Daniel Malinsky, and Niels Richard Hansen. Causal learning for partially observed stochastic dynamical systems. In *Proceedings of the 34th Conference on Uncertainty in Artificial Intelligence (UAI)*, 2018.
- Judea Pearl. *Causality*. Cambridge University Press, 2009.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. Causal inference on time series using restricted structural equation models. In *Advances in Neural Information Processing Systems*, 2013.
- Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. MIT Press, 2017.
- Karin Rathsman, Per Nilsson, and Søren Wengel Mogensen. A time series benchmark, February 2024. URL <https://doi.org/10.5281/zenodo.10641290>.
- Alexander Reisach, Christof Seiler, and Sebastian Weichwald. Beware of the simulated DAG! Causal discovery benchmarks may be easy to game. In *Advances in Neural Information Processing Systems*, 2021.
- Jakob Runge, Sebastian Bathiany, Erik Bollt, Gustau Camps-Valls, Dim Coumou, Ethan Deyle, Clark Glymour, Marlene Kretschmer, Miguel D Mahecha, Jordi Muñoz-Marí, et al. Inferring causation from time series in Earth system sciences. *Nature communications*, 10(1):2553, 2019.
- Stephen M Smith, Karla L Miller, Gholamreza Salimi-Khorshidi, Matthew Webster, Christian F Beckmann, Thomas E Nichols, Joseph D Ramsey, and Mark W Woolrich. Network modelling methods for fmri. *Neuroimage*, 54(2):875–891, 2011.
- Peter Spirtes and Kun Zhang. Search for causal models. In Marloes Maathuis, Mathias Drton, Steffen Lauritzen, and Martin Wainwright, editors, *Handbook of Graphical Models*, chapter 18, pages 439–470. CRC Press, 2018.

## Appendix A. Data Structure and Overview

The dataset is available online as a compressed archive (Rathsman et al., 2024). Metadata is provided in the file `systemoverview.csv` (in the following, *the system overview*). This file has a row for each PV containing

- PV name (Name),
- description of the PV (Description),
- measurement unit (Unit),
- subsystem index (Subsystem),
- subsystem description (Subsystem\_Description),
- sensor type index (SensorType), and
- sensor type description (SensorType\_Description).

The dataset is structured in a hierarchy with levels *period* (folder), *subsystem* (folder), and *PV* (csv-file). The file `Period_p/Subsystem/Name.csv` contains measurements from a single PV in Period  $p$ , see also Figure 6. Each measurement in such a csv-file is simply a timestamp and a measured value. The units of measurement are different for different PVs as specified in the system overview.

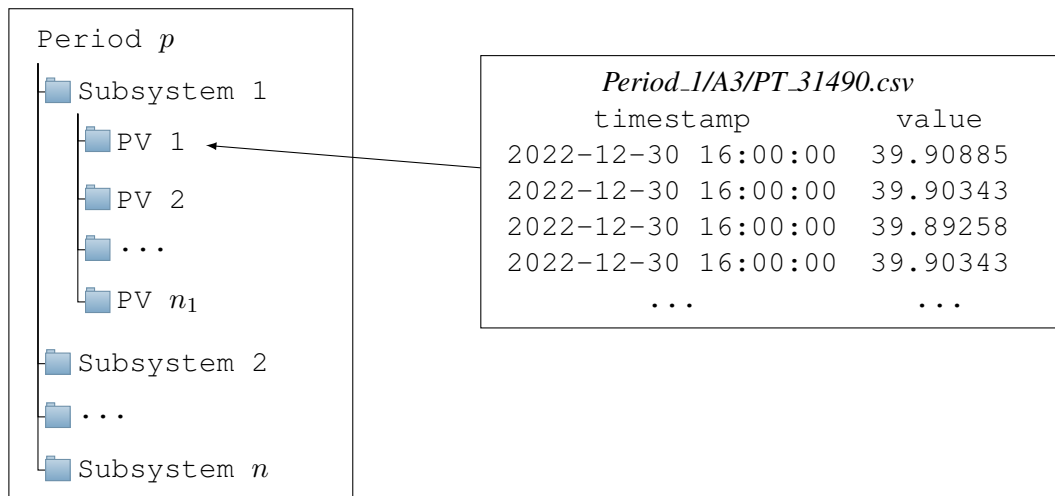


Figure 6: Overview of the data structure. A csv-file provides the data for a single PV in a time period  $p$ .



**Appendix B. Description of the Data**

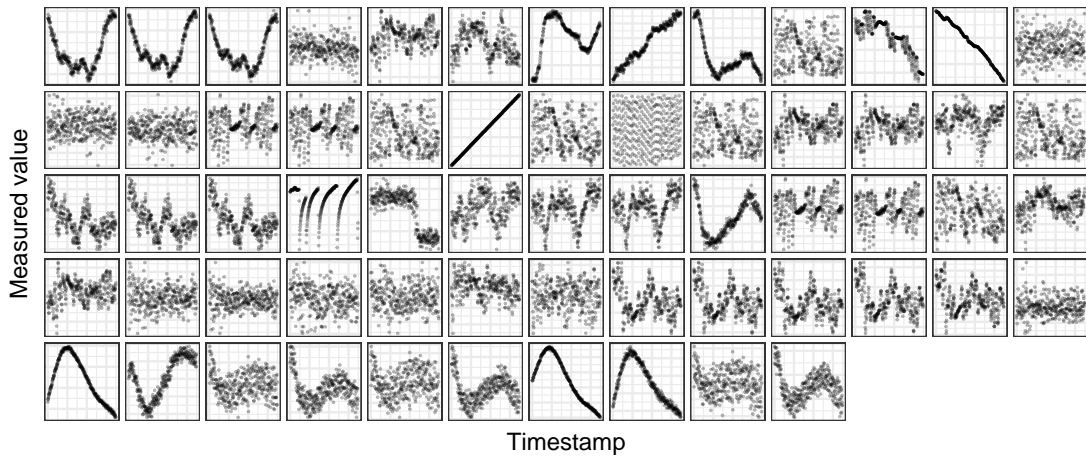


Figure 7: Aggregated data (one observation per 10 minutes). The PVs are the same as in Figure 5. This plot covers 55 hours of operation.

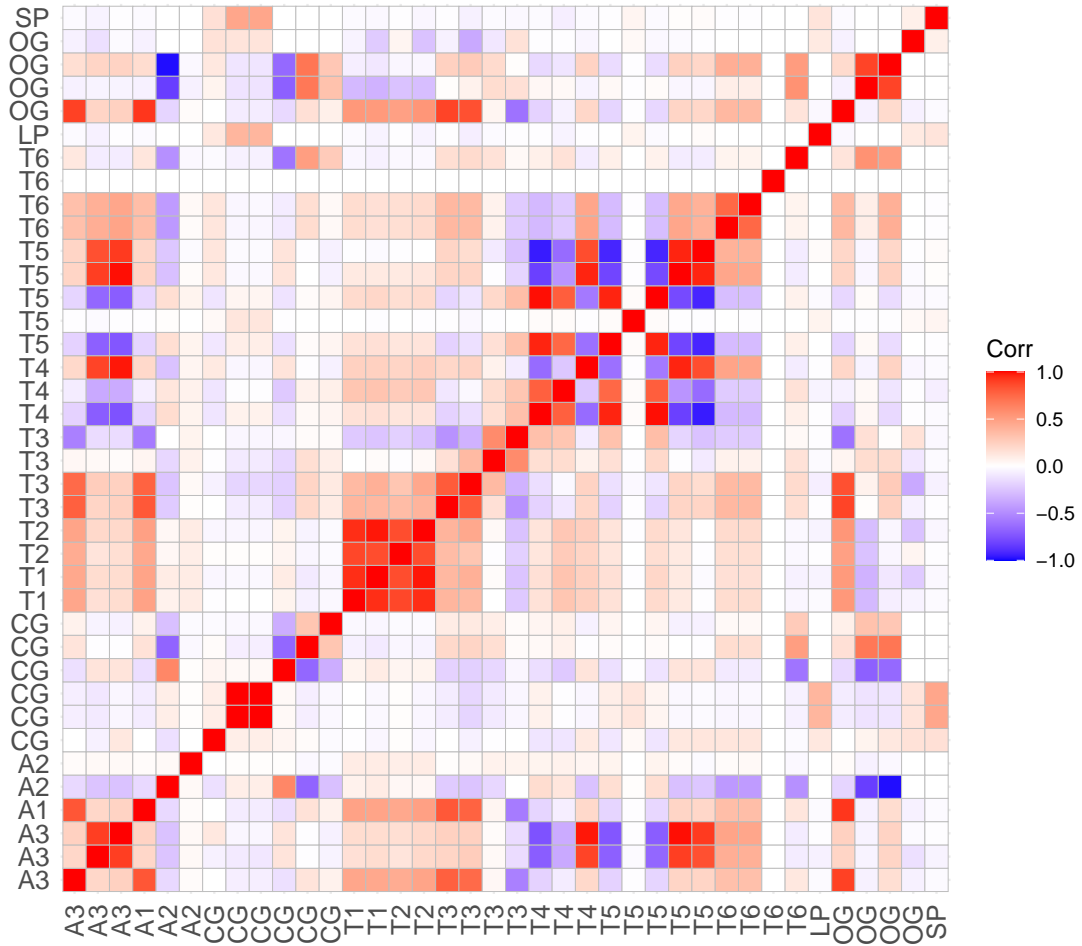


Figure 8: Correlation plot of 38 PVs (aggregated data, seconds). At each row and column, the subsystem of the PV is indicated, not the PV-name itself. Pearson correlation was computed for each pair of variables, not accounting for temporal structure. This means that correlations may be *spurious*. As expected from the causal graph, we see strong correlations between PVs from subsystems A3, T4, and T5, for example.

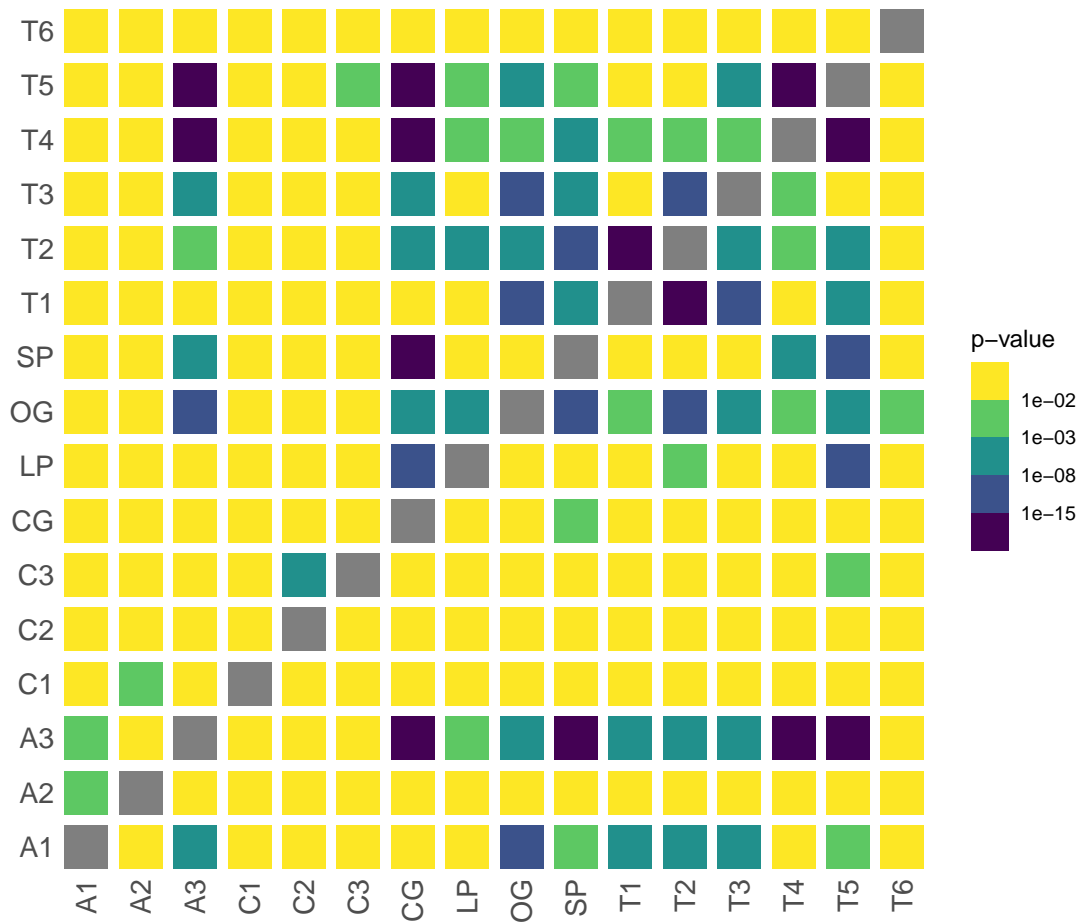


Figure 9: Tests of Granger noncausality based on one hour of data (aggregated to seconds). The color scale is discrete: p-values larger than 0.01 are yellow, p-values larger than 0.001 (but less than 0.01) are limegreen, etc. For subsystems  $i$  and  $j$ , the square in  $(i, j)$  (row index  $i$  and column index  $j$ ) contains the test result for testing if subsystem  $j$  is Granger-noncausal for subsystem  $i$ .