Space-Time Causal Discovery in Earth System Science: A Local Stencil Learning Approach

J. Jake Nichol[1,2*], Michael Weylandt[3], G. Matthew Fricke[1], Melanie E. Moses[1,4], Diana Bull[2], and Laura P. Swiler[2]

[1]Department of Computer Science, University of New Mexico
[2]Sandia National Laboratories
[3]Zicklin School of Business at Baruch College, City University of New York
[4]Santa Fe Institute

Key Points:

- We introduce Causal Space-Time Stencil Learning (CaStLe) for learning local causal dynamical structure underlying space-time data.
- CaStLe enables previously infeasible analyses of grid-cell-level Earth system data, significantly outperforming traditional methods.
- We demonstrate this new capability by recovering the space-time evolution of atmospheric aerosol flow weeks post-volcanic eruption.

*1451 Innovation Pkwy SE #600, Albuquerque, NM 87123, U.S.A.

Corresponding author: J. Jake Nichol, `jjaken@unm.edu`

Abstract

Causal discovery tools enable scientists to infer meaningful relationships from observational data, spurring advances in fields as diverse as biology, economics, and climate science. Despite these successes, the application of causal discovery to space-time systems remains immensely challenging due to the high-dimensional nature of the data. For example, in climate sciences, modern observational temperature records over the past few decades regularly measure thousands of locations around the globe. To address these challenges, we introduce **Ca**usal **S**pace-Time S**t**encil **Le**arning (CaStLe), a novel meta-algorithm for discovering causal structures in complex space-time systems. CaStLe leverages regularities in local space-time dependencies to learn governing global dynamics. This local perspective eliminates spurious confounding and drastically reduces sample complexity, making space-time causal discovery practical and effective. For causal discovery, CaStLe flexibly accepts any appropriately adapted time series causal discovery algorithm to recover local causal structures. These advances enable causal discovery of geophysical phenomena that were previously unapproachable, including non-periodic, transient phenomena such as volcanic eruption plumes. Regularities in local space-time dependencies are transformed into *informative spatial replicates*, which actually improves CaStLe's performance when applied to ever-larger spatial grids. We successfully apply CaStLe to discover the atmospheric dynamics governing the climate response to the 1991 Mount Pinatubo volcanic eruption. We provide validation experiments to demonstrate the effectiveness of CaStLe over existing causal-discovery frameworks on a range of geophysics-inspired benchmarks while identifying the method's limitations and domains where its assumptions may not hold.

Plain Language Summary

We introduce a new method for learning the dynamics of causal systems, that is, the physical rules that define a system's behavior. While this task, *causal discovery*, is not new, existing tools are ill-suited for many large geophysics datasets. Current state-of-the-art approaches use statistical techniques to search for causal relationships between all aspects of a system, examining billions of possible causal effects, or simplifying the data by focusing on the most important variables. Instead of an exhaustive search or over-simplifying the data, we incorporate basic physical principles—requiring effects to be "local" and "uniform"—to massively simplify the causal discovery problem. We demonstrate that our approach can recover known geophysical dynamics by applying it to the 1991 Mt. Pinatubo eruption, validating its ability to uncover space-time causal structure from observational data.

1 Introduction

Explaining the causal dynamics that govern geophysical phenomena is paramount in the Earth sciences. Climate models, for example, critically depend on understanding both local and global causal pathways to model the complex Earth system. Understanding short- and long-term consequences of the Earth system's behavior is essential for future model development, our scientific knowledge, and preparing for the future. More specifically, in atmospheric science, we know the initial state of specific wind modes, such as the quasi-biennial oscillation or the Brewer-Dobson circulation, dramatically affects the later evolution and impact of volcanic eruptions, major wildfires, or geoengineering efforts such as stratospheric aerosol injection (Hitchman et al., 1994; Jones et al., 1998; Aquila et al., 2014; Gray et al., 2018).

Traditional statistical methodologies, while providing valuable insights, often fall short of capturing the complex causal relationships inherent in geophysical systems. Causal models are hard-won and often represent the culmination of many decades of research. Causal discovery tools aim to accelerate the discovery of these relationships using philosophicallyand statistically-rigorous techniques to separate predictable, but indirect, statistical relationships from direct causal connections. Causal discovery has been successful across the sciences, providing new understandings of climate, biological, genetic, neural, and other dynamical systems (Ebert-Uphoff & Deng, 2012; Sugihara et al., 2012; Neto et al., 2010; X. Zhang et al., 2011; Kamiński et al., 2001; Tsonis et al., 2017). However, applying existing causal methods to space and time structured data remains limited due to the complexity and scale of such systems.

This work presents a novel causal discovery methodology that overcomes these challenges to recover networks describing local causal structures from gridded data. A fundamental insight driving the present work is that in many complex systems, global phenomena—whether climate teleconnections, brain functional networks, or ecosystem dynamics—emerge from countless repeated and structured local interactions. We can better understand how complex global patterns arise by accurately capturing these foundational local structures.

Today's Earth science measurement and modeling capabilities provide a wealth of data for studying our planet's complex dynamics. However, due to the immense complexity of these dynamics, simple analyses provide only a limited understanding of the data. Causal discovery tools offer the ability to understand finer mechanistic details via causal graphs' simplicity, interpretability, and flexibility. Causal discovery is a field that utilizes algorithmic causal inference to identify causal models as dependencies between fields of interest, which are often represented as a directed acyclic graph (DAG). Causal graphs let us analyze the space-time evolution of fields of interest and causal discovery can estimate them without requiring hypothesized physical models. Insights gleaned from causal discovery can further inform physical models, validate simulations against observational data, and identify future research questions.

While causal discovery show considerable promise for addressing problems in the Earth sciences, the enormous size and scope of Earth science data have limited its applications. For example, atmospheric data often contains hundreds of thousands of grid cells, each with several orders of magnitude fewer observations in time. That imbalance is one aspect of the *curse of dimensionality* (Bellman, 1957; Bühlmann & Geer, 2011), where high dimensionality relative to sample size challenges conventional statistical methods and renders many forms of inference, including causal discovery, unreliable without dimensionality reduction. Despite these obstacles, causal discovery has been successfully applied in Earth science (Deng & Ebert-Uphoff, 2014; Runge et al., 2015; Capua et al., 2019, 2020; Nowack et al., 2020; Krich et al., 2020; Galytska et al., 2022; Tibau et al., 2022; O'Kane et al., 2024; Zhao et al., 2024), primarily via dimensionality reduction techniques to reduce the number of relationships to estimate. Those contributions identified teleconnection pathways to recover large, periodic climate modes and their effects. While a dimensionality reduction approaches can be practical, the analysis of local effects has been considered challenging and generally avoided due to the curse of dimensionality (Ebert-Uphoff & Deng, 2012; Runge et al., 2015; Nowack et al., 2020).

In contrast to dimensionality reduction methods that marginalize large amounts of information, our work leverages the known locality in space-time systems to harness *informative spatial replicates*, i.e., repeating space-time relationships, without loss of local structural information, to identify local causal graphs. These advances enables us to approach problem classes in space-time systems that are typically intractable with prior art—both in terms of performance and algorithmic efficiency. We highlight two features of CaStLe that are useful contributions to causal discovery for geoscience problems: the ability to learn grid-level relationships instead of regional relationships from reduced dimensional data (e.g. principal components or modes) and the ability to handle dynamic, advective processes.

Prior causal discovery work in Earth science has primarily focused on large-scale regional phenomena, such as the El Niño Southern Oscillation. These patterns, generally consistent in their spatial distribution and periodic in nature, are well suited to global dimensionality reduction techniques, which project fields onto a small number of modes. While global teleconnections are crucial research areas, they ultimately emerge from local causal interactions. However, dimensionality reduction sacrifices critical local information, making it impossible to see how local structures give rise to global patterns. CaStLe reduces problem complexity in a fundamentally different way: By identifying and leveraging the repeating local structures, it preserves the relationships at the grid level while remaining applicable to spacetime systems that exhibit multiscale organization.

Typical dimensionality reduction approaches to causal discovery decrease the data space from many grid cells to a few regional modes and uses many observations, resulting in a *little p, large n* problem, where p is the number of variables and n is the number of data points. In contrast, phenomena that evolve dynamically in space or occur rarely, like volcanic plumes, are harder to analyze and often have few data points. Such problems are *large p, little n*. CaStLe makes causal discovery of the space-time evolution of these phenomena tractable for the first time by leveraging the gridded sample space, avoiding the marginalization that reduces many grid cells into a single time series per regional mode, and recovering interpretable space-time causal structures.

This work's primary case study is the 1991 Mount Pinatubo eruption. It injected a plume of aerosols into the stratosphere, which then advected around the tropical zone before dispersing northward and eventually diffusing around the globe. This example demonstrates the characteristics of the unique, transient problem class, has an established research history, and exhibits dynamics verifiable with a known causal driver: stratospheric wind.

We introduce a new Earth system causal network, the *causal stencil graph*, which describes local space-time causal structures between adjacent locations, and a new estimation methodology, **Ca**usal **S**pace-Time S**t**encil **Le**arning (CaStLe), that is capable of describing local mechanistic pathways in space and time between grid cells. Grid-level causal discovery in high dimensional space-time data has been previously considered intractable due to the curse of dimensionality (Nowack et al., 2020; Tibau et al., 2022). Though demonstrated with climate model output, our methodology applies to any space-time system where local physical interactions drive global behavior, including fluid dynamics, biological pattern formation, or material transport processes.

CaStLe combines modern causal discovery with classical physics-based principles, namely spatial and temporal locality, to accurately perform causal discovery on large spatial domains. Our novel local-coordinate-space projection does not marginalize any data points, such that local causal information is lost, which is a common sacrifice of other space-time dimension reduction techniques such as weighted averaging or principal component analysis (PCA). This preservation of local information is crucial because global-scale phenomena in complex systems emerge from interactions at smaller scales. By mapping these foundational causal pathways, CaStLe provides insights not just into immediate local effects but also into how these effects propagate and combine to create larger-scale patterns.

With these advances, CaStLe achieves remarkable improvements over state-of-the-art space-time causal discovery approaches. CaStLe is a flexible framework that can be implemented by adapting any given time series causal discovery algorithm to the stencil approach. Our approach performs excellently in high-dimensional data regimes, making it capable of describing the local space-time evolution of transient phenomena transporting over many grid cells.

The Earth system is rich with transient phenomena examples including forest fires, monsoons, coastal erosion, salt or freshwater incursions, inter-tropical convergence zone shifts, and atmospheric rivers. Aside from elucidating underlying dynamics, CaStLe can be used to identify and characterize causal change points, such as polar vortex disruption and ocean current disruptions. Additionally, understanding these local dynamic structures can give further insights into the construction and evolution of important macro phenomena such as the El Niño Southern Oscillation, the Quasi-Biennial Oscillation, and the Madden-Julian Oscillation. Table 1 in the Appendix summarizes the capabilities of CaStLe and their relevance to specific Earth science applications. These capabilities address analytical needs that have been challenging or infeasible with previous causal discovery approaches.

The remainder of this paper is organized as follows: Section 2 provides a brief background on causal discovery and its use in Earth science; Section 3 describes our case studies in the HSW-V and E3SMv2-SPA models and available data; Section 4 explains our novel CaStLe methodology; Section 5 demonstrates CaStLe's ability to recover known volcanic aerosol evolution in climate models of different resolution; and finally, Section 6 illustrates CaStLe's computational, and performance improvements over the state-of-the-art methods with synthetic data experiments.

Contributions

We introduce the CaStLe approach to causal discovery from space-time data. CaStLe allows the discovery of causal structures in high-dimensional spatial data, avoiding the need for dimension reduction techniques that dominate causal discovery of space-time data, e.g., the work by Nowack et al. (2020). By working in the raw data space, CaStLe's causal graphs are *inherently interpretable* and do not require mapping structures from the dimension-reduced space back onto the original data. We provide a theoretical analysis of CaStLe, showing that it has attractive computational and statistical properties and, rather remarkably, that CaStLe's accuracy actually increases on larger spatial domains. We apply CaStLe to two simulations of a major volcanic eruption and demonstrate how it can be used to better understand how stratospheric winds mediate the climate response to volcanic activity. Our first study is of a relatively simplified model to validate the methodology with proxy ground-truth. In our second study, we consider a more realistic model and find that CaStLe still provides consistent and valuable results, demonstrating its value for realistic atmospheric dynamics. Finally, extensive numerical experiments measure the advantages of CaStLe and demonstrate: i) significantly improved performance over existing causal discovery methods on a set of vector autoregressive (VAR) benchmarks; and ii) the use of CaStLe to identify the governing dynamics of Burgers' non-linear partial differential equation (PDE). While our case studies utilize climate model data, the methodology is domain-agnostic and can be applied to any high-dimensional space-time system meeting our locality and stationarity assumptions.

2 Background: Causal Discovery and Formal Mathematical Scope

Here, we provide a brief overview of the causal discovery field and the mathematical scope of our contributions. For a broader overview of causal discovery and its applications to Earth science, see the reviews by Glymour et al. (2019), Runge, Bathiany, et al. (2019), and Runge et al. (2023), and the book by Peters et al. (2017). Additionally, we outline the mathematical constraints and assumptions that define where our methodology can be applied in the class of space-time systems.

Causal discovery is a field of causal inference that seeks to recover causal dynamics from observational data. In the parlance of causal inference, *observational data* is data that is passively observed rather than data to which treatments (e.g. manipulations) have been applied. Observational data can be natural (e.g. physical observations) or synthetic (e.g. simulations). The present work exclusively pertains to untreated data, so we will use *observational* in this way.

While correlation does not imply causation, causal discovery is built upon Reichenbach's common cause principle (Reichenbach, 1956): if two quantities are correlated then one must cause the other or there is a third causal driver of the two. Causal discovery generally has two output classes: a causal graph/network (Pearl, 1995) or a structural causal model (Pearl, 1998). We focus on causal graphs, which are networks of variables (nodes) connected by edges that denote a causal dependence. Causal graphs can be more appealing than structural equation models because they are human-interpretable and do not require prior knowledge of the underlying causal function. In the study of Earth science, causal graphs may often be preferred to visually describe space-time relationships on the globe. Our contribution produces a novel type of causal graph, the causal space-time stencil, which is detailed in Section 4 and an example of which is in panel 4 of Figure 2.

2.1 Related Work: Causal Structure Learning

In recent decades, causal inference has been developed into a rigorous mathematical framework (Rubin, 1974; Pearl, 2000; Pearl et al., 2016). These developments made algorithmic discovery of causal structures from observational data possible (Spirtes et al., 1993; Peters et al., 2017; Glymour et al., 2019). Causal structures can be modeled with two common forms: structural causal models (SCMs) and causal graphs. Both describe a functional relationship between a variable X_j and its causal parents, denoted $\mathscr{P}(j)$.

For example, if X_i causes X_j, then it is said X_i is a parent of X_j and $i \in \mathscr{P}(j)$. Formally, Peters et al. (2017, p.83) defines an SCM as follows:

> *A **structural causal model (SCM)** consists of a collection of d **(structural) assignments***
>
> $$X_j := f_j(\boldsymbol{X}_{\mathscr{P}(j)}, \eta_j), \qquad j = 1, \ldots, d,$$
>
> *where $\mathscr{P}(j) \subseteq \{1, \ldots, d\} \backslash \{j\}$ are called **parents of** X_j: and a joint distribution $\mathbf{P}_\eta = P_{\eta_1,\ldots,\eta_d}$ over the noise variables, which we require to be jointly independent; that is $\mathbf{P}_\eta$ is a product distribution* [in our notation].

An SCM admits a unique causal graph, where $X_j \to X_i$ if $j \in \mathscr{P}(i)$ and $j \not\to X_i$ if $j \notin \mathscr{P}(i)$. While discovery of an SCM requires hypothesizing all f_j's, discovering a causal graph can be done without knowing the exact functions. Because a causal graph does not imply a specific function between variables, each may imply multiple SCMs. This does limit some of the inferential power of causal graphs, in exchange for more versatility.

Algorithms for discovering causal graphs have two primary classes: constraint-based and score-based algorithms. Constraint-based methods use statistical tests to compute conditional independence relationships between sets of variables. Once a set of independence relationships is established, it utilizes causal assumptions and reasoning to connect the variables with directed links. Score-based approaches are similar but use score optimization to determine causal dependence between variables. Both constraint-based and score-based algorithms produce causal graphs because they operate on graphical structures and independence relations rather than the explicit parametric relationships between variables required to specify a complete SCM.

Early causal discovery algorithms developed as two parallel traditions. The temporal Granger causality (Granger, 1969) methodology was an early innovation using time series data to determine if the past history of X aids the prediction of Y better than Y's history alone. If so, then X *Granger causes* Y. Independently, the constraint-based PC algorithm (named for its authors Peter and Clark) (Glymour & Scheines, 1986) and FCI (Spirtes & Glymour, 1991) developed out of the inductive causation (Pearl & Verma, 1992) framework and the earlier SGS algorithm (Spirtes & Glymour, 1991), significantly improving the efficiency of causal discovery using statistical structures in observed data. In time, other structural algorithms developed, such as LiNGAM (Shimizu et al., 2006), utilizing asymmetries in non-linear and non-Gaussian data for inferences, and NOTEARS (Zheng et al., 2018), a graph score-optimization-based method. Eventually, these two traditions converged as structural methods were developed to take advantage of temporally ordered data. Key advances included: hMRF (Liu et al., 2010), which uses hidden Markov models for estimation and is grounded in Granger causal structures, PCMCI (Runge, Nowack, et al., 2019) (and related PCMCI+ and LPCMCI), which improves PC to handle autocorrelated dependencies better, and DYNOTEARS (Pamfil et al., 2020), which extends the NOTEARS method to time series. More recently, a third tradition, causal representation learning, developed out of machine learning (ML) to leverage causal reasoning in ML models (Schölkopf et al., 2021). While still a developing field, it shows particular promise for estimating relationships in the presence of latent confounding.

The directed nature of time provides a powerful asymmetry to leverage, often sufficient to overcome the difficulties of autocorrelation, automatically orienting discovered relationships in time. In contrast, spatial data lacks an obvious uniform directional structure and poses challenges for causal discovery. As discussed in Section 1, while some approaches have incorporated domain-specific spatial constraints for point-measurement networks, none have developed a generalizable framework that leverages fundamental physical principles of locality to enable scalable causal discovery in high-dimensional gridded space-time systems.

2.1.1 Causal Discovery in Earth Science

We present a brief review of causal discovery for Earth science to position CaStLe within the literature. Please also see the extensive reviews by Runge et al. (2023) and Ali et al. (2024).

Ebert-Uphoff and Deng (2012) were the first to apply a causal discovery algorithm, PC-stable (Colombo & Maathuis, 2014), to the climate science domain. They were able to find a grid-cell-level causal teleconnection network in 50 year daily geopotential height data using the PC algorithm. Ebert-Uphoff and Deng (2014); Deng and Ebert-Uphoff (2014) further explored application requirements and climatological interpretations of the geopotential height analysis. In each paper, they note grid challenges related to the high expense of many grid cells, aggregation effects, and cell spacing. The first paper limits the number of grid cells to 800, while the subsequent analyses limited grid cells to 200 to minimize computational costs. While their results are compelling, they use extensive decadal data and recover patterns common to all 50 years. The fundamental difference between our work and Ebert-Uphoff and Deng's work is that they recover causal graphs from recurring atmospheric phenomena with sufficiently large datasets on relatively coarse-grained grids, whereas CaStLe recovers networks of isolated phenomena with many more grid cells and many fewer time samples per cell.

Runge et al. (2015) introduced an alternative approach to causal discovery of space-time Earth science data. They reduced the dimensionality with varimax-rotated principal component analysis prior to applying the causal discovery algorithm, producing a graph relating discrete, potentially remote, regions. Their causal graph is most similar to a teleconnection network between large areas on the globe. Nowack et al. (2020) utilized that framework to evaluate CMIP5 models. Particularly of note, they point out the challenges and strengths of Ebert-Uphoff and Deng (2012)'s grid-cell-level approach, "...

while an analysis at the grid-cell-level is more granular which, however, carries the challenges of higher dimensionality, will have a strong redundancy among neighbouring grid cells, and grid-level metrics will require handling varying spatial resolution among data sets."

Tibau et al. (2022) built on the dimensionality reduction approach, augmenting it to output grid-cell-level networks. They specifically delineate *mode-level* (dimensionality reduction or cell aggregation) and grid-level causal discovery. Their augmentation is called Mapped-PCMCI, which first applies dimensionality reduction, then computes a mode-level causal network with PCMCI, and finally maps the grid cells within the modes to each other using the network previously constructed. Their resulting network is one consisting of edges between grid cells, but the method assumes that cells within modes are fully connected, i.e., each 6 cell is dependent on all of its neighbors. In contrast, our work specifically seeks inter-cell spatial relationships. Finally, they also describe the failure of a traditional causal discovery approach for grid-cell-level data, "[if] we apply PCMCI directly at the grid-level, the low power of this high-dimensional and redundant estimation problem (see Section 2.2.2) leads to most links being missing."

Boussard et al. (2023) and Brouillard et al. (2024) developed the Causal Discovery with Single-parent Decoding (CDSD) algorithm within the causal representation learning framework and applied it to the climate science field. Like CaStLe, CDSD performs well in high-dimensional data settings but through a different mechanism. It performs dimensionality reduction by learning latent variables and enforcing a "single-parent" constraint where each grid cell belongs to exactly one latent factor. This naturally clusters grid cells into coherent, often contiguous regions and enables the discovery of causal relationships between these larger-scale patterns. In contrast to CaStLe's grid-level structure learning, CDSD identifies broader teleconnection pathways between regional climate modes. Thus, while CaStLe preserves the original grid structure to capture fine-grained causal dynamics, CDSD abstracts to a higher level by mapping the native grid space to an identifiable latent representation before performing causal discovery.

Several studies have addressed local-scale phenomena. Pfleiderer et al. (2020) applied causal discovery to identify precursors to seasonal hurricane frequency. They utilized the precursors to inform a predictive model. Polkova et al. (2021) identified local drivers of marine cold-air outbreaks in the Barents Sea. These demonstrate that existing causal discovery approaches can be valuable for seasonal and sub-seasonal phenomena. However, both marginalized large regions prior to analysis, reducing the space's dimensionality, and did not evaluate the space-time evolution of phenomena nor grid-level dynamics.

There are some examples of causal discovery algorithms leveraging spatial information. Zhu et al. (2016) developed pg-Causality that applies space-time pattern mining and a Gaussian Bayesian Network to seek local dependencies in the space-time propagation of air quality data. Sheth et al. (2022) developed STCD for understanding hydrological systems. They constrained the discovery of spatial structures by only allowing higher elevation nodes to be parents of lower elevation nodes because water follows the gravity gradient. While both cleverly use mined or known spatial structure to inform their causal discovery, they are both limited to use in sparse point-measured data from static base stations rather than gridded data. Further, these methods enforce constraints as filtering mechanisms, whereas CaStLe actively leverages spatial structure to enhance statistical power. Neither address the scalability challenges in high-dimensional gridded data.

2.1.2 Parallel Approaches in Neuroscience: Causal Discovery for High-Dimensional Spatial-Temporal Data

Other scientific domains face similar challenges with high-dimensional space-time data. Neuroscience, for example, needs to study mechanisms in brain interactions, and fMRI images may contain thousands to millions of pixels. The anatomy of the brain also exhibits locality constraints. Ramsey (2014) made computational optimizations to the Greedy Equivalence Search algorithm, including sparsity constraints and limiting the distance of potential parents, to recover graphs with millions of nodes. Saetia et al. (2021) marginalized regions of interest in the brain using spatial averaging and then applied the PCMCI algorithm to construct causal graphs. There is a common interest in recovering graphs of high-dimensional grid-level data throughout the sciences. Developing more tools that enhance the estimation and interpretability of causal graphs in these spaces will help advance our understanding of space-time structures across the sciences.

What is clear from prior work is that grid-level analyses are challenging, both statistically and computationally, due to how many grid cell dependencies need to be estimated, the enormous number of observations needed, and the redundant information content of nearby cells. As we present in the following sections, CaStLe adds to the literature as it overcomes the statistical and computational limitations of grid-level analysis by leveraging the known physical structure of spatial information to produce interpretable graphs describing local causal structures.

2.2 PDE-Like Systems

We seek to perform causal discovery from space-time data governed by consistent physical laws. As detailed in Section 4, CaStLe operates via two phases. The first restructures the given space-time data into a lower-dimensional local neighborhood space without marginalization or loss of any data points; the second is the causal discovery step. This section details the assumptions required for efficient use of spatial replicates that enable CaStLe's first phase, scalability properties, performance in high-dimensional settings, and interpretability. We note that the assumptions necessary for the second phase will be inherited from our meta-algorithm's chosen causal discovery method. In general, they will be the causal Markov condition, faithfulness, and often causal sufficiency, which we define formally in Appendix A.2.

We take PDE-like models as our starting point, and assume that all behavior in the given space are driven by a fixed set of dynamics that apply at infinitesimal time and spatial scales. Specifically, we assume that, for data observed in discrete space and time, the evolution of a single grid cell is controlled only by the values of its immediate spatial neighbors at the previous time step. Using causal discovery, we seek to determine which neighbors have a causal impact on a given grid cell and the direction of that relationship. Our analytical framework has similarities to the sparse identification framework initially developed by Brunton et al. (2016), though our approach builds upon causal discovery rather than sparse regression. Because our approach can use non-linear conditional independence tests, we can avoid the difficult dictionary construction step associated with sparse regression methods.

In contrast to causal discovery methods, other current research also focuses on approximating ordinary differential equations or PDE-like systems with operator learning approaches, such as operator neural networks (Li et al., 2020; Pathak et al., 2022; Hart et al., 2023). These Fourier Neural Operators (FNO) focus on generating accurate models of the PDE-like evolution of key variables over time and space. Their assumptions are rooted in several of the same fundamental physical principles of how PDEs propagate effects in space and time as CaStLe: locality in space and time and spatial stationarity. While CaStLe is not meant to be a predictive model, it captures important relationships between grid cells in an interpretable fashion, providing insights into the underlying causal structures.

2.3 Causal Discovery of Physical Dynamics: Dynamical Constraints

We state here four key assumptions that capture what we describe as a PDE-like system $\boldsymbol{X}_t$:

T1) Temporal Locality: for any $\tau \neq 1$, $X_{i,t-\tau} \not\to X_{j,t}$ for any spatial coordinates (i, j)

T2) Temporal Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over time. That is, $X_{i,t-1} \to X_{j,t} \Leftrightarrow X_{i,t-1+\tau} \to X_{j,t+\tau}$ for any time offset τ.

S1) Spatial Locality: if (i, j) are not neighbors (in a problem-specific sense) then $X_{i,t_1} \not\to X_{j,t_2}$ for any t_1, t_2.

S2) Spatial Causal Stationarity: the dynamics governing the evolution of $\boldsymbol{X}_t$ do not change over space. That is, $X_{i,t-1} \to X_{j,t} \Leftrightarrow X_{i+s,t-1} \to X_{j+s,t}$ for any spatial offset s.

Here, $\not\to$ denotes the absence of a direct causal relationship between two variables.

Therefore, if an SCM exists for a given system, then it will have a functional shape constrained by our assumptions: $X_t = f(X_{t-1}, \eta_t)$, for some vector of noise, η_t. In the context of an SCM, the constraints are: temporal locality (T1) adds lagged relationships between parent and child variables; spatial locality (S1) restricts possible parents to those in the spatial neighborhood of each variable (grid cell), that is, f_i is only a function of the neighborhood of i (f_i depends only on $\boldsymbol{X}_{\mathscr{P}(i)}$); and temporal/spatial causal stationarity (T2 & S2) require that there be only one function, f, for all space and time in the window/region of analysis.

Building on physical principles, Assumption T1 implies that causal dependencies follow the "arrow of time" while S1 disallows "action at a distance." Assumptions T2 and S2 serve to ensure that there is a consistent causal structure to target. Assumption S1 further requires that f_i is only a function of the neighborhood of i (f_i depends only on $\boldsymbol{X}_{\mathscr{P}(i)}$). We refer the reader to the book by Peters et al. (2017) for a more detailed discussion of how SCMs can be used to model physical systems.

We deliberately chose lag-1 temporal relationships in assumption T1 because they reflect fundamental physical principles: In the discretized form of PDEs, each element depends on the future state of the immediate past of its neighboring elements. The symmetry of the radius-1 neighborhood in assumption S1 and the single lag constraint in assumption T1 captures the essential causal dynamics in physical processes when temporal and spatial data resolutions are appropriately balanced.

While not descriptive of all possible systems, we assert these locality and stationarity assumptions are descriptive of any system governed or modeled after PDEs, cellular automata (Bhattacharjee et al., 2020), or Tobler's First Law of Geography (Miller, 2004; Walker, 2022). These assumptions reflect fundamental principles of locality and consistency that apply across numerous domains, from fluid dynamics to reaction-diffusion systems. However, for these to hold in practicality, one must also assume sufficient data is available to characterize locality and dynamics are smooth and non-turbulent, relative to the analysis frame. These assumptions imply that there is an optimal balance between temporal and spatial resolution sufficient to impose space-time locality. The exact value of this scaling is problem-dependent, as more rapidly evolving systems require higher temporal resolution, and we do not explore it further here. However, we note that similar concerns are well-studied in the design of numerical differential equation solvers where spatial and temporal discretizations must be chosen suitably consistently.

Section 4 and Appendix A detail how these assumptions are essential for our methodology, CaStLe, and discuss their limitations. Section 4.6 discusses strategies for managing those limitations. While CaStLe's framework assumptions (T1, S1, T2, S2) enable efficient use of space-time samples, the algorithm adapted for CaStLe's parent-identification phase will have additional causal assumptions.

Interestingly, CaStLe's spatial locality assumption (S1) creates an environment where, when properly implemented, causal sufficiency can be satisfied by construction. When we focus on learning only the parents of the center cell while including all potential spatial neighbors in the analysis, we automatically satisfy causal sufficiency for that specific node if S1 holds. While reliant on S1 holding, this is significant because causal discovery is notoriously the most challenging causal discovery assumption to ensure in real-world settings (Spirtes et al., 1993; Raghu et al., 2018). As we discuss in Section 4.5, sufficiency may be relaxed depending on which causal discovery algorithm is adapted for the parent-identification phase. However, satisfying it by construction may enable implementation choices with fewer compromises.

In the following sections, we discover grid-cell-level causal graphs under these five assumptions. Assumptions T1 and S1 allow us to significantly reduce the scope of the problem, as there are only 9 possible parents of a grid cell in 2D (8 neighbors and itself). Assumptions T2 and S2 suggest that we only need to determine a single local causal graph, because spatial stationarity allows us to extend it to the entire domain.

3 Data: The 1991 Mt. Pinatubo Eruption

Mount Pinatubo's eruption in 1991 was a massive, natural intervention in the climate, with effects that had a relatively high signal-to-noise ratio. The event launched 20 Tg of SO_2 gas into the atmosphere (Guo, Bluth, et al., 2004; Guo, Rose, et al., 2004; Kremser et al., 2016). The sulfate aerosols that resulted from these gases remained in the stratosphere for approximately two years, leading to stratospheric warming of ~ 1.5K and surface cooling of 0.2-0.5K (Dutton & Christy, 1992; Labitzke & McCormick, 1992; Parker, Wilson, Jones, Christy, & FOLLAND, 1996; Soden et al., 2002). This aerosol injection has recently been the object of much study, with some authors suggesting it as a natural proxy for proposed stratospheric aerosol injection (SAI) responses to global climate change (Trenberth & Dai, 2007). Recent work continues to characterize the nature of the response to the Pinatubo eruption, with the timing and spatial structure of the surface response being essential factors to inform policy decisions (Weylandt & Swiler, 2024).

Large volcanoes can impact climate quantities, such as surface temperatures, on timescales from months to years (Parker, Wilson, Jones, Christy, & Folland, 1996; Robock, 2000; Timmreck, 2012; Marshall et al., 2022). However, to evaluate whether CaStLe could recover the initial advection dynamics of volcanic aerosols, we focused on the period shortly after the eruption that includes stratospheric aerosol transport. The recent paper by Marshall et al. (2022) indicates: "Although global-scale climatic impacts following the formation of stratospheric sulfate aerosol are well understood, many aspects of the evolution of the early volcanic aerosol cloud and regional impacts are uncertain." This initial spread of aerosols in the stratosphere is a geophysical process, falling between synoptic weather patterns and longer-term impacts.

We utilized models of the event, combining stratospheric aerosol and wind data, as acase study to illustrate the analysis possible with CaStLe. Figure 1 is a high-level illustrative schematic of the this work's key ideas: We collect gridded space-time data, e.g. aerosol optical depth (AOD) measurements, and apply it to CaStLe to learn a causal

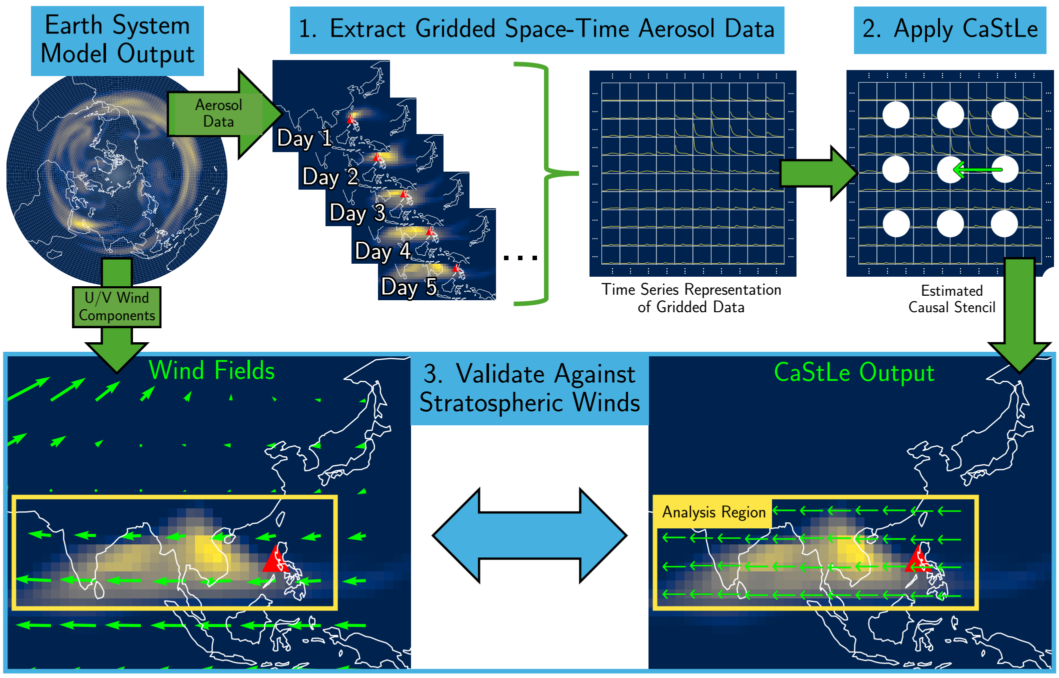


Figure 1: Schematic overview of the key elements of CaStLe and the process followed in its application to Mount Pinatubo's eruption of stratospheric aerosols. Beginning with Earth system model output, Step 1. is to collect stratospheric wind and aerosol data. Step 2. is to apply our novel CaStLe meta-algorithm to the aerosol data to obtain a causal graph describing the space-time evolution of the aerosols. Finally, we use the wind fields to help validate the causal graph results in Step 3.

stencil graph. We then map the stencil to the original grid space. Finally, we compare the data to ground-truth. To be clear, the ground-truth in our later case studies is a proxy, referring to the models' understood underlying dynamics, not the true realization of AOD in Earth's atmosphere or a mathematical representation of the dynamics. In Section 5, we compare to the wind fields carrying AOD as a proxy ground-truth. In Section 6, we compare CaStLe results from synthetic data to mathematically-known ground-truth.

3.1 Held-Suarez-Williamson-Volcanic

For our first case study, we utilized the limited-variability ensemble approach of the Held-Suarez-Williamson-Volcanic (HSW-V) model (Hollowed et al., 2024). HSW-V is an atmosphere-only model built in the Department of Energy's Energy Exascale Earth System Model version 2 (E3SMv2) (Golaz et al., 2022). HSW-V does not set out to replicate the historical Mt. Pinatubo eruption or any other, but uses the Mt. Pinatubo's eruption characteristics "to produce a plausible realization of a volcanic event, simulated with a minimal forcing set" (Hollowed et al., 2024). The model was developed specifically to facilitate basic research of attribution methodologies by providing realistic source-to-impact pathways of eruption quantities. We use this model to create a realistically complex dataset of stratospheric aerosol and wind dynamics with a clear ground-truth to demonstrate the capabilities of CaStLe and the correctness of its results.

We gathered aerosol optical depth (AOD), sulfate, and zonal (U) and meridional (V) wind fields for analysis. Only AOD is provided to CaStLe, while the sulfate, U, and V wind components are used for validating results, as detailed in Section 5. AOD is a derived quantity that measures the extinction of a beam of light through the atmosphere by atmospheric aerosols, i.e., it describes the amount of light occluded by atmospheric particles. One of the simplifying aspects of HSW-V is that all aerosol particles originate from SO_2 gas ejected by the volcano; this avoids confusing signals from other sources, such as smoke and dust, in the atmosphere.

The data collected from the HSW-V ensemble run are on a 2° grid with 6-hourly average observations. We selected AOD in grid cells between $-20°$ to 40°N and $-120°$ to 140°E, comprising 3,900 grid cells. We used the first three weeks post-eruption for our analysis.

3.2 Mt. Pinatubo in E3SMv2-SPA

For our second case study, we considered a simulation of the Mt. Pinatubo eruption in the fully coupled E3SMv2 model augmented with Stratospheric Prognostic Aerosol capability (E3SMv2-SPA) as detailed and validated by Brown et al. (2024). E3SMv2-SPA includes atmosphere, land, ocean, sea ice, land ice, and river components. AOD, U, and V wind fields are analogously collected from this dataset. However, in this model, aerosols are a natural feature, thus complicating the analysis of aerosol optical depth.

Data were collected on a daily temporal resolution for a 1° spatial grid. We selected grid cells between $-30°$ to 60°N and $-180°$ to 180°E. Analysis covered the first six months. Because this data has a coarser temporal resolution and finer spatial resolution than our study of HSW-V, we coarsened the CaStLe spatial grid to a 3° grid, resulting in 3,600 total grid cells. This helps ensure that the motion of aerosol particles between grid cells is measured within the one-day sample period.

4 Methodology: Causal Discovery with CaStLe

4.1 Notation

We first introduce notation used in the remainder of this paper. Data is observed on a spatial domain $\mathcal{D}$, which we typically take to be a finite subset of the real plane, $\mathbb{R}^2$. The causal structure generating this data can be represented by a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \mathcal{D}$. CaStLe represents local causal structure with a *stencil*, which we identify as a graph $\tilde{\mathcal{G}} = (\tilde{\mathcal{V}}, \tilde{\mathcal{E}})$ in a reduced coordinate space ($|\tilde{\mathcal{V}} = 9|$). In both the original and reduced spaces, let $\mathscr{P}(v)$ be the *potential* causal parents of v and let $\mathcal{P}(v)$ be the *actual* causal parents of v. We take $\mathcal{D}$ to be points on a regular grid of size $N \times N$, observed over T time steps, giving data $\boldsymbol{X} \in \mathbb{R}^{N^2 \times T}$. When transformed to the reduced space used by CaStLe, the resulting data matrix will be denoted $\tilde{\boldsymbol{X}} \in \mathbb{R}^{T(N-2)^2 \times 9}$. Quantities estimated from data are denoted with a hat, e.g., $\hat{\mathcal{P}}(v)$. We provide additional background on the interpretation of the causal graphs $\mathcal{G}, \tilde{\mathcal{G}}$ in Section 2.1 and formally specify the mapping between $\boldsymbol{X}$ and $\tilde{\boldsymbol{X}}$, or equivalently, between $\mathcal{V}$ and $\tilde{\mathcal{V}}$, in Section 4.3.

4.2 Causal Space-Time Stencil Learning

We now introduce the CaStLe paradigm for the causal discovery of local space-time dynamics. Under our assumptions, CaStLe identifies a *sketch* of the local causal dynamics, which we call a stencil. This stencil can then be used to construct the causal graph for the entire system (S2). The stencil is estimated in a reduced coordinate space, where we only examine the direct neighbors of a given grid cell (S1). We can pool information across time (T2) and space (S2) in order to estimate the stencil accurately, and the problem is tractable because we only seek causal parents which are local in time (T1). As we will see, this combination of reduced search space and pooled information provides a powerful approach to causal discovery and enables accurate causal discovery from high-dimensional grid-cell-level data.

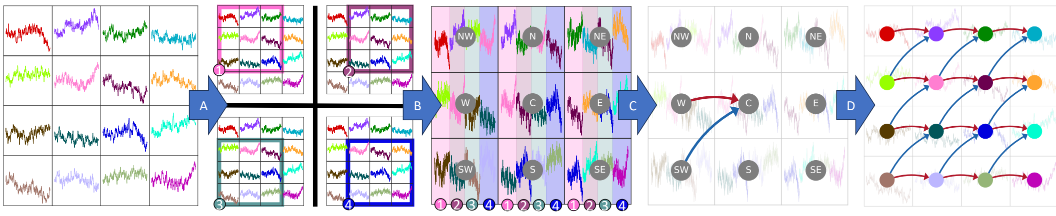


Figure 2: Illustration of CaStLe (Algorithm 1) as applied to space-time data on a 4×4 grid. *Step A* (§4.3.1): for every interior grid cell, its 3×3 (Moore) neighborhood is selected. (Note, all four 4×4 grids in the second panel are identical.) *Step B* (§4.3.1): Data are represented in a reduced coordinate space obtained by appending time series from each neighborhood according to its position relative to the neighborhood's center. *Step C* (§4.3.2): during the Parent Identification Phase (PIP), a causal discovery algorithm is used to estimate the parents of the center time series; the resulting graph forms the causal stencil. *Step D* (§4.3.3): the estimated stencil is expanded to its equivalent representation in the original space. Note that each *time chunk* (colored intervals in the center panel) in the reduced space corresponds to an interior grid cell of the original data, and that each edge in the final causal graph reflects to a stencil edge learned during the PIP. See §4.3 for details.

Having motivated the CaStLe approach to causal discovery from space-time data in Section 2.2, we now state it formally as Algorithm 1, describe its computational steps, and then analyze its statistical and computational properties.

4.3 The CaStLe Meta-Algorithm

4.3.1 Steps A-B: Projection to a Reduced Coordinate Space

CaStLe begins by transforming the given data from its original domain into a reduced coordinate space that captures the underlying causal dynamics' locality and spatial homogeneity. In this transformation, all data points are preserved, i.e., no marginalization or truncation occurs. This process is represented as Steps A and B in Figure 2 and Algorithm 1. In Step A, the local 3×3 (Moore) neighborhood of each interior cell is selected, and each cell is labeled by its location relative to the center cell (`S`, `NW`, `E`, etc.). This process creates $(N-2)^2$ sub-views in $\boldsymbol{X}_i \in \mathbb{R}^{T\times 9}$.

In Step B, these views are concatenated along the time dimension to create a reduced coordinate space data matrix $\tilde{\boldsymbol{X}} \in \mathbb{R}^{T(N-2)^2\times 9}$. Note, when concatenating the subviews, data are aligned by their coordinates relative to the neighborhood center so that, e.g., data from all `NW` cells are aligned upon concatenation, even though they originally come from different spatial locations. Although this transformation results in specific time series segments appearing in multiple reduced space cells, these repetitions do not eventually create spurious dependencies in the causal stencil, as CaStLe only seeks lag-1 dependencies. The repeated segments are well-separated in the temporal dimension, and no chunks appear in different cells in the same interval.

We depict this process on a 4×4 grid in the first half of Figure 2. In Step A, the four interior cells are sequentially highlighted, and their local neighborhoods are extracted, which are depicted in boxes colored according to the center used. In Step B, the local

Algorithm 1 CaStLe for Space-Time Data in 2D ($\mathcal{D} \subseteq \mathbb{R}^2$)

Inputs:

- Parent-Identification Phase subroutine PIP
- Gridded space-time data $\boldsymbol{X} \in \mathbb{R}^{T \times N^2}$

1. Step A: Extract 3×3 Moore Neighborhoods
   - For each interior point in the original space, construct local view of the data $\boldsymbol{X}_i = [X_{\cdot \mathscr{P}(i)}] \in \mathbb{R}^{T \times 9}$
2. Step B: Construct Reduced Space Data Matrix

$$\tilde{\boldsymbol{X}} = [\boldsymbol{X}_1^\top \, \boldsymbol{X}_2^\top \, \ldots \, \boldsymbol{X}_{(N-2)^2}^\top]^\top \in \mathbb{R}^{T(N-2)^2 \times 9}$$

3. Step C: Perform Parent-Identification in Reduced Space

$$\mathsf{PIP}(\tilde{\boldsymbol{X}}) = \tilde{\mathcal{E}} = (\hat{\mathcal{P}}(\mathtt{C}) \times \mathbb{R}^9) \subseteq \mathscr{P}(\mathtt{C}) \times \mathbb{R}^9$$

4. Step D: Expand Stencil Graph to Original Coordinate Space:
   - $\mathcal{E} = \emptyset \subseteq \mathcal{V}^2 \times \mathbb{R}$
   - For each $(p, w) \in \hat{\mathcal{E}}$:

$$\mathcal{E} = \mathcal{E} \cup \{(p(v), v, w) \text{ for } v \in \mathcal{V}\}$$

Outputs:

- Graph Stencil, $\tilde{\mathcal{E}}$
- Estimated Causal Graph, $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

data views are concatenated to form one set of time series, with each temporal *chunk* reflecting the color of the center cell of the underlying data view.

4.3.2 Step C: Parent-Identification Phase

CaStLe next examines the reduced coordinate space data representation, $\tilde{\boldsymbol{X}}$, to identify the stencil of the local causal dynamics. This is done by applying an augmentation of an arbitrary time series causal discovery algorithm to identify the parents of the center cell, `C`. We emphasize that we only seek the parents of `C`, not a full causal structure, in this step and refer to it as the *Parent Identification Phase* (PIP). Under assumption S1 (locality), all parents of `C` are present at this step, satisfying causal sufficiency, ensuring more accurate estimation of the causal stencil. By contrast, while the data of the parents for the exterior cells, e.g. `W`, is included in the reduced data space matrix, $\tilde{\boldsymbol{X}}$, it spreads across multiple columns, and accurate parent identification is not possible. The output of this process is a set of (up to) 9 weighted edges, corresponding to the parents of `C` (the eight neighboring cells and `C` itself).

We depict the PIP in Step C of Figure 2, where two parents of `C` are identified: `W`, which has a positive dependence on `C`, and `SW`, exhibiting negative dependence. Note that while the PIPs we implemented in testing—see Section 6.1—had no trouble with the *seams* connecting each time *chunk* in the reduced space, we propose an improved testing implementation in Appendix E to alleviate potential statistical testing issues.

4.3.3 Step D: Graph Reconstruction in the Original Space

Finally, CaStLe uses the stencil constructed in Step C to reconstruct the causal graph in the original data space, in a process that essentially reverses Steps A and B. Specifically, for each edge identified in $\tilde{\mathcal{E}}$, corresponding edges are added to grid cell in the original domain. We depict this in the final step of Figure 2 where the stencil is repeated throughout the entire 4×4 space, copying the two parents of `C` identified in Step C, to create a causal graph in the original space. Note also that we use the stencil to identify parents for both interior and boundary cells, omitting edges that go "off-grid" when applying the stencil to boundary cells.

4.4 Theoretical Properties

CaStLe has many advantages over classical causal discovery algorithms in gridded space-time settings. By reducing the causal discovery problem to identifying the causal parents of the center cell (`C`) in the reduced space, CaStLe achieves significant improvements in both the computation necessary to infer the causal graph and the statistical quality of that graph. As previewed in Section 2.2, the PIP's focus on identifying only the parents of the center cell creates an important connection to the causal discovery assumption of causal sufficiency. Because we include all spatial neighbors (as defined by our locality assumption S1) in the conditioning set, all potential parents of the center cell are present in the analysis. If our spatial locality assumption holds, causal sufficiency is automatically satisfied within each local stencil analysis. This represents a key advantage of the CaStLe framework - while the Markov condition and faithfulness remain necessary assumptions for the PIP algorithm, our implementation leverages spatial structure to ensure causal sufficiency by construction.

Below, we briefly outline the theoretical implications and their contributions to CaStLe's remarkable performance and algorithmic improvements. Their derivation, a deeper analysis, and a discussion on graph estimation asymptotic consistency are provided in Appendix B. We discuss CaStLe's asymptotic consistency in Appendix C, which shows that CaStLe converges on the correct causal stencil as grid size increases, given a PIP consistent in increasing time samples. These properties illustrate the mathematical justification for CaStLe's empirical correctness and improvement over the state of the art shown in the following sections.

CaStLe yields significant improvements to both *time complexity*, a measure of an algorithm's computation time as it scales with input size (e.g., number of time steps, graph nodes), and *statistical complexity*, a measure of estimation performance given larger sample sizes. Following the complexity analysis of Kalisch and Bühlmann (2007), we show that traditional causal discovery approaches are bounded by $\mathcal{O}(np^3 2^p) = \mathcal{O}(T(N^2)^3 2^{N^2}) = \mathcal{O}(TN^6 2^{N^2})$, for T time samples and $N \times N = N^2$ grid cells. Since CaStLe computes on the smaller *reduced coordinate space*, and only seeks causal parents of one node, rather than parents of all nodes, several terms become constants, resulting in $\mathcal{O}(np^3 2^p) = \mathcal{O}(T(N-2)^2 \times 9^3 \times 2^9) = \mathcal{O}(TN^2)$. CaStLe's computational complexity is $\mathcal{O}(TN^2)$, a major improvement over existing approaches. For more details on this derivation, see Appendix B.1. By leveraging locality and spatial replicates, CaStLe identifies causal structure for the entire graph ($\mathcal{O}(N^4)$ possible edges) in N^2 time. Kalisch and Bühlmann (2007, Appendix B) show that the probability of the PC algorithm incorrectly estimating the true graph is bounded by $\approx \mathcal{O}(N^{2N^2})$, whereas we find that CaStLe's error probability scales as $\approx \mathcal{O}\left(\frac{N^2 T}{e^{N^2 T}}\right)$. From this, as the grid size grows larger, we see PC is less likely to estimate the correct causal graph, while CaStLe is more likely to estimate the correct graph. Furthermore, both of these effects are exponential, implying significant performance differences even on moderately sized graphs; this change from a regime of exponential decay to super-exponential growth in graph recovery performance makes local causal graph recovery feasible, finally enabling the tools of causal discovery to scalably explore grid-level Earth science dynamics in commonly high-dimensional settings.

4.5 Methodological Limitations

CaStLe's assumptions may pose challenges in some domains of interest, and violations of these assumptions can affect the CaStLe output. For example, large-scale homogeneity can be difficult to achieve in geosciences, which is the primary rationale for the spatial-blocking strategy that we implement for our application in Section 5. Locality assumptions (T1 & S1) create a framework where the causal Markov condition can be effectively applied to local structures, while causal stationarity assumptions (T2 & S2) create consistency in these structures across space and time. However, the PIP algorithm we use within CaStLe additionally requires standard causal discovery assumptions, particularly the causal Markov condition and faithfulness, which is a separate non-trivial assumption. We list causal sufficiency as an assumption, however, if the others hold then it follows that all of the causal parents of the stencil's center are in its immediate neighborhood, so sufficiency is satisfied by construction. Alternatively, causal sufficiency may be relaxed if the chosen PIP is an algorithm that does not rely on sufficiency, such as the FCI algorithm (Glymour et al., 2019). As such, violations of CaStLe's assumptions relate directly to violations of the causal Markov condition, faithfulness, and causal sufficiency. Both Spirtes et al. (1993, p. 29) and Runge (2018) discuss assumption violations in causal discovery and some examples of how they manifest in resulting graphs. We have included a more detailed discussion on each assumption and their limitations in Appendix A.

4.6 Strategies for Addressing Limitations

To address the limitations of CaStLe's assumptions, several practical strategies can be employed. One effective approach is the use of spatial blocking to create subdivisions where dynamics are more uniform, thus mitigating the violation of spatial causal stationarity (S2). The selection and size of these blocks are highly domain-dependent and can be guided by subject matter expertise. An automated approach may be sufficient for certain dynamics, such as stratospheric dynamics, but more manual approaches may be necessary for surface-level dynamics where blocks are chosen based on topological assumptions. In specific areas of interest, blocks can be manually created to avoid topological boundaries such as coastlines, rivers, and mountain ranges, ensuring that the assumptions of spatial homogeneity are better satisfied.

Additionally, strategies such as variograms can be used to test for spatial statistical stationarity, providing heuristics for effective blocking. In future work, an iterative block size estimation approach could be considered. Varying the block size serves as a form of *stability check*, a technique widely applied in ML to ensure robustness of discoveries to parameter choices and modeling assumptions (Allen et al., 2023). However, it is important to note that there may not always be a single optimal block size due to the complex nature of spatial dynamics. Instead, there may be a range of valuable block sizes depending on the needs for analysis and the limitations of the setting. Because CaStLe is data efficient, it may be better to tend towards smaller blocks, which are more likely to be homogeneous, but possibly at the cost of some interpretability.

Deep learning and space-time feature engineering approaches may be fruitful directions for future research on automated block-identification. Methods such as δ-MAPS (Fountalis et al., 2018), feature extraction with convolutional neural networks (Nukavarapu et al., 2023), and spatiotemporal cluster analysis (Davis et al., 2025) are strong starting points. These computational approaches could automate the identification of optimal spatial blocks, reducing reliance on manual delineation and subject matter expertise while preserving the statistical properties necessary for valid causal discovery with CaStLe.

By employing these strategies and acknowledging their limitations, the robustness and applicability of CaStLe in various domains can be significantly enhanced, allowing for more accurate causal discovery in complex space-time systems. In general, more data at higher spatial and temporal resolutions will make satisfying the assumptions easier. The appeal of CaStLe is when one is interested in small-scale local dynamics, it is preferable to analyze raw gridded data directly, because marginalization can introduce statistical artifacts.

Appendix I provides an empirical investigation of how violations of each assumption affect CaStLe's performance when applied to our E3SMv2-SPA case study. Our analysis reveals that CaStLe is surprisingly robust to moderate assumption violations. While violations of spatial and temporal causal stationarity (particularly with overly large blocks or extended time intervals) introduce more false positives and reduce interpretability, CaStLe often still identifies key true causal pathways. This robustness to moderate assumption violations further expands the practical utility of CaStLe in realistic Earth science applications where perfect adherence to assumptions is rarely possible.

5 Results: Discovering Atmospheric Dynamics in Global Climate Models

As described in Section 3, we applied CaStLe to output of the Held-Suarez-Williamson-Volcanic atmosphere model, tuned to accurately reproduce the observed Pinatubo response (Hollowed et al., 2024), and the E3SMv2-SPA model including the eruption. In this section, we describe how we applied CaStLe to these case studies and present the results.

5.1 Validation with HSW-V

We first note important implementation considerations, particularly how CaStLe's assumptions are satisfied. In general, if assumptions T1, T2, S1, and S2 are uncertain, either because of data availability or dynamical instability, then assumptions can be verified using subject matter expertise. In this study of Mt. Pinatubo, we describe how we carefully managed each assumption prior to applying CaStLe.

In order to be sure CaStLe's assumptions of temporal locality, temporal causal stationarity, and spatial locality (T1, T2, and S1) held in the dataset's 2° grid resolution (corresponding to approximately 214 km at 15 degrees N), we used atmospheric wind speeds at the time of the eruption, which were recorded at 25 m/s on average at 30 hPa; cf. Figure 1 in Thomas et al. (2009). That speed translates to a theoretical maximal aerosol travel distance of 540 km over a 6-hour period, meaning aerosols should move fast enough to traverse one 2° grid cell per time step.

Spatial causal stationarity, assumption S2, is indeed violated considering the globe holistically. We resolved this challenge by using a spatial blocking strategy to create subdivisions in which dynamics were more uniform, and applied CaStLe within each separately. As noted in Section 4.6, the selection of blocks and their size is a potential challenge and is highly domain-dependent. We conducted a sensitivity analysis of block sizes, which is presented in Appendix H, and determined that dynamics were consistent in various of block sizes. We chose a middle size, $20° \times 20°$, for this analysis to balance more nuanced outputs (smaller sizes) with less risk of false positives (larger sizes). This case study was selected for its relatively simple advective dynamics to clearly validate CaStLe and demonstrate its results in an atmospheric setting. We observe that stratospheric winds vary smoothly and slowly, without hard boundaries, which enables us to use a regular grid of blocks. Other settings, such as surface level analyses, the blocking strategy will certainly require special treatment to avoid analysis across hard dynamical boundaries, such as coastlines and mountain ranges. In Appendix H, we also demonstrate that blocking alone is not sufficient for non-CaStLed approaches to succeed.

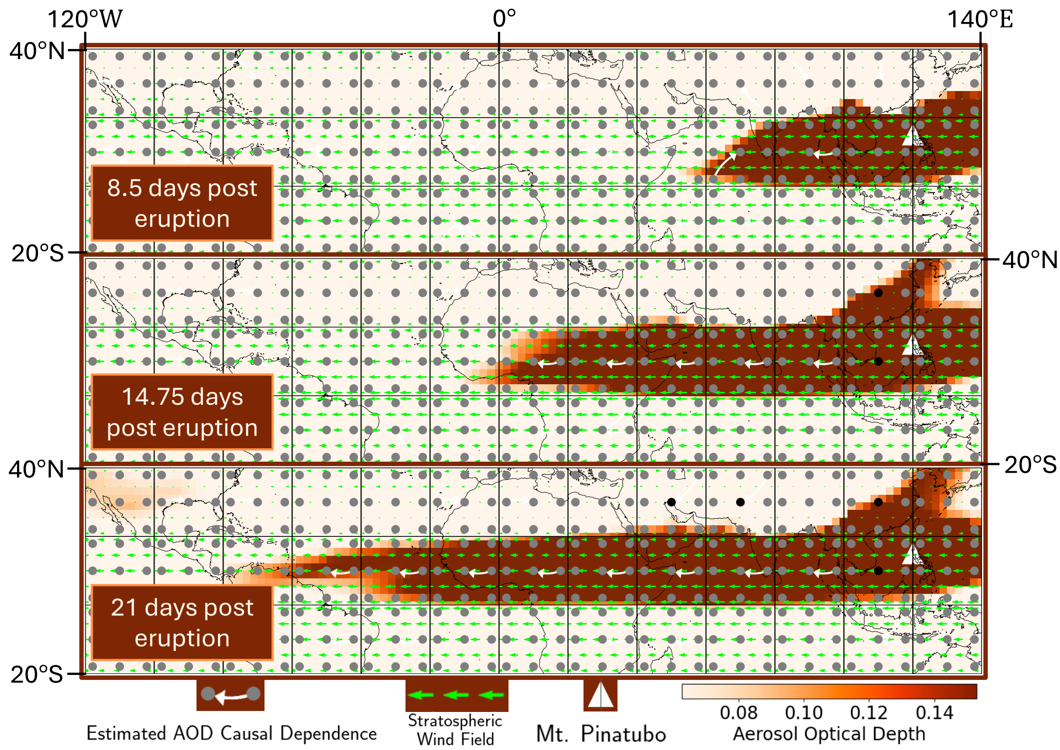


Figure 3: Application of CaStLe-PC-Stable to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a six hour time lag, the time resolution of the HSW-V dataset. On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells. Additional details are given in Section 5.

We applied CaStLe within each block separately and visualized the resulting causal stencil for each grid cell in Figure 3. In Appendix H, we provide a brief sensitivity analysis of specific block sizes and also demonstrate that blocking alone is not sufficient for non-CaStLed approaches to succeed.

We chose CaStLe's PIP to be the PC-Stable-Single algorithm because in our validation experiments in Section 6.1.2, we found it to be the marginally more effective PIP. However, those experiments showed any tested PIP algorithm is effective. PC-Stable-Single is the PC-Stable causal discovery algorithm (Colombo & Maathuis, 2014) adapted to find the causal parents of only one node; its pseudocode is provided in Appendix L. Specific CaStLe parameterizations are given in Appendix G. In Appendix J, we present similar results using DYNOTEARS for CaStLe's PIP.

Our proxy ground-truth in this case study was stratospheric winds that cause suspended aerosols to advect through space. We display dominant wind fields throughout the space to validate the resulting graphs. Our dataset included wind components in 72 pressure levels in the HSW-V dataset, so we display column-averages of the levels at the levels where volcanic sulfate was most prevalent. Specifically, we chose pressure levels containing more than 5 µg of sulphate Kg air, which were between $\sim$ 6-114 hPa. With this, we effectively captured the stratosphere and 56% of all sulfate aerosols in all atmosphere levels. By comparing winds in at the stratospheric levels where most of the sulfur was present, we can directly compare CaStLe's discovery of AOD's space-time evolution to wind data in the same locations.

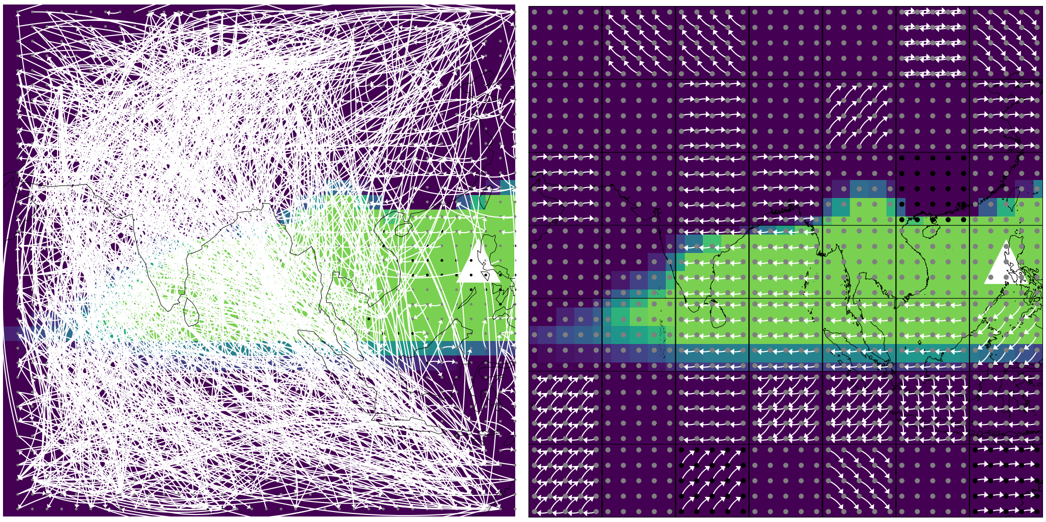
Comparing the wind and recovered stencils in Figure 3, it is clear to see that CaStLe is able to accurately reconstruct the prevailing stratospheric winds using only AOD observations. As these wind fields are the key drivers of aerosol dispersal, it is clear that CaStLe can accurately capture the dynamics dictating the spatial pattern of the Pinatubo response. The CaStLe stencils best capture the underlying wind fields when AOD levels are high. When there are few particles in a region, it is challenging to determine wind by solely observing dispersal patterns. We also observe a zonal (East-West) pattern driving the aerosol dispersion, with Pinatubo aerosols transported nearly fully around the equator within 3 weeks, while meridional (North-South) dispersion taking much longer. This alignment between CaStLe-derived causal structures and observed wind patterns demonstrates the method's effectiveness in reconstructing the physical mechanisms driving aerosol transport, particularly in regions with sufficient particle density to enable clear detection of dispersal trajectories.

5.1.1 Comparative Analysis of CaStLe Versus Traditional Approaches on HSW-V

The current state-of-the-art causal discovery methods cannot tractably approach this study of Mt. Pinatubo's aerosol short-term evolution. As described in Section 1, dimensionality reduction techniques commonly used to make them tractable are suitable for spatially static, periodic space-time patterns. However, they are not good solutions for studying a dynamic, transient pattern because modes derived from those techniques are space-timely invariant. Moreover, they are meant to capture large-scale teleconnections, rather than local dynamics that eventually give rise to global phenomena such as teleconnections. For a detailed demonstration of why dimensionality reduction approaches, such as PCA and PCA-varimax, are insufficient for capturing local causal structures in space-time systems like volcanic eruption plumes, see Appendix F.

Traditional approaches attempted without dimensionality reduction suffer from the *curse of dimensionality* when applied to short-term global-scale phenomena because there are more grid cells than temporal observations. They also struggle to identify local connections in the massive search space they seek, where every grid cell may be dependent on any other grid cell; i.e., they are not constrained by local causal structure. Finally, their efficiency scales poorly as the grid size gets larger, requiring a lot of time to execute on relatively small grids. We present specifics below and discuss time complexity in depth in Section 4.4 and Appendix B.1.

Here, we demonstrate the disparity in performance between traditional approaches and CaStLe for our HSW-V case study using the PC algorithm. The reasons for the disparity are explored in Sections 1 and 2. Because PC did not terminate within 48 hours on the full spatial region studied in Section 5.1, we restricted the analysis space the area between 20° to 50°N and 55°W to 120°E in the first 8.5 days after the eruption. On the 2° grid, the given space is equivalent to a 35×35 grid, or 1,225 grid cells. Since temporal observations were 6-hourly, there were 34 time series samples per grid cell.



(a) PC algorithm results

(b) CaStLe results

Figure 4: Causal maps inferred from the PC algorithm applied naively to all grid cells and CaStLe's equivalent results immediately to the west of Mt. Pinatubo; a 35 × 35 grid between −20° to 50°N and 55° to 125°E in a 8.5 day span after the eruption. All links represent a six hour time lag, the time resolution of the HSW-V dataset. As expected, PC struggled with the high dimensionality and the discovered dependencies do not conform to the ground-truth understanding that aerosols advected towards the west. It also fails to identify local dynamics, instead drawing most connections over great distances. The PC analysis was computed in 729 minutes on 1,600 grid cells, while the CaStLe analysis was computed in 0.46 seconds.

Figure 4 shows the results of the PC causal algorithm and CaStLe-PC-Stable applied to a large section of grid cells for the HSW-V problem. Figure 4a illustrates that PC is incapable of reconstructing a graph with any meaningful physical interpretation. There are some local dynamics found, but they are dominated by the many links across disparate locations. PC was implemented here with the partial correlation conditional independence test, a test alpha-value of 0.00001, and a p-value threshold of 0.05 to remove links below that threshold in the final graph. P-values were corrected using the Benjamini-Hochberg procedure prior to final thresholding.

In Figure 4b, CaStLe was applied to 10°-by-10° blocks, rather than the 20°-by-20° blocks in Figure 3. The smaller block size enables more link density and nuanced results, with the possibility of more mistakes. In this illustration, we chose to display the stencils mapped back to the original space for each block to compare to PC more fairly and demonstrate how much more sparse CaStLe's results are. We found that CaStLe was again able to recover the westward aerosol transport from Mt. Pinatubo. Because HSW-V only models aerosols from the volcano, there is little to no aerosol signal outside the plume, and results in these areas will be less reliable.

Additionally, the run-time of the PC algorithm is demonstrably poorer than CaStLe. The PC algorithm experiment in Figure 4a PC took 65 minutes to execute for a 35×35 grid size. In contrast, the CaStLe experiment in Figure 4b completed all blocks serially in 0.46 seconds on the same data. Further, for each of the panels in Figure 3, CaStLe computed the 39 stencils for the 3,900 grid cells in a total of 10 seconds. These empirical data points are explained by CaStLe's improved theoretical properties, as detailed in Section 4.4 and Appendix B.

5.2 Extending to More Complexity: E3SMv2-SPA Modeled Aerosols

Given the intended simplicity of the HSW-V model, we also evaluated a simulation of the Mt. Pinatubo eruption in E3SMv2-SPA. More complex graphs arise with a more complex model, providing an opportunity for more nuanced analysis and discovery, but with a higher chance of false positives and false negatives. E3SMv2-SPA is a fully coupled model, so AOD results from many sources including the volcanic eruption and Saharan dust. As such, we expect results to be somewhat noisier, however, as we demonstrate below, CaStLe is still able to identify important features of transport. Because of this additional complexity, we focus on CaStLe as an exploratory tool and leave additional analysis to future work. However, even with the added complexity, CaStLe can obtain compelling results consistent with dominant stratospheric winds as well as the dynamics discovered in our study of HSW-V.

We used 15° spatial blocks so that CaStLe operates on a 5×5 grid space per block. This size strikes a balance in the trade-off that a smaller block-grid enables more nuance in the final output, and larger block-grids take advantage of more spatial replicates to multiply sample size. We chose to study the eruption in two distinct 20-day intervals spanning a six month period to understand the changing evolution of the plume.

Similarly to HSW-V, we utilize the U and V wind fields to visually validate the CaStLe results. In this case, we did not average over multiple altitudes, instead opting to simply use the 50 hPa wind fields; this altitude was shown in Brown et al. (2024, Figure S6) to contain significant levels of the sulfate aerosols.

Figure 5 depicts the results of our experiment on E3SM. Again, we applied CaStLe-PC-Stable to construct causal stencils for each given spatial block. We selected two intervals of interest from our results to show here. Day 15 is June 15, 1991, the day of the eruption, so the top row of Figure 5 is the first 20 days after the eruption. The bottom row was selected to illustrate later dynamics when aerosols have circumnavigated the tropical zone and more northward advection is present. Days 175-195 are November 22 to December 12, 1991, a little over six months after the eruption.

In the more challenging setting of the fully-coupled E3SMv2-SPA model, our results in the first weeks are still generally consistent with those in HSW-V presented in Section 5.1, showing that CaStLe is largely robust to greater complexity. We note that visually identifying the sulfate aerosol plume is much more difficult in this case as the background AOD is quite strong. A solution may be to apply CaStLe to AOD anomalies (computed by subtracting grid cell long-term AOD means from the signal in the analysis period), thus potentially removing background variability from the analysis. However, our goal in this work is to present CaStLe as applied to raw data to illustrate what it can and cannot accomplish in complex, heterogeneous settings.

Regardless, we observe that tropical westward advection is present throughout both studied time periods, but more complexity is present in other regions, in part due to the background AOD. Six months later, the aerosols and winds are in a different regime. We observe northward and southward causal structures in the northern latitudes matching dominant wind fields in the area, with CaStLe stencils still consistent in the tropics. Additionally, CaStLe recovers dynamics moving aerosols northwards above central Asia and southwards through western North America. Causal structures are recovered more often and more accurately where stronger winds coincide with more aerosol presence, building a map of significant aerosol movement. A more complex model and smaller block sizes illustrate more nuanced dynamics, and there is more to learn from these; however, we leave deeper atmospheric dynamics analysis to future work.

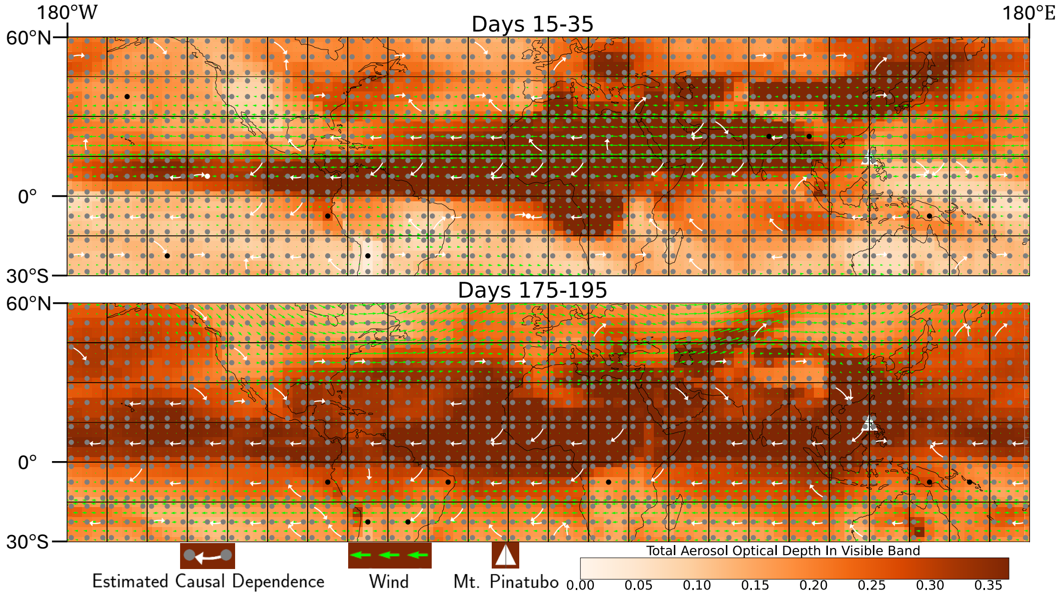


Figure 5: Application of CaStLe-PC-Stable to E3SMv2-SPA simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only total aerosol optical depth (AOD). Autodependencies are shown with black nodes where grid cells cause themselves, and gray nodes where there is no autodependence. All links represent a one day time lag, the time resolution of the E3SMv2-SPA dataset. The heatmap depicts AOD from any source at 50 hPa. The top panel depicts learning from the first 20 days after eruption, which began on day 15. The bottom panel depicts learning approx 6 months after the eruption over a 20-day time period. In the more challenging setting of the fully-coupled E3SMv2-SPA model, our results in the first weeks are still generally consistent with those in HSW-V presented in Section 5.1, showing that CaStLe is largely robust to greater complexity. In the bottom panel, the aerosols and winds are in a different regime. CaStLe stencils are still consistent in the tropics and now begin to recover dynamics pushing aerosols northwards above central Asia and southwards through western North America. A more complex model and smaller block sizes illustrate more nuanced dynamics, and there is more to learn from these, however, we leave deeper atmospheric dynamics analysis to future work.

6 Validation and Benchmarking

In this section, we demonstrate the effectiveness of the CaStLe approach to space-time causal discovery, highlighting its ability to identify structure in low-signal and data-sparse regimes. We first demonstrate the benefits the CaStLe approach can provide to *any* causal discovery algorithm using a synthetic linear-Gaussian dynamics benchmark; we then apply CaStLe to an important non-linear PDE problem, showing that we can determine the underlying advective forcing.

6.1 Evaluating CaStLe: A Comparative Analysis

We demonstrate the effectiveness of CaStLe using a set of local interaction models (LIMs), building upon the comparison framework introduced by J. J. Nichol et al. (2023). In summary, we defined a stencil for each experiment that dictates how each grid cell depends on its nine neighbors (including itself). A LIM is a special case of an SCM, which simulates the evolution of a gridded space by computing the current state of each grid cell based on a predefined function of the historical states of its neighbors. In the linear case, this is most simply accomplished with vector autoregression (VAR) models, where the coefficient is sparse, only containing nonzero entries where a desired dependence exists between neighbors. The function is defined by a linear function of coefficients in the given stencil. Our results appear in Figure 6, which shows that CaStLe provides significant improvements in graph recovery regardless of the causal discovery algorithm used in the parent identification phase.

6.1.1 Data: Benchmark Construction

In order to compare different causal discovery algorithms with a common set of benchmarks, we begin by generating coefficient matrices parameterizing spatially homogeneous and statistically stationary VAR(1)s that satisfy our key assumptions S1 and S2. We generate coefficient matrices for these VARs, $\tilde{M}$, using the following sampling scheme:

1. Generate a random 3×3 *local dynamics matrix*, M, with d non-zero elements, one of which is the central element (autocorrelation). Each of the d non-zero elements, $\{a_i\}_{i=1}^{d}$, have a random value $1.0 \geq \text{coefficient}_i \geq s_*$.
2. Expand M to $\tilde{M}$ on a grid of size $N \times N$ (cf. Step D of Algorithm 1 or Figure 2-2 of J. J. Nichol et al. (2023))
3. If $|\lambda_{\max}(\tilde{M})| \geq 1$, scale $\tilde{M}$ by $|\lambda_{\max}(\tilde{M})|$.
4. If $m < s_* \; \forall m \in \tilde{M}$, reject, else accept.

where $|\lambda_{\max}(\tilde{M})|$ is the maximum absolute eigenvalue of $\tilde{M}$, which when above 1.0 indicates the system is numerically unstable (Strang, 2016, p.307). We note that this process is essentially an accept-reject scheme used to sample from the set of statistically stationary & spatially homogeneous VARs on a 2D grid with minimum signal strengths $s_* \geq 0.1$ and fixed sparsity levels in the range $d \in \{1, 2, \ldots, 9\}$. After each $\tilde{M}$ is generated, we create a single realization, using standard Gaussian noise applied independently, cellwise at each time step.

6.1.2 Method Comparison: Highlighting CaStLe's Strengths

On each realization, we apply one of three causal discovery algorithms, in both CaStLed and non-CaStLed form: i) the `PC` algorithm of Spirtes and Glymour (1991) as adapted to time series by Runge, Nowack, et al. (2019, Algorithm S1 with $q = 1$); ii) `PCMCI`, an autocorrelated time series extension of `PC` developed by Runge, Nowack, et al. (2019); and iii) the `DYNOTEARS` approach of Pamfil et al. (2020), itself a time series adaption of the `NOTEARS` approach of Zheng et al. (2018). We additionally compare each of these against a simple sparse VAR approach, where we estimate VAR coefficients directly using ordinary least squares (OLS) and truncate coefficients with magnitude less than s_*; this approach is not necessarily causal, but it is the exact model of our data generating process and provides a useful point of comparison.

We compare the estimated graph structure with the true graph derived from the sparsity pattern of $\tilde{M}$ and report the average Matthews' Correlation Coefficient (MCC) (Matthews, 1975) and F_1 score over 30 replicates. We used an adapted MCC formula derived by J. J. Nichol et al. (2023), which accounts for edge cases in which the denominator would be zero, but is otherwise defined as:

$$\text{MCC} = \frac{(\text{TP} \times \text{TN} - \text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \tag{1}$$

where TP, FP, TN, and FN are true positive count, false positive count, true negative count, and false negative count, respectively. Here, a positive is a graph edge that exists, and a negative is a graph edge that does not exist. The MCC graph similarity measure is sometimes preferable to the more common F_β Score (β is chosen such that recall is considered β times as important as precision), which is dependent on the ratio of positive to negative test cases; we treat link positives equally to link negatives, hence our preference for MCC. Figure 6 includes the F_1 score due it its common use in causal discovery, but results are similar.

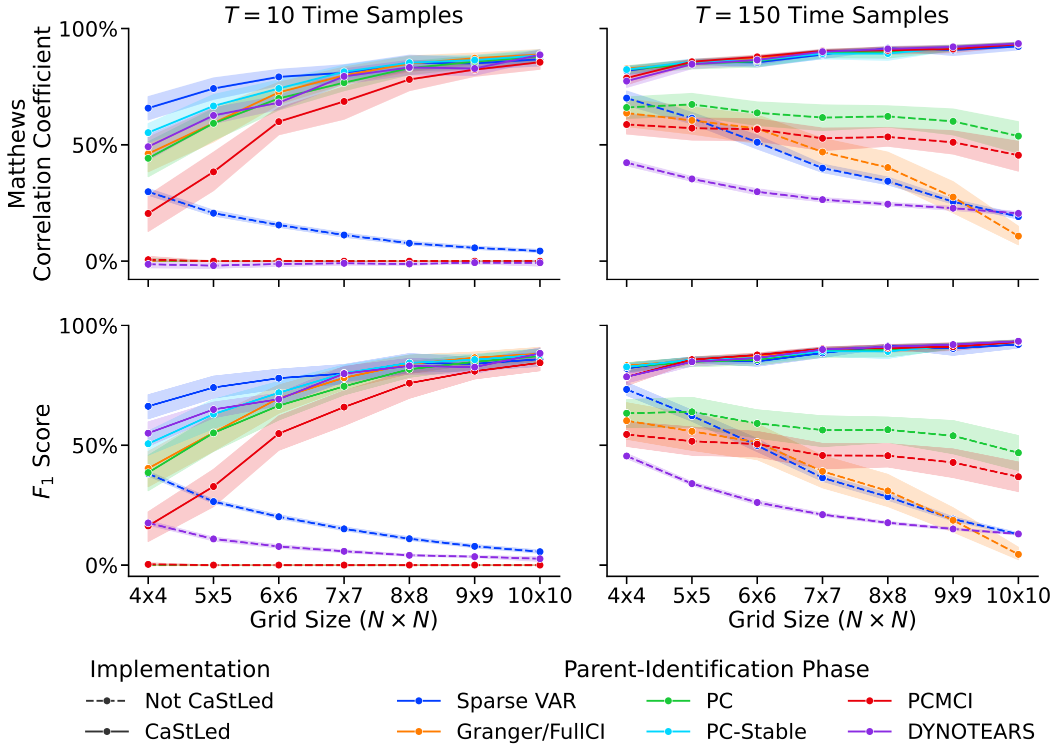


Figure 6: Comparison of CaStLed and non-CaStLed causal discovery approaches on linear-Gaussian dynamics, including Granger causality or FullCI (orange), PC (green), PCMCI (red), and DYNOTEARS (purple), as well as a statistical model of the data generating process (blue) presented with both MCC and F_1 metrics. In the low-sample size regime (T=10, left) CaStLed approaches can accurately recover the underlying causal graph, with performance increasing on larger grid sizes (solid lines); by contrast, non-CaStLed approaches are unable to perform better than mere chance (dashed lines). Even a model based on the underlying data generating process (Sparse VAR, blue) is significantly outperformed by its CaStLed counterpart. In the high-sample size regime (T=150, right), non-CaStLe approaches have improved performance but still compare unfavorably with their CaStLed counterparts.

In Figure 6, we depict CaStLe performance results on a 2D VAR with ground-truth link density $d = \frac{4}{9}$. We show two extremes of sample size: a low-sample regime of $T = 10$, which is barely enough to identify the local dynamics of 9 cells, and a high-sample regime of $T = 150$. Our results are quite striking: in the low-sample regime, the CaStLed versions of each algorithm can accurately infer graph structure, with near-perfect performance on grids of size 10×10. By contrast, the performance of the non-CaStLed versions is essentially no better than random guessing, with only the sparse VAR able to exhibit any skill, and then only on small grids. In the high-sample regime, the CaStLed variants perform well on all grid sizes, with CaStLe-PC consistently achieving perfect recovery; the non-CaStLed variants perform better, as expected, but their performance still decays quickly as the spatial grid grows.

While the stronger performance of the CaStLed variants is noteworthy, the exhibited trends are even more important and highlight the true strength of the CaStLe approach: CaStLed approaches *improve* on larger grids while traditional approaches suffer. While Figure 6 shows results for the fixed link density $d = \frac{4}{9}$, we present results for all other link densities in Appendix K.

Having established CaStLe's strong performance on linear dynamics, we also validated its effectiveness on non-linear systems that more closely resemble realistic physical processes in Earth science. Specifically, we applied CaStLe to the advection-diffusion dynamics of Burgers' equation, a fundamental non-linear PDE that models a combination of advective and diffusive processes. Unlike our VAR benchmarks, which are discrete linear models with random initializations, Burgers' equation presents continuous non-linear dynamics that allow us to evaluate CaStLe's ability to recover spatial propagation patterns under controlled conditions. Our analysis demonstrates that CaStLe successfully identifies the underlying advection angle across a range of diffusion conditions, further supporting its applicability to complex space-time systems. This non-linear validation's complete methodology and results are presented in Appendix D.

7 Discussion

We have introduced CaStLe, a novel causal discovery meta-algorithm tailored for analyzing grid-level space-time data sets arising in Earth science. CaStLe can be directly applied to grid-level data and does not require pre-processing and spatial dimension reduction, allowing it to capture dynamics in the natural domain of the data rather than a derived (PCA-type) space. This distinction is crucial because global-scale phenomena across many complex systems, whether climate teleconnections, ecological patterns, or fluid dynamics, emerge from networks of local causal interactions that are often lost in dimensionality reduction approaches. While demonstrated with Earth science case studies, CaStLe is fundamentally domain-agnostic, applicable to any space-time system governed by local physical interactions, from fluid dynamics and heat transfer to biological pattern formation.

CaStLe can overcome the limitations of existing causal discovery approaches in Earth science's space-time data, filling a significant gap. By leveraging realistic assumptions of locality and homogeneity, CaStLe creates "spatial replicates" to substitute large observational domains for lengthy time series. This process transforms the spatial causal discovery problem from the high-dimensional (many variables, few observations) to the low-dimensional (few variables, many observations) regime, allowing accurate and efficient discovery of underlying causal dynamics. A key aspect of CaStLe is the causal *stencil* graph, a simplified representation of the local dynamics driving larger global behaviors. This notion of a stencil is particularly well-suited for systems able to be modeled by PDEs, as PDE-type dynamics inherently enforce both locality and homogeneity, as well as the sufficiency assumptions necessary for causal discovery to be *truly causal.*

We used these insights to identify the space-time evolution of volcanic aerosols that erupted from Mount Pinatubo in the HSW-V and E3SMv2-SPA models. We found that CaStLe found the expected path of advection in both models and more nuanced dynamics, including northward and southward dispersion, in E3SMv2-SPA. We showed that CaStLe outperforms its peers in the causal discovery of synthetic benchmarks generated by vector autoregressive structural causal models. Additionally, as detailed in Appendix D, we found that CaStLe could accurately identify the advection angle in our Burgers' equation benchmark, demonstrating that it can filter out the "noise" of diffusion.

Our brief theoretical analysis of CaStLe in Section 4.4 and in Appendix B, demonstrates two regimes of consistent estimation for CaStLe, i.e., CaStLe recovers the true causal dynamics: long time series ($T \to \infty$) or large grid sizes ($N \to \infty$). This starkly contrasts existing approaches, whose performance rapidly deteriorates as $N \to \infty$. Several other important theoretical questions remain open, including the optimal relationship between sampling rates and grid resolution, behavior under mild violation of the key assumptions, and the correct target of inference for systems without clear advective dynamics (e.g., the chemical evolution of atmospheric aerosols).

We have focused our attention on space-time data observed on regular 2D grids, but we believe that this assumption can be relaxed to adapt CaStLe for a broader range of observational structures. CaStLe can also be adapted to multivariate space-time data (more than one observation at each point) by including more comeasured variables in CaStLe's transformation of the region to the reduced coordinate space, enabling causal discovery of the space-time interactions of multiple species on the grid-level, which is a particularly exciting avenue of future research and application to Earth system dynamics. Developing data-driven methods for evaluating block sizes based on output robustness will enable more automatic application of CaStLe, requiring less subject matter expertise. Finally, causal representation learning is a nascent field combining the estimation power of machine learning with the strength of inference of causal discovery. Applying these techniques in CaStLe's parent-identification phase or for discovering spatial embeddings for regional block analysis is an exciting potential direction for future work.

Because our assumptions are readily satisfied by many physical systems, CaStLe can be applied quite broadly in the physical sciences. It may find value in any space-time system in which quantities at every point in space impact their adjacent spatial neighbors. In the Earth system, it may be of particular interest for studying forest fires, ocean dynamics, salt/fresh water incursions, and coastal erosion, for example. For atmospheric rivers, CaStLe could identify pathways of moisture transport and evolution; for wildfire spread, it could reveal causal relationships between local weather conditions and fire behavior; for drought propagation, it could track how soil moisture deficits spread across regions. CaStLe's preservation of local causal structures while efficiently handling high-dimensional data offers advantages over approaches requiring dimension reduction. For datasets where the temporal sampling is too coarse relative to the spatial resolution, extending to a radius-2 neighborhood might be appropriate while still maintaining our core assumption of locality. This extension would preserve the fundamental CaStLe methodology—only the dimensionality of the reduced coordinate space would increase. Additionally, CaStLe provides a promising framework for Earth system model evaluation (Nowack et al., 2020; J. J. Nichol et al., 2021), potentially identifying where models produce correct outcomes through incorrect causal mechanisms.

While climate science typically studies large, long-term phenomena, the community increasingly recognizes the importance of understanding multi-scale interactions (Diffenbaugh et al., 2005; Palu, 2019; Agarwal et al., 2019; Z. Zhang et al., 2022). Teleconnections present an exciting challenge for future applications of CaStLe. These statistical dependencies between distant regions appear to violate locality but physically result from countless local interactions that are often unobserved or unmodeled. A two-stage methodology could be effective for tackling this challenge. First, apply CaStLe to discover local causal stencils, and then apply a complementary causal discovery technique to connect the discovered local processes across scales. This approach could bridge the gap between local and global causal discovery in climate science.

Complex space-time systems present apex challenges for causal discovery, combining chaotic dynamics, high dimensionality, noisy observational records, and complex underlying physical processes. CaStLe represents the first successful application of causal graph discovery to learn grid-cell-level causal structures in Earth systems. By preserving local causal structures while efficiently handling high-dimensional data, CaStLe presents a path toward connecting micro-scale interactions with macro-scale phenomena, potentially offering new insights into how global patterns emerge from local causal mechanisms.

There are rich future research directions, including multivariate analysis and automated block size selection. The feasible discovery of local causal stencils presents an exciting new frontier for causal discovery of space-time data, particularly in the Earth sciences.

Appendices

Table 1: Capabilities of CaStLe for Earth science applications. This table summarizes the key methodological advantages of CaStLe and their relevance to specific Earth science phenomena, highlighting applications where grid-level causal discovery enables analyses that were previously infeasible with prior causal discovery approaches.

Capability	**Description**	**Relevant Applications**
Local mechanism discovery	Global phenomena emerge from local causal interactions. Previous approaches use dimensionality reduction, losing this local information.	Volcanic plume transport (Sjolte et al., 2021), wildfire propagation & plume transport (Baranowski et al., 2021), atmospheric rivers (Payne et al., 2020; Baño-Medina et al., 2025; Higgins et al., 2025)
Transient, non-periodic phenomena	CaStLe effectively identifies grid-level causal pathways.	Volcanic eruptions, heat waves (Keellings & Moradkhani, 2020), wildfires (Driscoll et al., 2024)
High-dimensional data settings	CaStLe leverages spatial replicates to make high-dimensional problems tractable.	Gridded Earth science data from: regional climate modeling, satellite observation analysis, climate re-analysis products (Ali et al., 2024, Table 3)
Earth system model evaluation and comparison	CaStLe enables comparison of causal mechanisms between models and observations at the grid level, potentially identifying where models produce correct outcomes through incorrect causal mechanisms.	Grid-level causal model evaluation that identifies local mechanism differences between models and observations, extending beyond previous approaches that were limited to regional-scale analysis (Nowack et al., 2020; J. J. Nichol et al., 2021)

Appendix A Understanding Assumptions

In this section, we outline the key assumptions underpinning the CaStLe framework and their relationship to causal discovery assumptions.

A.1 CaStLe Assumptions

CaStLe operates via two complementary sets of assumptions:

1. **CaStLe Framework Assumptions (T1, S1, T2, S2)**: These enable efficient use of spatiotemporal data by leveraging locality and stationarity to transform a high-dimensional problem into a tractable one.
2. **Causal Discovery Assumptions**: The causal discovery algorithm used within CaStLe's Parent Identification Phase requires its own set of assumptions - typically the Causal Markov Condition, Faithfulness, and Causal Sufficiency.

While these assumption sets are conceptually distinct and serve different purposes, they work together to enable scalable causal discovery in high-dimensional space-time systems.

In review, our framework introduces four key assumptions to capture a "PDE-like" system $\boldsymbol{X}_t$, creating an environment where local space-time dynamics can be efficiently learned:

T1) Temporal Locality: restricts causal influence to the most recent past state, one time lag, aligning with how PDEs are discretized.
T2) Temporal Causal Stationarity: ensures consistent causal structure over time.

S1) Spatial Locality: limits causal influence to immediate spatial neighbors.
S2) Spatial Causal Stationarity: ensures consistent causal structure across space.

These assumptions enable CaStLe to leverage "spatial replicates"—treating each local neighborhood as providing information about the same underlying causal process. This transforms what would be a high-dimensional, data-sparse problem (many variables, few observations) into a data-rich problem (few variables, many observations).

A.2 Causal Discovery Assumptions

Separately, the causal discovery algorithm used within CaStLe's PIP require its own assumptions. The three foundational assumptions of causal discovery are provided below, verbatim from Runge (2018). In depth discussion of each is discussed by Spirtes et al. (1993, Ch. 3), and Peters et al. (2017, Ch. 6.5). They are discussed in terms of directed graph *separation* ($\bowtie$), where variables are separated when all causal paths between them are "blocked" by conditioning variables, preventing information flow through the graph structure. Separation is detailed more thoroughly by Runge (2018, Section III B.).

- **Causal Markov condition**:
  The joint distribution of a time series process $\boldsymbol{X}$ with graph $\mathcal{G}$ fulfills the Causal Markov Condition if and only if for all $Y_t \in \boldsymbol{X}_t$ with parents $\mathcal{P}_{Y_t}$ in the graph

$$\boldsymbol{X}_t^- \setminus \mathcal{P}_{Y_t} \bowtie Y_t \mid \mathcal{P}_{Y_t} \implies \boldsymbol{X}_t^- \setminus \mathcal{P}_{Y_t} \perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t}, \tag{2}$$

  that is, from separation in the graph (since the parents $\mathcal{P}_{Y_t}$ separate Y_t from $\boldsymbol{X}_t^- \setminus \mathcal{P}_{Y_t}$ in the graph) follows independence.
  This includes its contraposition

$$\boldsymbol{X}_t^- \setminus \mathcal{P}_{Y_t} \not\perp\!\!\!\perp Y_t \mid \mathcal{P}_{Y_t} \implies \boldsymbol{X}_t^- \setminus \mathcal{P}_{Y_t} \not\bowtie Y_t \mid \mathcal{P}_{Y_t}, \tag{3}$$

  from dependence follows connectedness.
  – A variable is conditionally independent of its non-effects given its direct causes.
- **Faithfulness**:
  The joint distribution of a time series process $\boldsymbol{X}$ with graph $\mathcal{G}$ fulfills the Faithfulness condition if and only if for all disjoint subsets of nodes (or single nodes) $A, B, S \subset \mathcal{G}$ it holds that

$$X_Y \perp\!\!\!\perp X_Z \mid X_S \implies Y \bowtie Z \mid S, \tag{4}$$

  that is, from independence follows separation, which includes its logical contraposition

$$Y \not\bowtie Z \mid S \implies X_Y \not\perp\!\!\!\perp X_Z \mid X_S, \tag{5}$$

  from connectedness follows dependence.

  - Every conditional independence in the data must correspond to a separation in the causal graph (no accidental cancellations).

- **Causal sufficiency**:
  A set $W \subset V \times \mathbb{Z}$ of variables is causally sufficient for a process $\boldsymbol{X}$ if and only if in the process every common cause of any two or more variables in W is in W or has the same value for all units in the population.

A.3 Relationship Between Assumption Sets

While CaStLe assumptions (T1-S2) and causal discovery assumptions serve different purposes, there are important interactions between them:

- CaStLe assumptions create an environment where causal discovery becomes tractable in some high-dimensional gridded settings.
- CaStLe assumptions do not guarantee causal discovery assumptions will be satisfied.
- For example, even in perfectly stationary systems (T2, S2 satisfied), faithfulness can be violated through counteracting mechanisms, as demonstrated in Runge (2018).
- Similarly, the Causal Markov Condition is a property of the joint distribution that cannot be derived from locality assumptions.

Instead of replacing causal discovery assumptions, CaStLe's assumptions create a context where causal discovery methods can be applied efficiently to high-dimensional space-time data.

A.3.1 CaStLe's Implementation and Causal Sufficiency

One meaningful connection exists between CaStLe's implementation and causal discovery assumptions: When CaStLe focuses on identifying only the parents of the center cell while including all potential spatial neighbors (per assumption S1), causal sufficiency is automatically satisfied for that specific node by construction - assuming S1 holds true.

This is a significant benefit, as causal sufficiency is typically the most difficult assumption to guarantee in practice (Spirtes et al., 1993; Raghu et al., 2018). While CaStLe cannot guarantee faithfulness or the Markov condition holds, its design cleverly leverages spatial structure to address causal sufficiency within each local analysis.

A.4 Potential Violations and their Manifestations

Violations of CaStLe's assumptions can occur in various ways, leading to different manifestations in the causal discovery process. Violations of CaStLe's assumptions can affect results in different ways:

1. Violations of Temporal/Spatial Locality (T1, S1): If causal effects extend beyond immediate neighbors, CaStLe will miss these connections, creating false negatives.
2. Violations of Stationarity (T2, S2): If dynamics change across space or time, CaStLe's stencil will represent only an average pattern, potentially creating both false positives and negatives.
3. Even with CaStLe assumptions holding, traditional faithfulness violations can occur through cancellation effects or deterministic relationships.

Below, we provide examples of how these assumptions can be violated and their potential impacts, drawing on the discussion by Runge (2018).

A.4.1 Temporal and Spatial Locality (T1, S1)

- *General Violation*: These assumptions can be violated by any process that introduces dependencies beyond immediate temporal or spatial neighbors.
- *Example – Time Aggregation*: Time aggregation can violate temporal locality by introducing dependencies across multiple time steps. Runge (2018) discusses how time aggregation can cause such violations (Section IV.B, Example 4). Figure 5 in Runge (2018) illustrates the impact of time aggregation on causal inference.
- *Example - Spatial Aggregation*: Similarly, spatial aggregation can violate spatial locality by introducing dependencies across non-neighboring spatial units.

A.4.2 Temporal and Spatial Causal Stationarity (T2, S2)

- *General Violation*: These assumptions can be violated by any process that introduces changes in the causal relationships over time or space.
- *Example – Counteracting Mechanisms*: Counteracting mechanisms or heterogeneous processes can violate these stationarity assumptions. If the data contains opposing generating processes (e.g., different hemispheres in climate data), the faithfulness assumption may be violated. This results in unstable and inconsistent causal relationships. Runge (2018) discusses such violations in Section IV.C, Example 5, and provides an illustration in Figure 6.

Understanding potential violations and their manifestations is crucial for applying our framework effectively in realistic scenarios. Section 4.6 outlines practical strategies to mitigate these violations.

Appendix B Statistical and Time Complexity

In this section, we elaborate on Section 4.4 and provide a more detailed discussion of the time-complexity (Appendix B.1) and statistical (Appendix B.2) properties of CaStLe. Additionally, we provide analyses giving conditions under which CaStLe is (asymptotically) guaranteed to recover the true causal graph, independent of the specific PIP used.

B.1 Time Complexity

Steps A, B, and D of CaStLe consist primarily of copying and rearranging of data, so we focus our analysis on the complexity of Step C, which dominates the runtime of CaStLe. Because CaStLe can use a variety of PIPs within Step C, we begin with a general analysis of the worst-case time complexity of causal discovery algorithms. Throughout, recall that a runtime complexity $\mathcal{O}(f(n))$ implies there exists a fixed constant $C \geq 0$ such that that the algorithm terminates in at most $Cf(n)$ steps *for any input of size* n.

Kalisch and Bühlmann (2007) and Runge (2018) discuss the time complexity of causal discovery, particularly the PC algorithm. Much of constraint-based causal discovery is descendant of PC, and it represents a valuable baseline for comparing the computational complexity of CaStLe and prior work. Causal discovery is largely bounded by how long it requires to determine independence between nodes (bounded by samples and size of conditioning sets of nodes) and how many times it needs to do so (generally bounded by the number of nodes). Runge (2018) cite the time complexity of a single conditional independence test using ordinary least squares (linear partial correlation), while Kalisch and Bühlmann (2007) explore bounds on the number of tests in PC. Our analysis is consistent with theirs, which we derive from first principles.

Consider causal discovery in p-dimensions (p measured variables) with n samples; suppose further that it is known, *a priori*, that any node in the causal graph has at most degree q: that is, no element has more than q causal parents. An exhaustive search for the causal parents of a *single node* will require evaluating $\sum_{i=0}^{q} \binom{p}{i} = \mathcal{O}(2^p)$ possible sets of parents; repeating this process for all p nodes evaluation of up to $\mathcal{O}(p2^p)$ possible causal graphs. If we construct graphs using statistical tests for linear partial (conditional) correlation, each test can be performed in $\mathcal{O}(n\,p\,\min\{n,p\}) = \mathcal{O}(np^2)$ time (the time required to fit an OLS regression to n observations and p variables using a direct method such as an SVD or QR factorization), yielding an overall runtime of

$$\mathcal{O}(np^2 * p2^p) = \mathcal{O}(np^3 2^p).$$

This analysis is quite loose, and as Runge (2018) notes, the complexity of a *single* linear conditional independence test can be reduced to $\mathcal{O}(np^2q^2)$ when efficient algorithms are used. Far stronger guarantees can be provided for specific causal discovery algorithms that more efficiently search the space of possible graphs. Regardless, even this rough analysis will be sufficient to demonstrate the algorithmic improvements attained by CaStLe.

We now consider the specific context of causal discovery from gridded time series data. Here, we have $n = T$ total observations and have $p = N^2$ features of our data. Direct application of causal discovery to this data gives a worst-case complexity of

$$\mathcal{O}(np^3 2^p) = \mathcal{O}(T(N^2)^3 2^{N^2}) = \mathcal{O}(TN^6 2^{N^2}),$$

so the complexity of standard causal discovery methods grows *super-exponentially* with the size of the grid. For the purposes of direct comparison to CaStLe, where $p = N^2$, we assume PC's $\tau_{max} = 1$. By contrast, the reduced space where CaStLe's PIP operates has $T(N-2)^2$ observations and only $p = 9$ features, yielding a *polynomial* worst-case runtime of

$$\mathcal{O}(np^3 2^p) = \mathcal{O}(T(N-2)^2 * 9^3 * 2^9) = \mathcal{O}(TN^2).$$

Even for grids of relatively modest size, this improvement can be significant: consider a small 30×30 grid; at 1° resolution, this covers approximately 1.5% of the globe. Unstructured causal discovery methods need to consider approximately $30^6 * 2^{30}$ possible graphs, while CaStLe needs to evaluate only $9^3 * 2^9 = 373{,}248$ graphs, representing an improvement of approximately 2×10^{12}-fold. Specific PIPs may provide less dramatic improvements, but it is clear that CaStLe can be expected to be millions-if not billions-of times more efficient than existing approaches.

Note that in our application scenarios, CaStLe is always applied to a square $N\times N$ grid. However, more generally we can consider p grid cells. Traditional causal discovery will be bounded by

$$\mathcal{O}(Tp^3 2^p),$$

while CaStLe will be bounded by

$$\mathcal{O}(Tp).$$

Thus, if grid cells scale linearly, CaStLe scales linearly in both samples and grid cells.

B.2 Statistical Consistency

Statistically, we see that CaStLe can achieve significantly improved estimation performance compared to a full graph inference approach. Rather than give a general analysis, we rely on the prior work of Kalisch and Bühlmann (2007) to compare CaStLe-PC with the standard PC algorithm. Using the same definitions of n, p, q as in our previous analysis, Kalisch and Bühlmann (2007, Appendix B) show that the probability of the PC algorithm incorrectly estimating the causal graph incorrectly is bounded above by

$$P[\hat{\mathcal{G}} \neq \mathcal{G}] = \mathcal{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right).$$

In our setting, this gives an error probability of

$$\mathcal{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right) = \mathcal{O}\left((N^2)^{N^2+2}(T-N^2)e^{-c(T-N^2)}\right) = \mathcal{O}\left(N^{2N^2}e^{cN^2} * Te^{-cT}\right)$$

for PC applied in the original data space. It is clear that this quantity grows rapidly in N, consistent with the intuition that causal discovery algorithms struggle when applied to larger spatial domains. By contrast, this analysis implies that the error probability of CaStLe-PC scales as

$$\mathcal{O}\left(p^{q+2}(n-q)e^{-c(n-q)}\right) = \mathcal{O}\left(9^{9+2}(T(N-2)^2-9)e^{-c(T(N-2)^2-9)}\right) = \mathcal{O}\left(\frac{TN^2}{e^{TN^2}}\right)$$

Quite surprisingly, this *decreases* with the graph size (N), implying that CaStLe actually achieves *better performance* when applied to larger spatial domains. We demonstrate the remarkable practical effect of this scaling in Section 6.1. Similar improvements can be shown for any base causal discovery algorithm (and associated PIP) for which precise estimates of statistical convergence rates are available.

Appendix C Asymptotic Consistency

We examine the asymptotic consistency of CaStLe, with a particular focus on the Parent Identification Phase (PIP). Asymptotic consistency is a fundamental property that ensures the accuracy of causal graph estimates as the number of observations increases. We begin by establishing the technical assumptions necessary for our analysis, specifically those related to the p-values generated by the PIP for edge existence. These assumptions are critical for maintaining control over both false positive and false negative rates, thereby ensuring the reliability of our causal inferences. The central theorem we present demonstrates that, under these conditions, CaStLe achieves asymptotic consistency as the number of nodes approaches infinity. In the case of Bayesian score optimization causal discovery, such as DYNOTEARS, Bayesian posterior probabilities can be used in lieu of p-values with suitable minor modifications to the combination procedure. The proof is structured into three parts, addressing the independence of observations, the application of Fisher's method for combining p-values, and the implications of using overlapping regions. Through this analysis, we aim to reinforce the validity of our algorithm and its effectiveness in uncovering causal relationships in gridded space-time data structures.

Technical Assumption (P1):

- The Parent Identification Phase, PIP$(\cdot)$, produces p-values for edge existence, which satisfy the following:
  - For every non-edge (i,j) $(j \notin \mathscr{P}(i))$, $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) = u$ for all $u \in [0,1]$; that is $p_{\text{PIP}}^{(i,j)} \sim \mathcal{U}([0,1])$ is uniformly distributed.
  - For every edge (i,j) $(j \notin \mathscr{P}(i))$ and every $T > T_0$, there exists $\pi_{(i,j)}^{T}(u) > 0$ such that $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) \leq \max\{0, u - \pi_{(i,j)}^{T}(u)\} < u$ for all $u \in [0,1]$.

  Taken together, these require that the PIP$(\cdot)$ control the false positive rate at the nominal significance level used and that the false negative rate is less than the false positive rate.

Here, T_0 is a minor technical assumption to allow the PIP to have non-trivial accuracy: we use it to exclude trivial cases like $T = 1$, in which no time series causal discovery mechanism can be accurate.

Additionally, note that we typically assume that the PIP$(\cdot)$ is asymptotically consistent, so that $\pi_{(i,j)}^{T}(u)$ is bounded above 0 for all u as $T \to \infty$. This can be used to prove T-asymptotic consistency of CaStLe, but in this section we aim only to prove N-asymptotic consistency.

Theorem: Suppose $\mathcal{D}$ is an $\mathbb{R}^{T\times N\times N}$ realization of a data-generating process satisfying T1-S2. Suppose also that $\mathsf{PIP}(\cdot)$ is a parent-identification-phase satisfying P1. Then, there exists a T_0 such that for any $T \geq T_0$, CaStLe is asymptotically consistent as $N \to \infty$; that is, the causal graph estimated by CaStLe converges to the true causal graph generating $\mathcal{D}$ with probability 1.

Proof. This proof proceeds in three parts:

- First, we argue that, for large N, well-separated (non-overlapping) spatial regions can be considered IID realizations.
- Next, we argue that the application of Fisher's method leads to asymptotic consistency of CaStLe.
- Finally, we argue that "infill" of the overlapping regions does not invalidate the asymptotic consistency.

At a high level, we argue that, because it is T-asymptotically consistent, there exists some T_0 where the PIP has non-trivial power. We then apply standard statistical methods for combining several weak p-values to obtain a global strong p-value. The technical bookkeeping of our argument serves primarily to deal with the fact that we use overlapping spatial regions and cannot assume independence of the individual p-values; we overcome this by selecting regions that are sufficiently spatially separated to be statistically independent on the time scale considered.

Without loss of generality, we focus on asymptotically consistent estimation of a single edge, say $(\mathsf{East}, \mathsf{Center})$. Extension to all 9 stencil edges follows immediately by a standard union bound argument.

Part I: For analytical simplicity, we divide the spatial region into square regions of size $(5+2T)\times(5+2T)$. On a grid of size $N \times N$, there are $B_{N,T} = \lfloor N/(5+2T) \rfloor$ such regions. We apply the $\mathrm{PIP}(\cdot)$ to the center 3×3 region of each region separately, obtaining $B_{N,T}$ p-values for the existence of the edge. Because these central regions are separated by (at least) $2T{+}2$ grid cells and causal effects exist at a distance of at most $2T$ under our data generating model, these p-values can be treated as statistically independent. (This is essentially the same argument used by Goerg and Shalizi (2013), though their application is quite different.)

Part II: Given $B_{N,T}$ independent p-values, we then apply Fisher's method for combining p-values. Specifically, given a set of p-values for edge non-existence, Fisher's method controls the *familywise error-rate*, rejecting the global null (no edges anywhere). By our assumption of spatial homogeneity, if an edge exists in at least one region, it must exist everywhere, so Fisher's method precisely tests for edge existence in the stencil.

Recall that Fisher's method constructs a test statistic $T = -2\sum_{b=1}^{B}\log p_b$ and tests it against a null χ^2_B distribution. We consider two cases:

1. If the edge does not exist, each p-value is $\mathcal{U}([0,1])$ by construction and the test statistic T follows its null distribution. So long as the global significance level used for Fisher's test α_{Fisher} is converging to 0 as $N \to \infty$, we have asymptotic consistency for edge absence.
2. If the edge does exist, each p-value is less than α with probability $(1{+}c)\alpha$ for some c strictly positive. We then have that T has a *non-central* χ^2 distribution, which is asymptotically distinguishable from a (central) χ^2 at all significance levels as $N \propto B \to \infty$.

Taken together, these guarantee the the output of Fisher's method is asymptotically consistent for both edge presence and edge absence.

Part III: In practice, we apply CaStLe not to disjoint regions but to overlapping regions. As discussed elsewhere, the region-discretization strategy and the use of Fisher's method are such that this does not cause "cross-contamination" or invalid tests of edge existence. We note here that this strategy also does not invalidate asymptotic consistency of CaStLe. Specifically, we note that, with overlapping regions, the p-values used in Fisher's method may no longer be assumed independent.

In this case, however, this is not an issue as they exhibit positive dependence (as they are taken from overlapping data). As such, the true degrees of freedom of T under the null are less than the nominal degrees of freedom; this leads Fisher's method to be (if anything) overly conservative in finite samples. Hence, for the case of edge absence, the nominal significance level is understated and we retain consistency as long as we take $\alpha_{\text{Fisher}} \xrightarrow{N\to\infty} 0$; for the case of edge presence, it suffices to note that the true sampling distribution is still asymptotically distinguishable from the null (since each individual p-value is powerful), so we retain consistency. □

We note that Fisher's method may not be the optimal method for combining p-values. In particular, Holm's method allows for arbitrary dependence of the p-values, likely yielding better performance at finite N, but we do not pursue this approach here as the implementation and theoretical analysis are somewhat more difficult. As with Fisher's method, Holm's method controls the error rate of the global null which, under our assumptions of causal stationarity, is precisely the correct null for accurate stencil estimation.

Additionally, we note that the p-values produced by the PIP under the null do not need to precisely satisfy a uniform distribution; conservative p-values decrease the value of Fisher's statistic T, thereby lowering the rate of false positives.

Remark: If PIP$(\cdot)$ is strongly asymptotically consistent as $T \to \infty$, it must satisfy assumption P1.

Proof. We argue by contradiction. Suppose that PIP$(\cdot)$ were not asymptotically consistent and that the false positive rates and false negative rates of the PIP were equal (or worse, the false negative rate was greater than the false positive rate). Specifically, assume that there exists a true edge (i, j) and some $\pi_- > 0$ such that $\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq u) > \pi_- + u$ for all T and all u. For the PIP to guarantee no false positives, we must take $\alpha \to 0$ as $T \to \infty$. But this would imply that there remains an asymptotic π_- probability of a false negative ($\mathbb{P}(p_{\text{PIP}}^{(i,j)} \leq \alpha) > \alpha + \pi_i \geq \pi_- > 0$), contradicting our assumption of asymptotic consistency.

□

Appendix D Application to Non-Linear Dynamics: Continuous Systems via Burgers' Equation

This appendix extends our validation of CaStLe to non-linear dynamical systems through application to Burgers' equation, demonstrating the method's effectiveness beyond the linear systems discussed in the main text.

Having established the strong performance of CaStLe on discrete models of linear dynamics, we turn to a far more challenging domain: continuous models with non-linear PDEs. Specifically, motivated by our interest in turbulent atmospheric dynamics, we consider Burgers' equation, a PDE used to model a combination of advective (directed flow) and diffusive processes (Burgers, 1948). While initially developed to model fluid flows, Burgers' equation has been successfully applied to a variety of fields, such as turbulence, non-linear wave propagation, traffic flow, cosmology, gas dynamics, and more (Bonkile et al., 2018). In the following experiments, we again implemented CaStLe's PIP with the PC-Stable-Single algorithm.

We note that the interaction of PDE dynamics with causal language is rather subtle: while PDEs are imbued with a "forward" direction in time, the actual numerical methods used to solve them include "forward" and "backward" steps in the underlying integrators as well as sophisticated interpolation schemes. Our focus here is not on finding a causal model for the PDE solution per se, but on identifying the structure of the underlying advection. This choice is motivated in part by the results of Rubenstein et al. (2018), who explored the related problem of identifying causal models from deterministic ordinary differential equations (ODEs). As they note, there is not generally a single causal graph corresponding to an ODE, with different models being appropriate at equilibrium or under various interventions. Given the additional complexity of PDEs, we believe that identifying the underlying advection angle provides the most meaningful causal representation of Burgers-type dynamics, particularly as it relates to our volcanic eruption aerosol case study.

D.1 Burgers' Equation: Model and Parameters

In two dimensions, Burgers' equation can be written as:

$$\frac{\partial u}{\partial t} + \underbrace{u\left(\alpha\frac{\partial u}{\partial x} + \beta\frac{\partial u}{\partial y}\right)}_{\text{Advective Dynamics}} = \underbrace{c\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right)}_{\text{Diffusive Dynamics}} + f \tag{6}$$

where α, β are the advection coefficients in the x, y directions, capturing directed flow dynamics; c is the diffusion coefficient; and f is a forcing term representing additional mass being injected into the system. In order to create a closed system with no exogenous forcings, we take $f = 0$ uniformly throughout this section.

The left panel of Figure D1 shows three different solutions to Burgers' equation at different advection angles (θ), advection strength ($M = \sqrt{\alpha^2 + \beta^2}$), and diffusivities ($c$), each with the same initial conditions. Examining the time evolution of these solutions (left to right), we see that the high-advection low-diffusion systems (top) exhibit a clear direction of flow, while it is far more difficult to find direction in low-advection high-diffusion systems (bottom). We take inferring the angle of advection as our principal task: given an observed solution u to Equation (6), can we determine the angle of the underlying advective dynamics?

D.2 Advection Angle Estimation

Given a CaStLe-estimated stencil, we infer the angle of underlying advection in the following manner: i) identify each potential parent edge of `C` with a vector, taking the angle of the underlying edge in the reduced space as direction and the (signed) strength of the underlying relationship as magnitude; ii) sum these vectors to obtain an aggregate estimate of the advective dynamics; iii) take the angle of the vector sum as an estimate of the underlying advection angle. In pseudo-code, we can write this as

$$\hat{\theta} = \texttt{atan2}\left(\sum_{l \in \mathscr{P}(\texttt{C})} e_l \sin\theta_l, \sum_{l \in \mathscr{P}(\texttt{C})} e_l \cos\theta_l\right).$$

Here `atan2` is the signed arctangent function, $\mathscr{P}(\texttt{C}) = \{\texttt{NW}, \texttt{N}, \ldots, \texttt{W}\}$ represents all potential parents the center cell, e_l represents the strength of that edge (0 for non-present

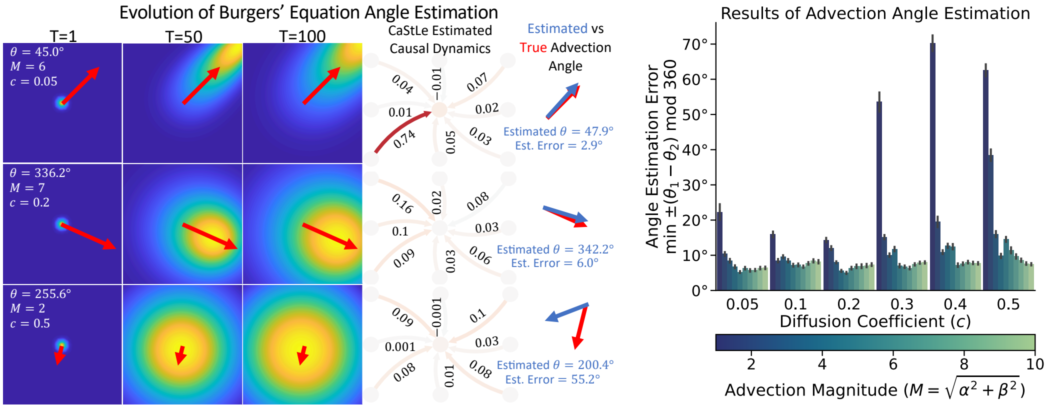


Figure D1: Application of CaStLe-PC to advection estimation from non-linear PDE dynamics. In the left panel, the first three columns depict realizations of Burgers' equation under different advection-to-diffusion regimes; the fourth column depicts the causal stencil identified by CaStLe-PC; and the final column compares the estimated advection angle with the true advection angle. The right panel depicts the accuracy of CaStLe-PC under various signal-to-noise conditions. Each combination of advection and diffusion rates were tested with 500 angles sampled uniformly from $[0°, 360°)$. In low-diffusion (high SNR) scenarios, CaStLe-PC can identify the underlying advection clearly (top row of left panel and yellow-green columns in right panel). By contrast, in low-advection (low SNR) scenarios, CaStLe-PC struggles to accurately identify the underlying advective dynamics (bottom row of left panel and blue bars in right panel). Even in highly diffusive scenarios, CaStLe-PC is able to accurately estimate the underlying advection when it is sufficiently strong (around $M/c \geq 20$) as shown in the middle row of the left panel. Additional details are given in Appendix D.

edges), and θ_l represents the angle of that edge $(135°, 90°, \ldots, 180°)$. This process allows us to estimate all angles instead of just the eight angles present in the stencil structure.

D.3 Experimental Setup

In order to assess the effectiveness of CaStLe-PC in a variety of regimes, we generate (approximate) solutions to Equation (6) with 500 angles sampled uniformly from $[0°, 360°)$, advection magnitudes varying from 1 to 10 and diffusion coefficients from 0.05 to 0.5. The diffusion-free ("noiseless") case of $c = 0$ is numerically unstable. To compute the simulated Burgers' dynamics, we use MATLAB's default PDE solver (`pdesolve`) on a circular mesh of radius 3 and 100 time steps equally spaced between $t = 0$ and $t = 1$. Then we interpolated the finite-element solution onto a grid of size 25×25, covering the square $[-1, 1]^2$, yielding spatial points that are approximately 0.1 units apart. We restrict our solution to avoid any boundary conditions. Finally, we apply CaStLe-PC and the aforementioned advection angle estimation method, and compare the estimated angle to the true angle. We demonstrate three realizations of this process in the left-hand panel of Figure D1.

D.3.1 Angle Estimation Results

Our results appear in the right panel of Figure D1, where we plot the difference in the true and estimated angle, taking care to account for the "wrapping" behavior of angle-valued data. We see that stronger advection (higher SNR) consistently leads to improved estimation (downward trend within each group), with estimated angles consistently within 10° for advection magnitude 5 or greater. Comparing across different levels of the diffusion coefficient c, we note that higher c increases the angle estimation error, as we would expect in the higher-noise regimes. For low advection magnitude and $c \geq 0.3$, we see an average error approaching the "pure guessing" value of 90°. Even at high diffusion levels ($c = 0.5$), moderate advection magnitudes of 5-6 are sufficient to ensure accurate estimation. From these, we see that CaStLe-PC is able to consistently recover advection structure across a wide range of SNR regimes. As demonstrated in Appendix F, traditional dimension reduction approaches such as PCA and PCA-varimax, when combined with standard causal discovery methods, fail to accurately capture the advection dynamics in Burgers' equation, particularly in identifying the correct advection angle. This highlights CaStLe's unique ability to preserve and extract meaningful causal structures from nonlinear PDE systems that would otherwise be lost through dimensionality reduction.

The takeaway from these results is that CaStLe can not only generalize to continuous, non-linear models of advection and diffusion, but it can successfully infer the direction of causality in any advective-diffusive system, given that the diffusion is not so large as to dominate advection. Further, each simulation has only one signal surrounded by large areas without data or causal information. Despite this sparsity and the presence of regions where diffusive information flow might suggest incorrect advection angles, CaStLe successfully identifies the correct advection angle when analyzing the full space. CaStLe is asked to learn from the full space, but successfully hones in on the correct advection angle. With these results, we believe CaStLe can be applied to a broad range of space-time systems with advective-diffusive properties to better understand their dynamics.

Appendix E Proposed Modification of Statistical Methods for CaStLed Data

While essentially any consistent PIP may be used in Step C, we anticipate that most PIPs will be derived from already existing causal discovery algorithms. Often, these algorithms are statistical in nature and it may be inappropriate to apply them directly to $\tilde{\boldsymbol{X}}$ due to the *seams* connecting each time *chunk*. For a statistical method, which computes a p-value for each potential edge (smaller p-values leading to present edges), we suggest the following *chunk testing* modification:

1. For each chunk $b \in \{1, \ldots, (N-1)^2\}$, let p_b be the p-value resulting from the PIP applied to that chunk.
2. Compute $T = -2\sum_b \ln p_b$
3. Let $p_{\text{agg}} = 1 - \chi^2_{2(N-1)^2}(T)$ where $\chi^2_k(x)$ is the cumulative distribution function (CDF) of a χ^2 random variable with k degrees of freedom evaluate at x.
4. If $p_{\text{agg}} < p_*$, identify a parent.

This method adapts Fisher's classical method for combining independent p-values to our setting. In practice, however, we have found that for sufficiently large T, this *chunking* is unnecessary as the proportion of *seams* in $\tilde{\boldsymbol{X}}$ goes to zero, and the PIP identifies the correct causal structure despite the small fraction of points of misspecification $(1/T)$.

Appendix F Limitations of Dimensionality Reduction for Space-Time Causal Discovery

We demonstrate the limitations of dimensionality reduction approaches such as PCA and PCA-varimax when applied to space-time causal discovery of advective-diffusive processes. Causal discovery methods in Earth science often employ these techniques to reduce the high dimensionality of gridded data before applying causal discovery algorithms. While effective for identifying large-scale teleconnections, we show that these approaches fail to capture the local causal structures that are essential for understanding space-time dynamics at the grid-cell level.

To illustrate these limitations, we apply PCA and PCA-varimax dimension reduction followed by PCMCI causal discovery—–the procedure described by Runge et al. (2015), Nowack et al. (2020), and Tibau et al. (2022) and employed in subsequent work—to each of our case studies: Burgers' equation, HSW-V, and E3SMv2-SPA. Our analysis reveals that while dimensionality reduction techniques can identify dominant modes of variability, they struggle to preserve the spatial relationships between neighboring grid cells, thus obscuring the local causal pathways that CaStLe is specifically designed to recover.

For the PCMCI step, we explored multiple lag values in our experiments and found that the results were consistently unable to capture the directional advection structure regardless of lag parameter choice. This suggests that the limitation is a fundamental constraint of the dimensionality reduction approach. In the results below, we show the simplest case with a maximum lag of 1.

Figure F1 shows the PCA analysis of Burgers' equation, where four EOFs capture approximately 91% of variance but the resulting PCMCI causal graph fails to recover the directional advection process, demonstrating PCA's inability to preserve local causal structures. Figure F2 shows similar limitations with PCA-Varimax applied to the same Burgers' equation data, where despite the rotation enhancing spatial localization of patterns, the causal graph still cannot represent the known directional advection dynamics. Figure F3 illustrates PCA applied to the HSW-V volcanic aerosol dataset, where four EOFs explain 85% of variance but produce a causal graph that misrepresents the known transport mechanisms. Figure F4 demonstrates that even with varimax rotation, which provides more spatially distinct patterns in the HSW-V dataset, the resulting causal graph cannot capture the directional flow of volcanic aerosols. The EOFs were reordered according to the identified centroids' longitude to improve interpretability. Figure F5 shows the application of PCA to the E3SMv2-SPA climate model data, where nine EOFs account for 87% of variance, yet the PCMCI causal graph fails to detect the underlying atmospheric circulation patterns. Figure F6 reveals that PCA-Varimax rotation of the E3SMv2-SPA data, with EOFs similarly reordered by longitudinal position for interpretability, still fails to recover the known directional transport processes, further confirming the limitations of dimensionality reduction for space-time causal discovery.

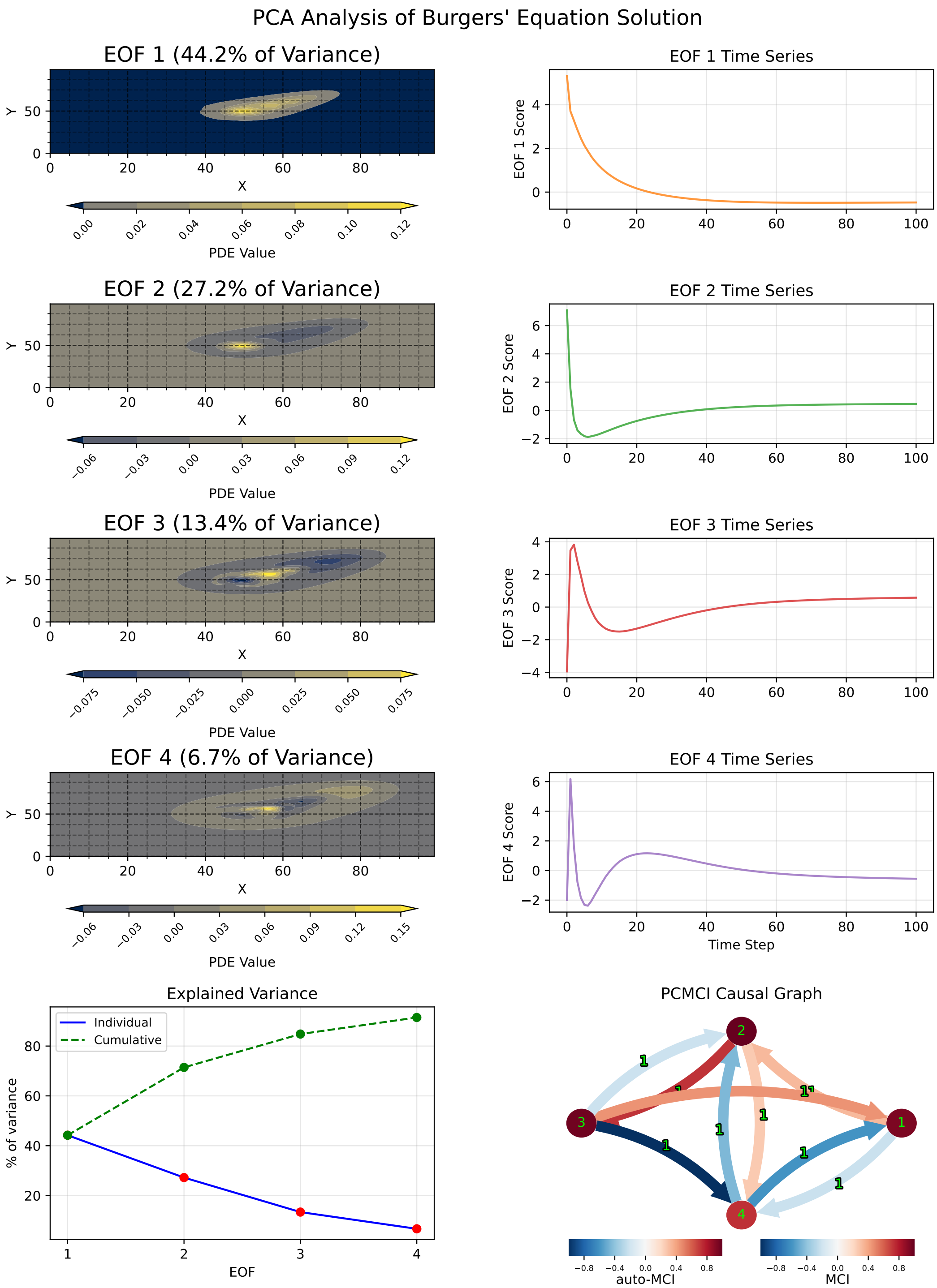


Figure F1: PCA study of Burgers' equation solution ($\theta = 45°$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture ≈91% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time systems.

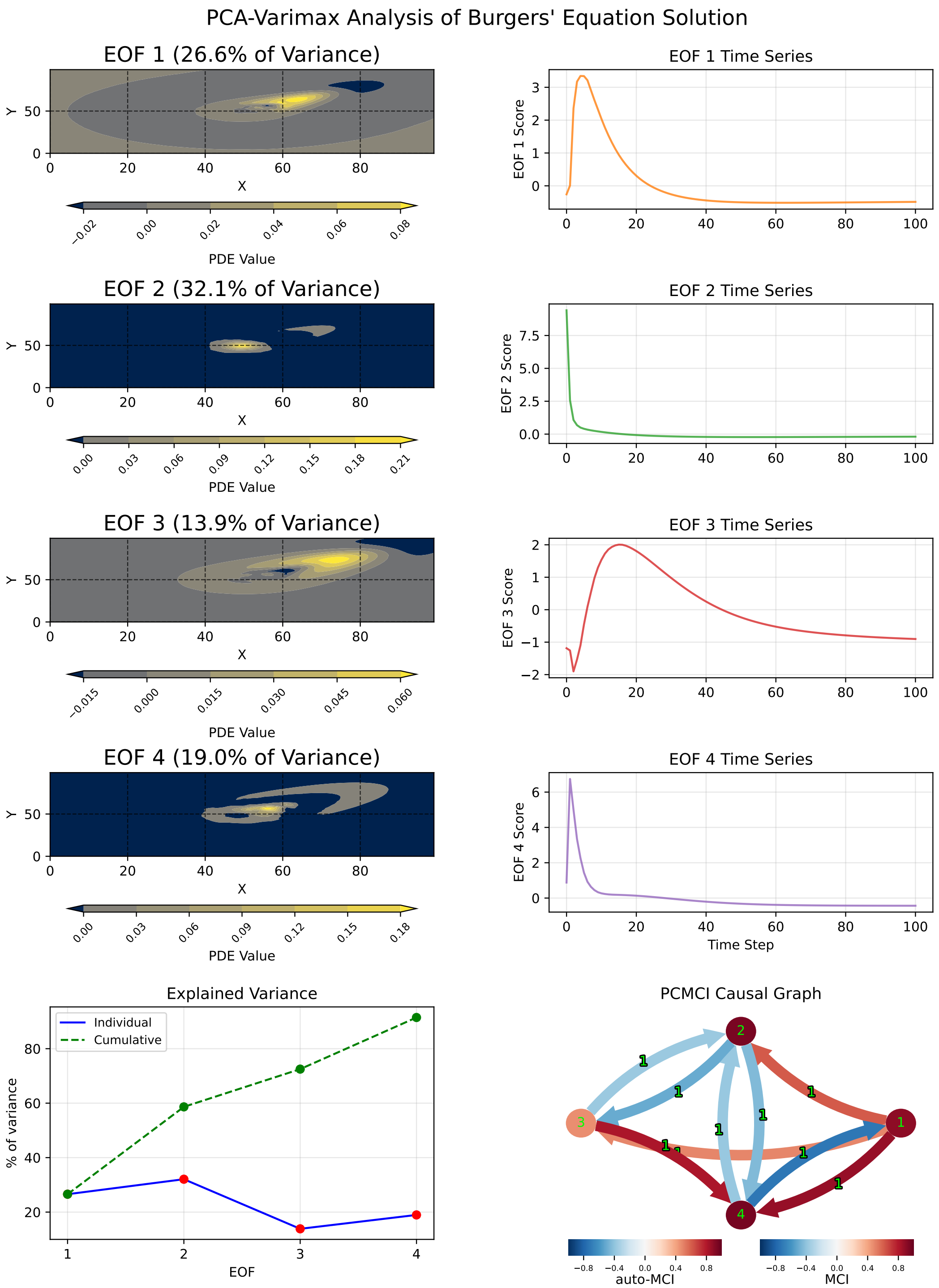


Figure F2: PCA-Varimax study of Burgers' equation solution ($\theta = 45°$, $M = 6$, $c = 0.05$). Four empirical orthogonal functions (EOFs) capture ≈91% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying PDE, highlighting limitations of this approach for local causal structures in space-time systems.

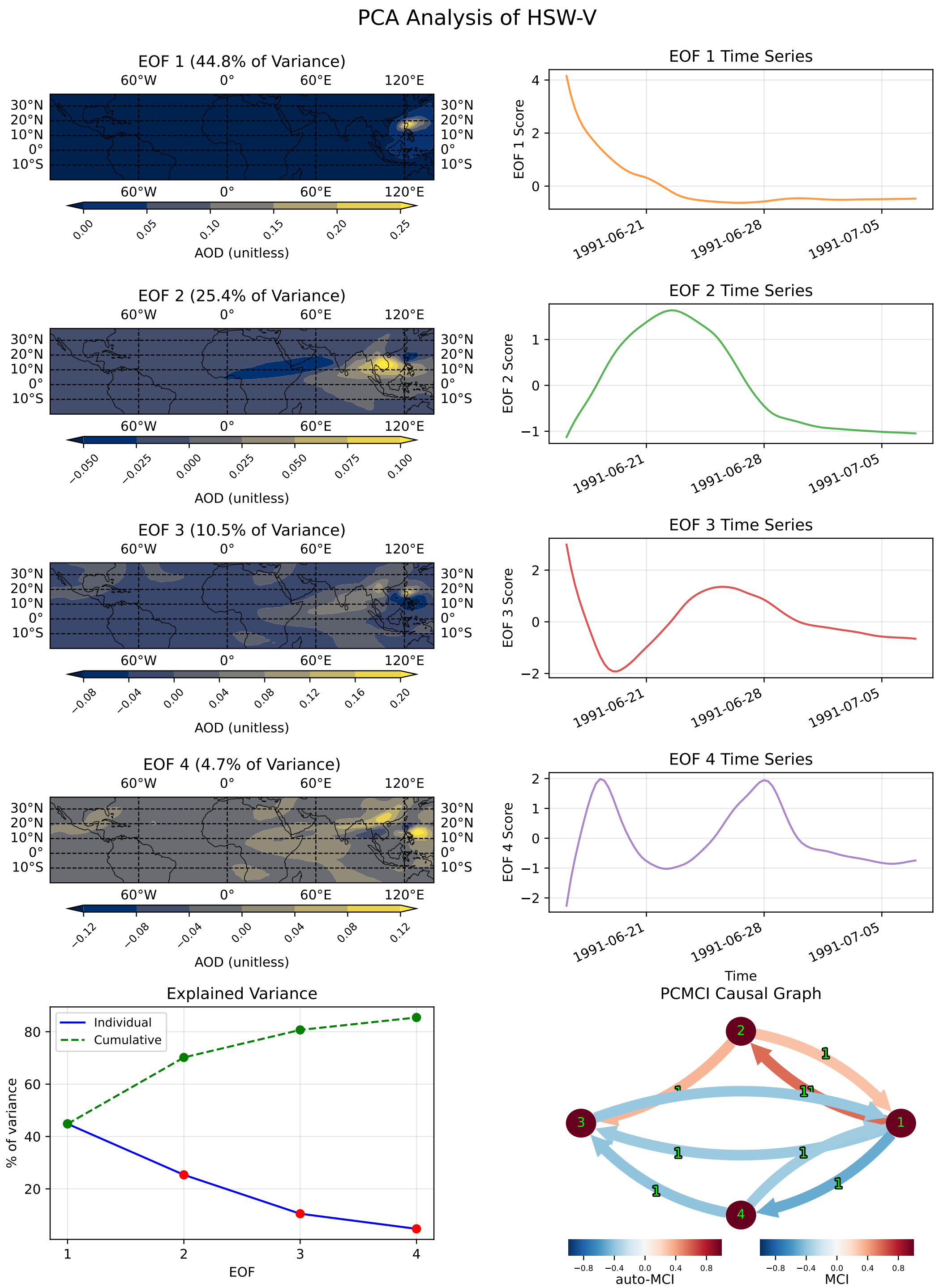


Figure F3: PCA study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture ≈85% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

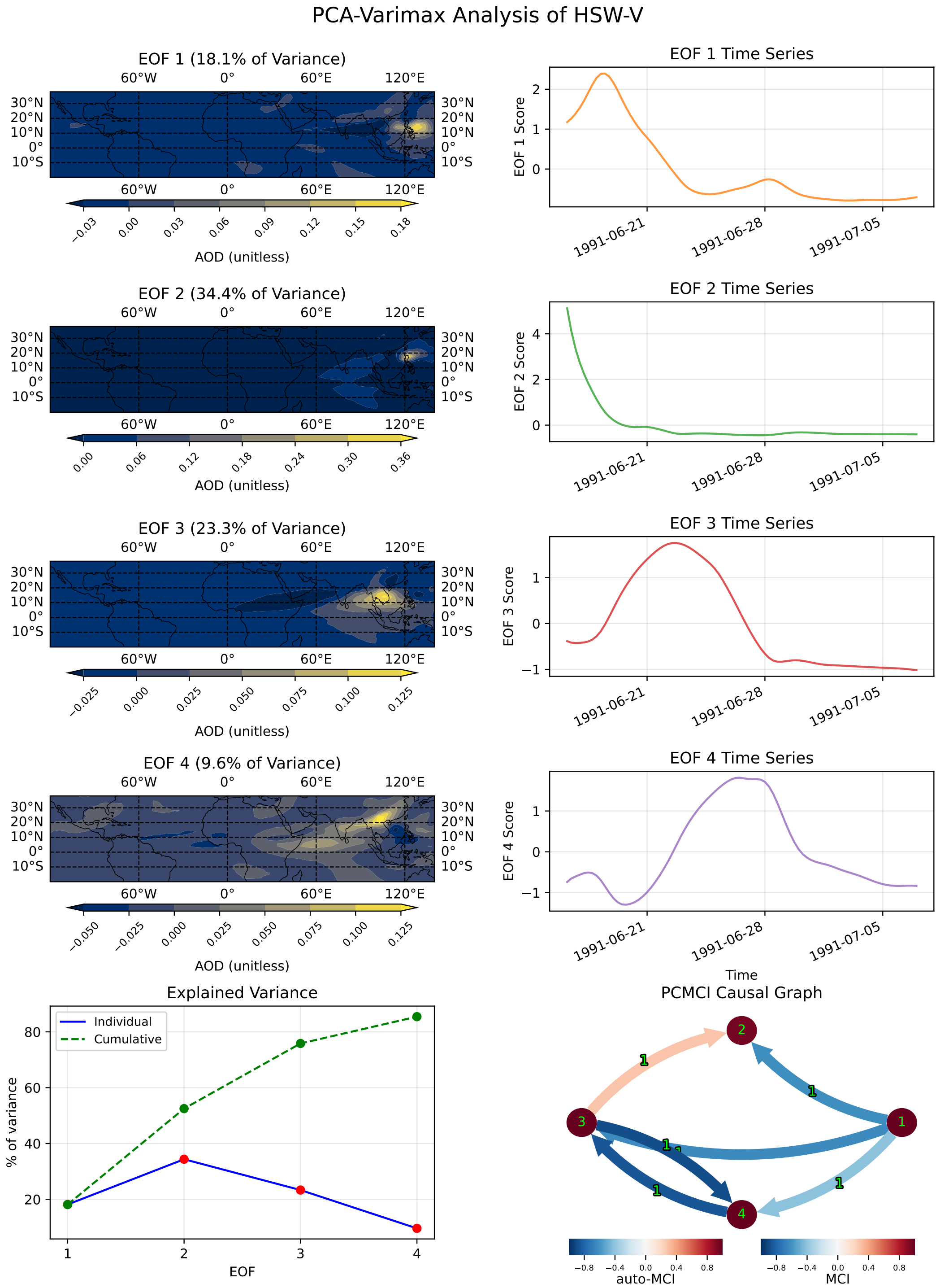


Figure F4: PCA-Varimax study of the HSW-V dataset, in the time interval 21 days post-eruption. Four empirical orthogonal functions (EOFs) capture ≈85% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

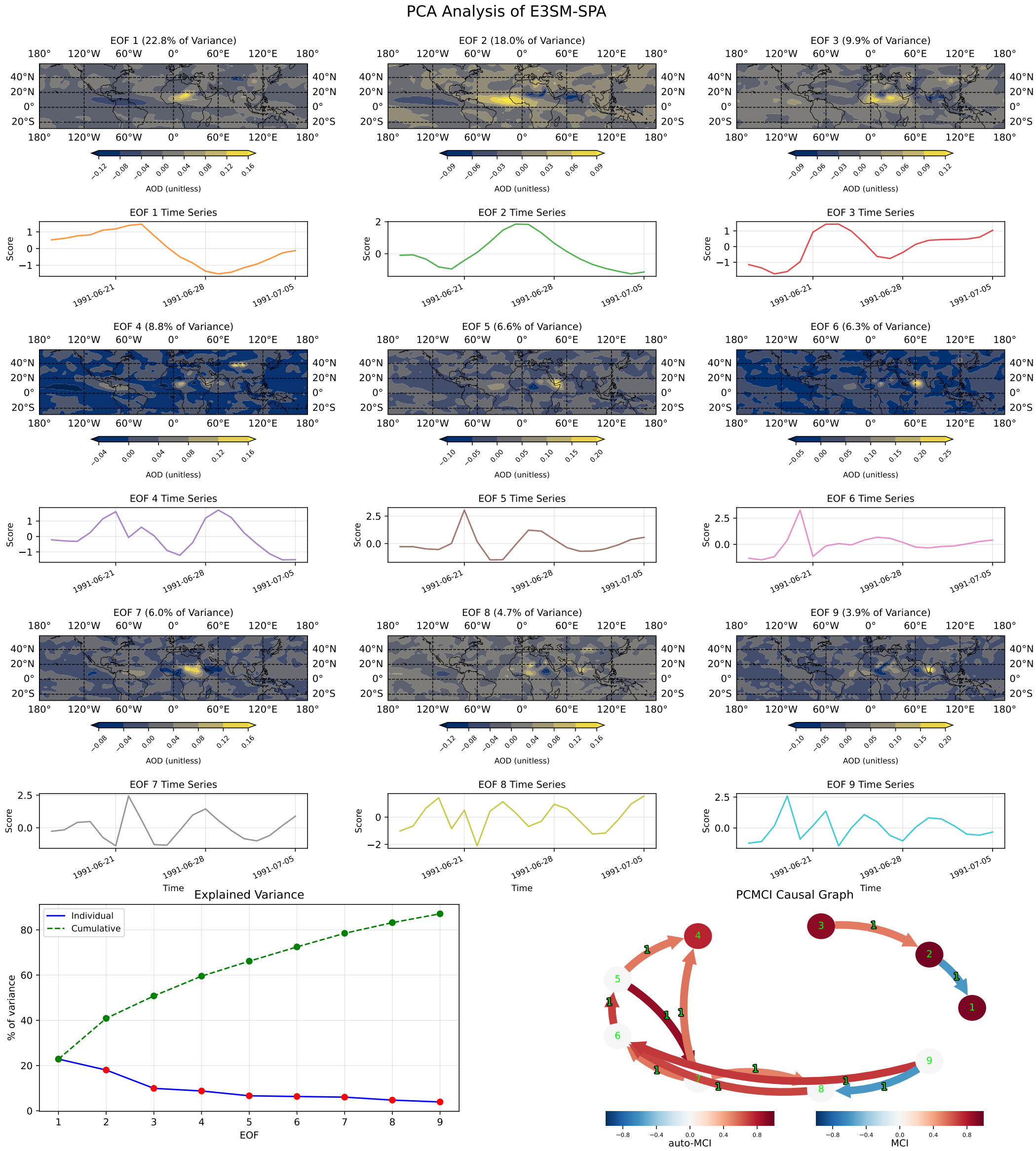


Figure F5: PCA study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture ≈87% of variance, with spatial patterns (left) and temporal evolution (right). The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

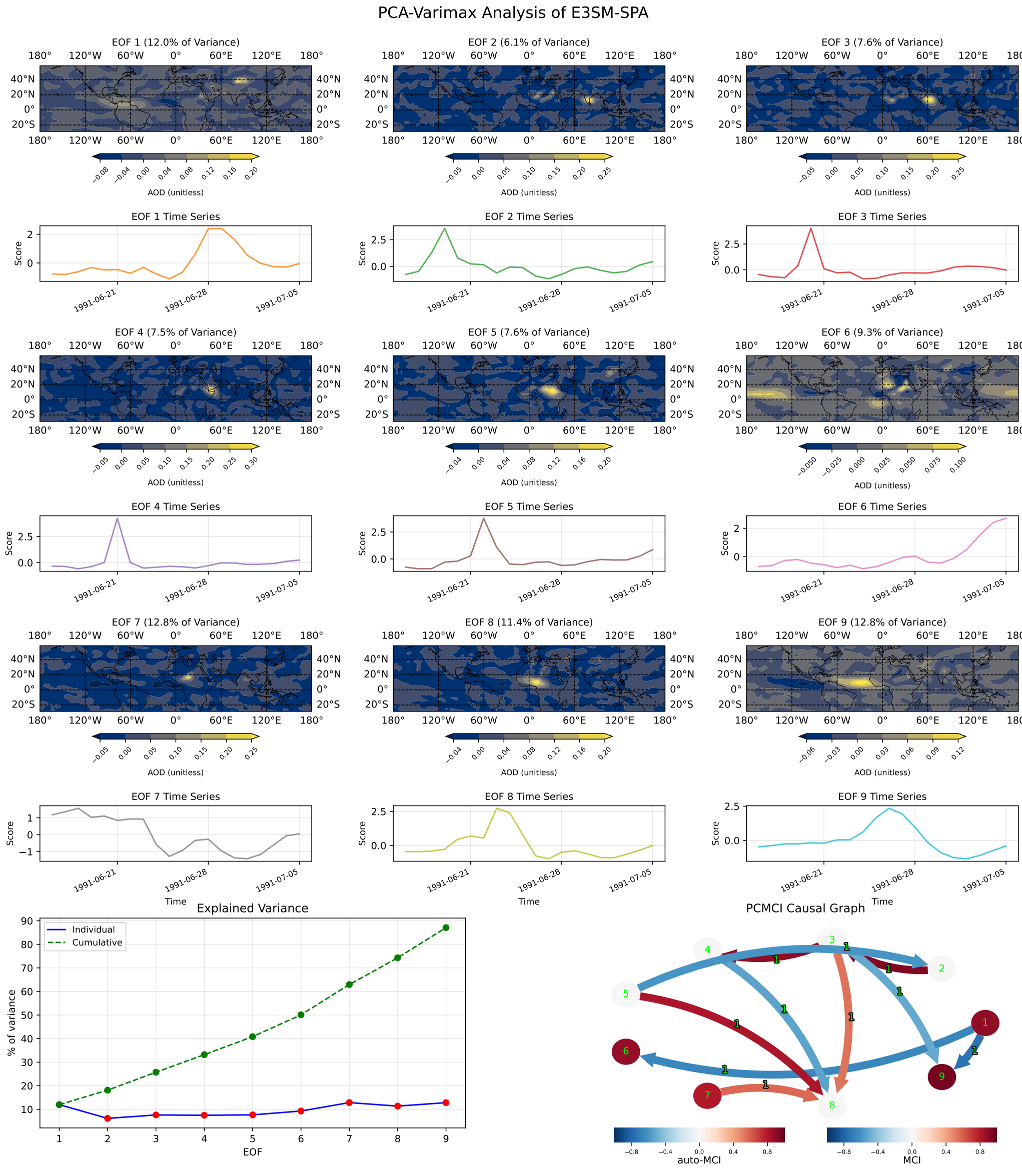


Figure F6: PCA-Varimax study of the E3SMv2-SPA dataset, in the time interval of days 15-35. Nine empirical orthogonal functions (EOFs) capture ≈87% of variance, with spatial patterns (left) and temporal evolution (right). Since varimax rotation does not preserve the explained variance ordering, we reordered EOFs according to the identified centroid's longitude. The bottom panels show explained variance distribution and PCMCI causal graph, which fails to accurately represent the known directional advection process in the underlying system, highlighting limitations of this approach for local causal structures in space-time systems.

Appendix G Additional experimental details for Section 5

CaStLe inherits several of the runtime parameters of the underlying PIP used. In Section 5, we set these values at relatively stringent threshold to highlight the most robust and important dynamics and to yield a highly interpretable graph; additional weaker dynamics can be recovered by relaxing these choices at the (potential) cost of additional false positive edges and less interpretability. Data-driven optimization of these parameters is difficult, though the validation strategies suggested by Allen et al. (2023) may be useful here. Specifically, we set a p-value threshold of 1×10^{-5} and removed estimated partial correlations of magnitude less than 0.35; we note here that, due to the adaptive search heuristics used by the PIP, the p-value threshold applied here is not a proper measure of statistical significance, but only a *heuristic* measure of estimated strength. We note that our resulting interpretations are generally quite robust to specific choices of these values.

Appendix H Analysis of Spatial Blocking

Here, we briefly investigate two impacts of spatial blocking, of the kind used in Section 5. Spatial blocking is a process in which regions of the global space are separated into blocks where CaStLe is applied individually and independently. This can be done for the sake of interpretability and to help ensure the spatial causal structure is uniform and homogeneous in the blocked space, satisfying Assumption S2.

First, we consider the impact of block size on the HSW-V case study. In our demonstration in Section 5.1, we approached block size heuristically, and we chose a relatively large block size to demonstrate correctness saliently. We found that results are generally robust to larger and smaller block sizes in the HSW-V case. In Figure H1, we show that the recovered dynamics in each stencil are generally the same over space for each block size. We see that larger block sizes are easier to interpret at a glance, while smaller sizes describe more nuance. We also found that results were generally robust to block size in the E3SMv2-SPA case.

Second, we consider the impact of a blocking strategy for causal discovery generally by comparing results of the PC algorithm to one block in E3SMv2-SPA to CaStLe-PC's results from the same data. Our comparison of CaStLe and the PC algorithm in Figure 4 make it clear that CaStLe captures the spatial evolution of Mt. Pinatubo's plume much more effectively and about 80,000 times faster. However, one may be concerned that sparsity and correctness could be achieved with blocking alone. In Figure H2a, PC struggles to estimate an interpretable and physically meaningful graph of the dependence structure in this area because of the signal redundancy between nonadjacent grid cells and that there are only 20 observations per grid cell and 25 grid cells. Figure H2b illustrates much better performance from CaStLe, in which CaStLe learns a stencil from the region and projects it back into the original grid space.

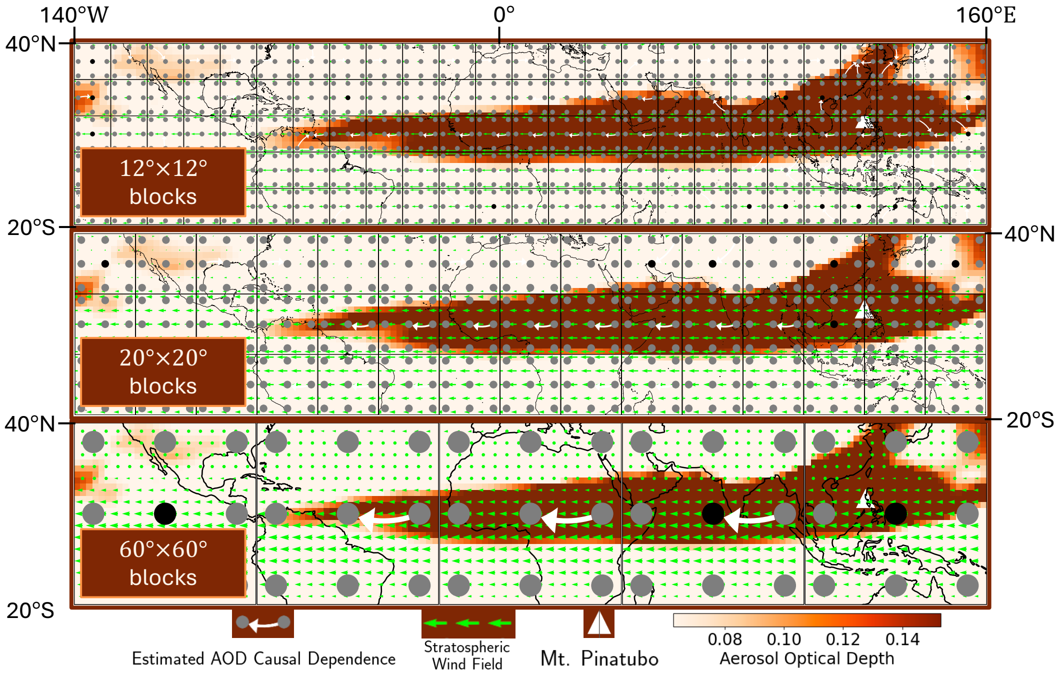
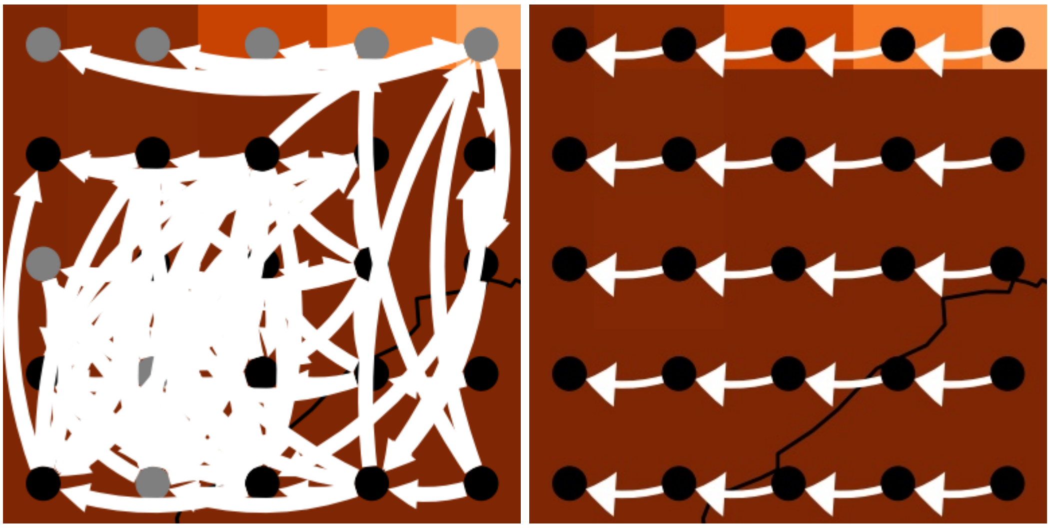


Figure H1: Results of CaStLe applied to HSW-V 21 days after the Mt. Pinatubo eruption with three different block sizes, 12° × 12°, 20° × 20°, and 60° × 60°. We find that results are generally consistent over the same area for each block size, with smaller block sizes allowing for additional nuance in some areas. Note that the 20° × 20° block panel is similar to the results shown in Figure 3, but more longitudes were added to get a space factorable by more integers, such as 12, 20, and 60.



(a) PC algorithm results (b) CaStLe results

Figure H2: The PC algorithm and CaStLe applied to E3SMv2-SPA in the $15° \times 15°$ block between 15° to 30°N and 75° to 90°E. from the day of the eruption to 20 days later. PC struggles to estimate an interpretable and physically meaningful graph of the dependence structure in this area. In contrast, CaStLe is able to identify an interpretable dependence structure that represents the local dynamics within the space.

Appendix I Analysis of Assumption Violation Examples

Here, we evaluate the impacts of potential violations of CaStLe's assumptions in our study of E3SMv2-SPA from Section 5.2.

I.1 Time Resolution is Too Coarse (Assumption T1)

The dataset's time resolution can determine if the temporal locality assumption (T1) holds. If the time resolution is too coarse, the temporal causal structures may be marginalized out or unmeasured. Dependencies between neighboring grid cells may not be manifested in the sparse time sampling. Here, we explore how our study of E3SMv2-SPA from Section 5.2 changes after coarsening the temporal resolution.

We coarsened the time resolution by two, from a daily to a two-daily resolution.

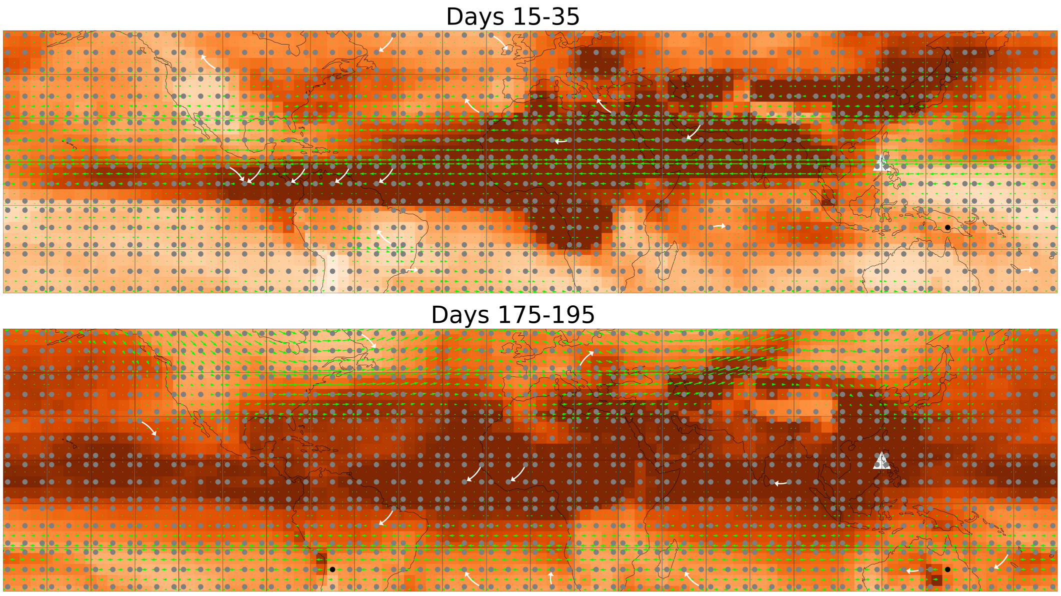


Figure I1: Results of using a coarsened temporal resolution (two-daily) in the E3SMv2-SPA study. CaStLe finds many fewer links in this setting. It is clear that when time is too coarse, causal structures fail to be detected. However, the remaining links that are found are largely true positives, suggesting that CaStLe is relatively robust to coarser time sampling.

Figure I1 demonstrates that CaStLe finds much fewer links when the time resolution is too coarse. However, the links that are detected are mostly consistent with known advective processes.

I.2 Time Interval is Too Long (Assumption T2)

When the time interval is too long, there may be too many causal structures in the data. This violates temporal causal stationarity (T2). Here, we investigate such a scenario.

We first computed causal stencils for an extended period, between day 15, the day of the eruption, to day 65. This is 30 days longer than our initial analysis from the start of the eruption.

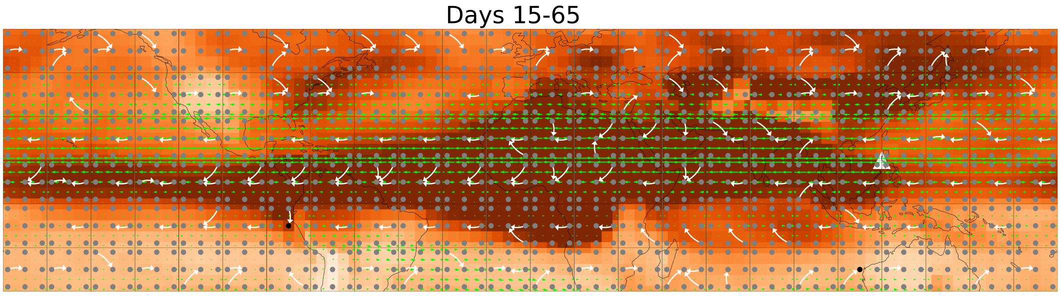


Figure I2: Results of applying CaStLe to a longer time interval from day 15 to 65. CaStLe identifies more links, indicating it is learning too many causal structures in the data, but still finds many of the true positives we found in our initial study. This indicates that many of the blocks in this interval have temporal causal stationarity, leading CaStLe to perform adequately.

We then computed causal stencils for the entire period between day 15 to day 215, roughly six months later.

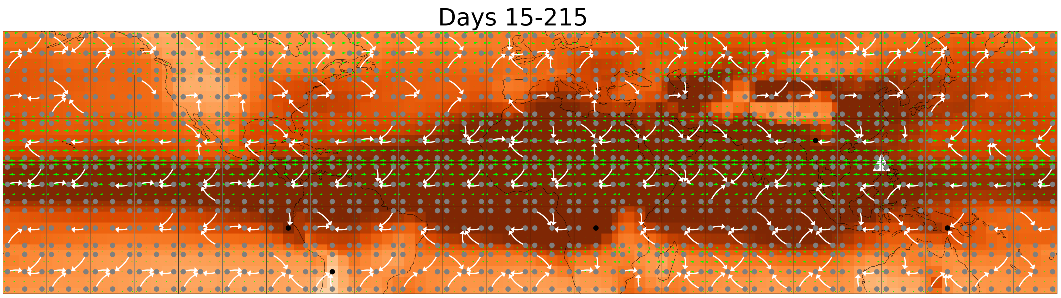


Figure I3: Results of applying CaStLe to a time interval that is too long and contains too many causal structures, day 15 to 200. We see that CaStLe identifies many links in each block. Comparing them to the winds is ineffective because the wind arrows are averages over the whole period rather than reflections of how they change in time, which CaStLe is learning from. With such a density of links, it is further challenging to know which are correct and which are spurious.

Figure I2 shows that when the time interval is longer, CaStLe identifies more links, indicating it is learning too many causal structures in the data, but still finds many of the true positives we found in our initial study. Figure I3 demonstrates the challenges of applying CaStLe to a time interval that contains too many difference causal structures. CaStLe identifies many links, creating uninterpretable stencils. The winds are a poor comparison because each arrow is a temporal average for that location, which is not representative over the entire interval. CaStLe may be capturing many spurious links or capturing all of the many fluctuating dynamics over the interval. Resulting is are uninterpretable stencils with unknown true and false positives. However, there are some blocks in the equatorial regions with sparse stencils. That indicates that dynamics were relatively stationary over the period.

I.3 Grid Resolution is Too Coarse (Assumption S1)

An appropriate grid resolution is important for satisfying the spatial locality assumption (S1). If the grid is too coarse then the underlying spatial structure may be marginalized out or unmeasured. If it is too small, causal relationships may appear outside the stencil neighborhood, requiring a radius-2 neighborhood implementation. Here, we investigate a grid resolution that is too coarse.

We coarsened the grid to 9°, rather than the 3° used in Section 5.2. Given that, to maintain 5×5 grid cells per block, each block is again $45° \times 45°$.

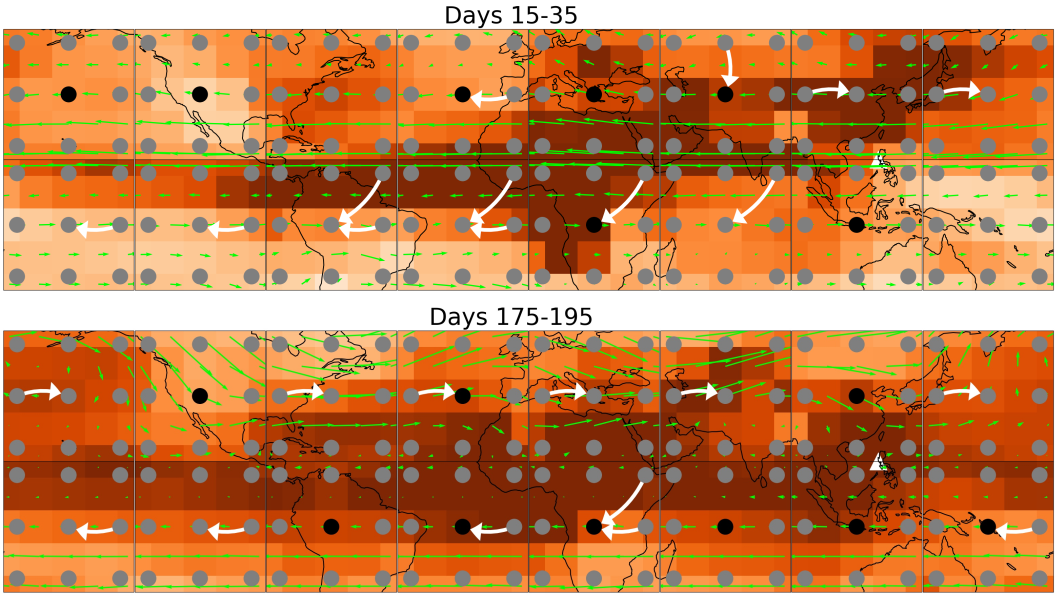


Figure I4: Results of using a coarse grid (9°) in the E3SMv2-SPA study. We find that CaStLe performs very well overall. There are few false positives and it clearly captures the overall advection dynamics of the system.

In Figure I4, we see that CaStLe performs very well overall. There are few false positives and it clearly captures the overall advection dynamics of the system.

We also coarsened the grid to 18°, resulting in $90° \times 90°$ blocks. In Figure I5, CaStLe performs well in the early time interval, clearly identifying the east-to-west advection pattern. However, in the later time interval, it finds no spatial structures apart from autodependencies in each block. This is likely because the east-to-west advection is weaker in this period and the grid is too coarse to capture the narrower bands of northward advection that dominates the interval.

We find that CaStLe is very robust to this assumption violation. It captures all of the most dominant advection patterns, while struggling to find smaller, weaker ones.

I.4 Block Sizes are Too Large (Assumption S2)

In Appendix H, we found that CaStLe's output was robust to very large and very small block sizes. Spatial blocks are intended to isolate regions such that only one underlying spatial causal structure exists in the block. If the blocks are too large, then Assumption S2 may be violated.

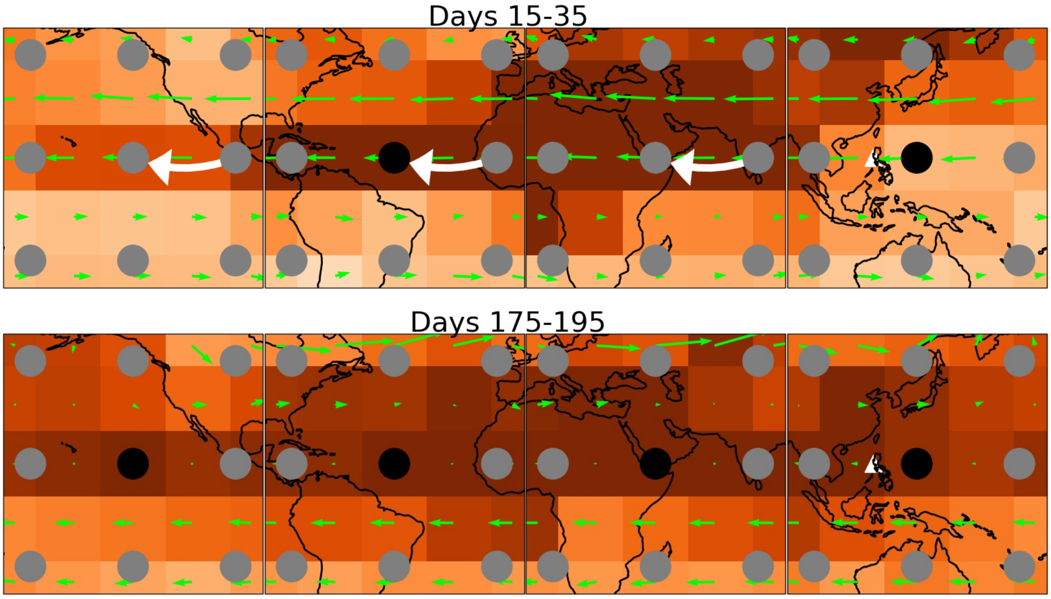


Figure I5: Results of using a coarse grid (18°) in the E3SMv2-SPA study. CaStLe performs well in the early time interval, clearly identifying the east-to-west advection pattern. However, in the later time interval, it finds no spatial structures apart from autodependencies in each block. This is likely because the east-to-west advection is weaker in this period and the grid is too coarse to capture the narrower bands of northward advection that dominates the interval.

In Figure I6, we used block sizes equal to $45° \times 45°$. Here, each block has 15×15 grid cells. This is in contrast to the 5×5 grid cell, $15° \times 15°$ blocks used in Section 5.2.

We find that while true positives remain, several false positives appear. Some positives may be the result of identifying multiple causal structures correctly within the space, while others may be confused results found because of the high density of links. In further testing with intermediate block sizes, we found that CaStLe is moderately robust to this assumption violation. As block sizes approach a more appropriate size, false positives diminish and true positives remain.

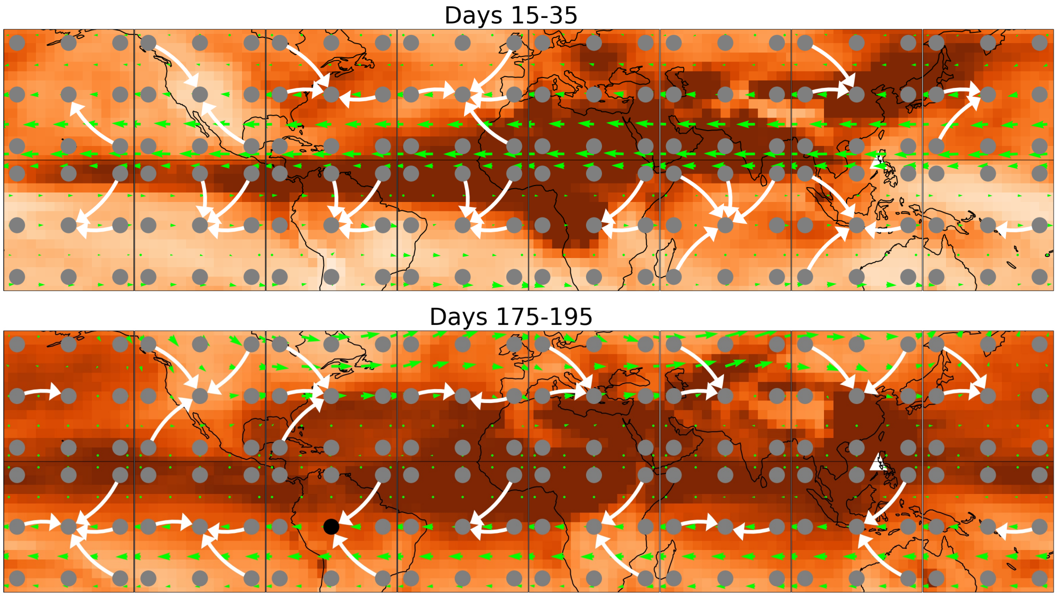


Figure I6: Results of using block sizes too large in the E3SMv2-SPA study. We see that many true positives are found, but many false positives as well. CaStLe seems to identify multiple contradictory causal structures within many cells, which may lead to more spurious links discovered. Even where links appear correct, they are largely uninterpretable in the presence of contradictions.

Appendix J Additional GCM Results

Figure J1 depicts results of implementing CaStLe with the Bayesian score optimization causal discovery algorithm, DYNOTEARS. We also presented results of DYNOTEARS applied to our VAR benchmark in Section 6.1. Here, we show that CaStLe-DYNOTEARS is able to recover comparable results to the CaStLe-PC-stable results shown in Section 5.1.

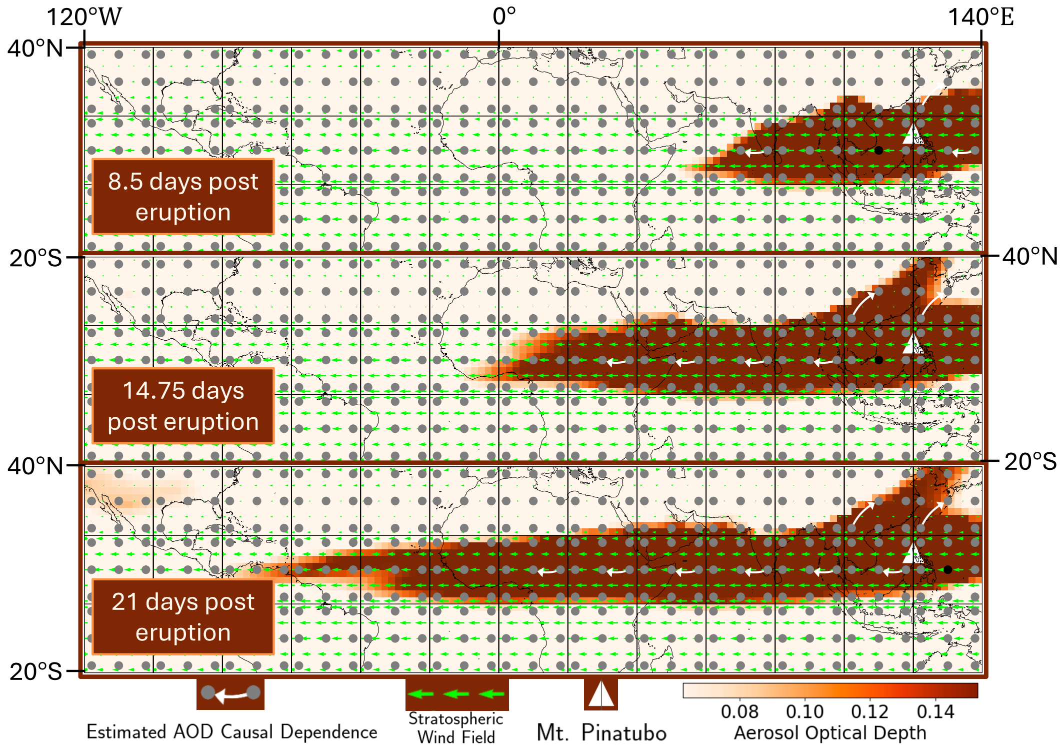


Figure J1: Application of CaStLe-DYNOTEARS to HSW-V simulation of the 1991 Mt. Pinatubo eruption. The stencils estimated by CaStLe (white) capture the underlying high-altitude wind fields (green) using only satellite-measured AOD, with near perfect accuracy in high aerosol regions (red-orange). On longer horizons (bottom row), CaStLe is able to recover equatorial wind currents as far away as South America, half-way around the world from Mt. Pinatubo (white triangle). CaStLe accurately identifies the prevailing westerly atmospheric winds because it was able to identify the space-time dependence between neighboring grid cells.

Appendix K Additional VAR Results

In Section 6.1, we demonstrated the strong performance of CaStLe on VAR-generated space-time data with fixed sparsity level $d = 4$; in particular, CaStLed variants uniformly improve over the performance of equivalent unstructured causal discovery algorithms. We repeat this analysis for a variety of sparsity levels in Figures K1 and K2 for the MCC and F_1 score similarity metrics, respectively. As in Figure 6, the CaStLed variants continue to significantly outperform across all sparsity levels, d; furthermore, as noted above, we observe that CaStLe can correctly estimate the underlying grid even on as few as $T = 10$ time samples when a sufficiently large grid is observed; non-CaStLe methods struggle on larger grid sizes, consistent with our analyses in the previous section. A time limit of 48 hours of wall-clock time was applied for each individual graph estimation: performance properties of methods that did not terminate during this window are not shown (*e.g.*, DYNOTEARS with $d = 6; T = 10; N = 10$).

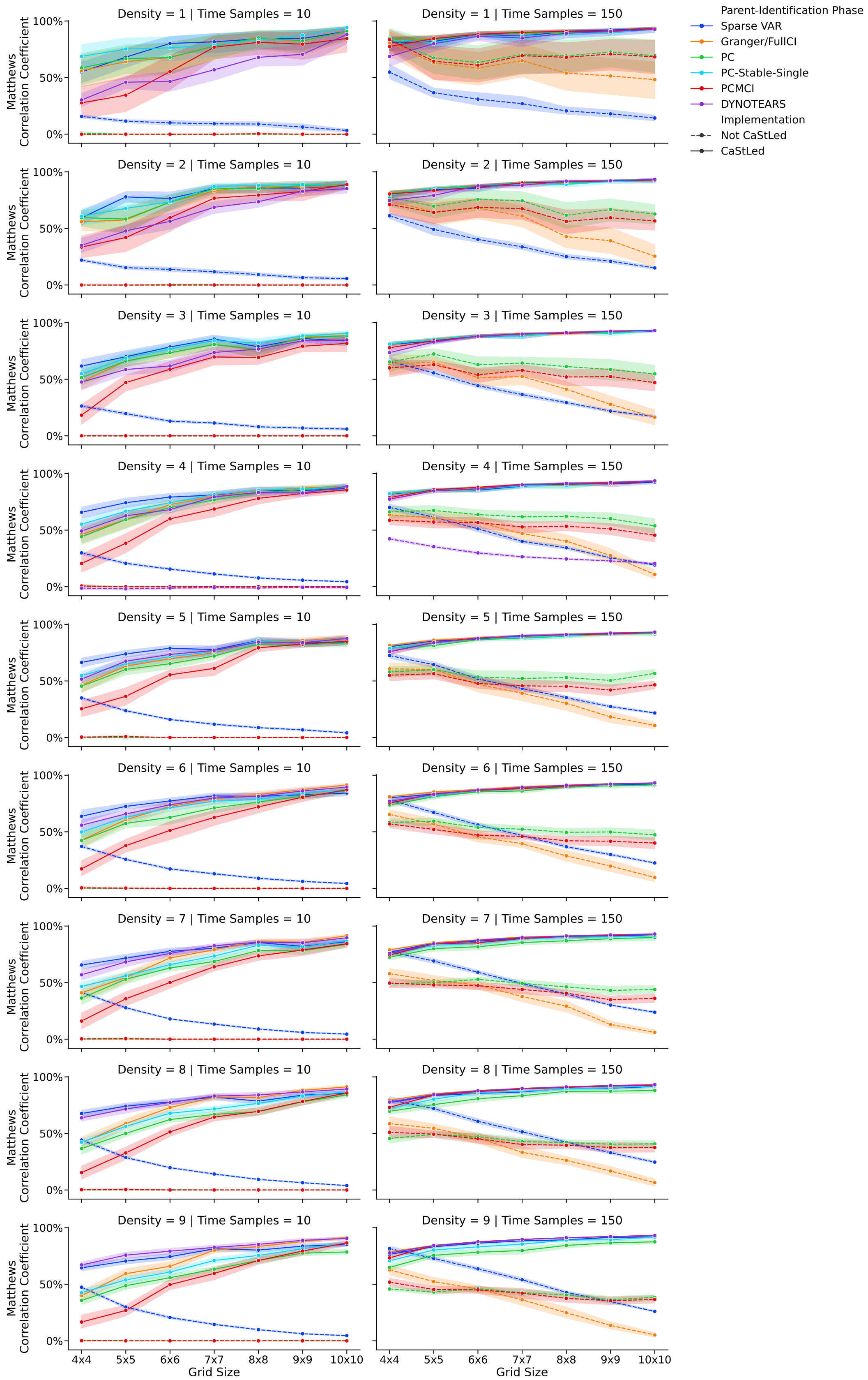


Figure K1: Matthews correlation coefficient (MCC) comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.1 for experimental details.

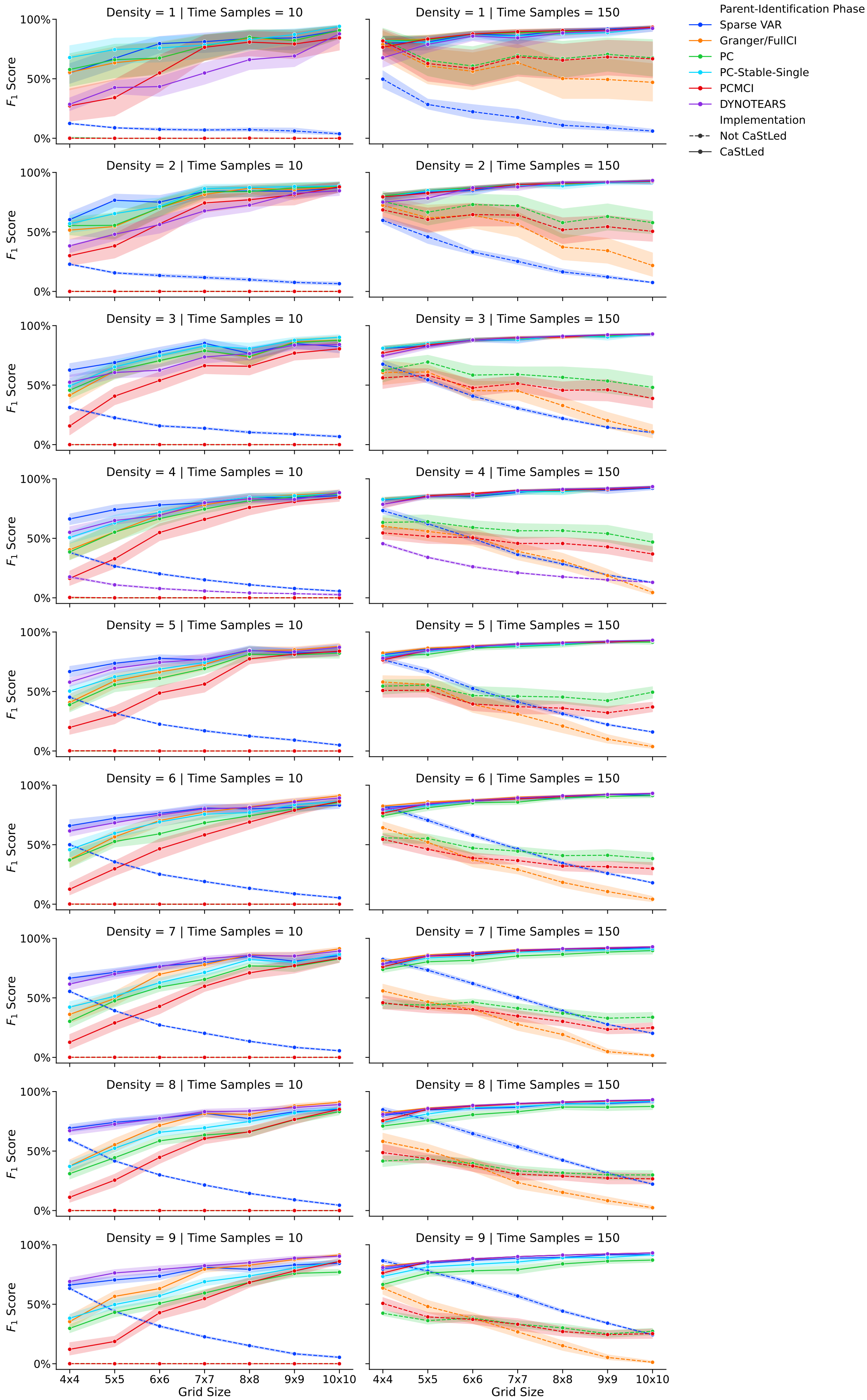


Figure K2: F_1 score comparison between CaStLed and non-CaStLed causal discovery approaches on 2D VAR dynamics for each sparsity level, including Granger causality (orange), PC (green), PC-Stable-Single (cyan), PCMCI (red), DYNOTEARS (purple), and a statistical model of the data generating process (blue). See Section 6.1 for experimental details.

Appendix L PC-Stable-Single

For the convenience of the reader, we include pseudo-code for the PC-Stable-Single algorithm of Runge, Nowack, et al. (2019), itself an adaptation of the PC-Stable algorithm of Colombo and Maathuis (2014). We use this as the PIP used for the CaStLe-based analyses shown in Sections 5.1.1, 5.2, and Appendix D. As our experiments in the proceeding section show, PC-Stable-Single exhibits small, but consistent improvements over alternative PIP choices.

Algorithm 2 PC-stable-single

Precondition: Time series dataset $\mathbf{X} = \{X^1, X^2, ..., X^N\}$, selected variable X^j, maximum time lag τ_{max} (default $\tau_{max} = 1$), significance threshold α_{PC}, maximum condition dimension p_{max} (default $p_{max} = N_{\tau_{max}}$), maximum number of combinations q_{max} (default $q_{max} = 1$), conditional independence test function I.

function CI$(X, Y, \mathbf{Z})$
    Test $X \perp\!\!\!\perp Y | \mathbf{Z}$ using test statistic measure I
    return p-value, test statistic value I
Initialize set of parents $\widehat{\mathcal{P}}(X_t^j) = \{X_{t-\tau}^i : i \in \{1, ..., N\}, \tau \in \{1, ..., \tau_{max}\}\}$
Initialize dictionary of test statistic values $I^{min}(X_{t-\tau}^i \to X_t^i) = \infty \ \forall X_{t-\tau}^i \in \widehat{\mathcal{P}}(X_t^j)$
for $p = 0, ..., p_{max}$ **do**
    if $|\widehat{\mathcal{P}}(X_t^j)| - 1 < p$ **then**
        Break for-loop ▷ Algorithm has converged
    for all $X_{t-\tau}^i$ in $\widehat{\mathcal{P}}(X_t^j)$ **do**
        $q = -1$
        for all lexicographically chosen subsets $\mathcal{S} \subseteq \widehat{\mathcal{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}$, with $|\mathcal{S}| = p$ **do**
            $q = q + 1$
            if $q >= q_{max}$ **then**
                Break from inner for-loop
            Run CI test to obtain (p-value, I) $\leftarrow CI(X_{t-\tau}^i, X_t^i, \mathcal{S})$
            if $|I| < I^{min}(X_{t-\tau}^i \to X_t^i)$ **then** ▷ Store min. I of parent among all tests
                $I^{min}(X_{t-\tau}^i \to X_t^i) = I$
            if p-value $> \alpha_{PC}$ **then** ▷ Removed only after all $X_{t-\tau}^i$ have been tested
                Mark $X_{t-\tau}^i$ for removal from $\widehat{\mathcal{P}}(X_t^i)$
                Break from inner loop
    Remove non-significant parents from $\widehat{\mathcal{P}}(X_t^i)$
    Sort parents in $\widehat{\mathcal{P}}(X_t^i)$ by $I^{min}(X_{t-\tau}^i \to X_t^i)$ from largest to smallest
return $\widehat{\mathcal{P}}(X_t^i)$

Open Research Section

The data generated and used for our HSW-V, VAR, and PDE experiments in Sections 5.1, 6.1, and Appendix D are available on Zenodo via `https://doi.org/10.5281/zenodo.12701546` with GNU Lesser General Public License v3.0 or later (J. Nichol, 2024). The data used for the E3SMv2-SPA experiments in Section 5.2 can be found in Brown et al. (2024). The code for generating data, running experiments, and generating figures has been archived in J. J. Nichol (2025). Future versions of CaStLe may be found at `https://github.com/jjakenichol/CaStLe`.

Acknowledgments

We thank Kara Peterson, the Deputy Principal Investigator of the CLDERA (CLimate impact: Determining Etiology thRough pAthways) project at Sandia National Laboratories (SNL), for helping to make this work possible. We also thank Joey Hart at SNL for helping with 2D Burgers' equation modeling and Tom Ehrmann at SNL for his help in understanding the atmospheric dynamics we sought to capture. We thank everyone on CLDERA's simulation team, especially Benj Wagman, Hunter Brown, and Joe Hollowed, for developing the E3SMv2-SPA and HSW-V models, preparing the data, and providing their expertise. Finally, we thank the reviewers who devoted their time to helping us significantly improve the communication of this work.

This work was supported by the Laboratory Directed Research and Development program at Sandia National Laboratories, a multi-mission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC (NTESS), a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration (DOE/NNSA) under contract DE-NA0003525. This written work is authored by employees of NTESS. The employees, not NTESS, own the right, title, and interest in and to the written work and is responsible for its contents. Any subjective views or opinions that might be expressed in the written work do not necessarily represent the views of the U.S. Government. The publisher acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this written work or allow others to do so, for U.S. Government purposes. The DOE will provide public access to results of federally sponsored research in accordance with the DOE Public Access Plan.

This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government.

References

Agarwal, A., Caesar, L., Marwan, N., Maheswaran, R., Merz, B., & Kurths, J. (2019). Network-based identification and characterization of teleconnections on different scales. *Scientific Reports*, *9*(1), 8808. doi: 10.1038/s41598-019-45423-5

Ali, S., Hasan, U., Li, X., Faruque, O., Sampath, A., Huang, Y., … Wang, J. (2024). Causality for Earth Science – A Review on Time-series and Spatiotemporal Causality Methods. *arXiv*. doi: 10.48550/arxiv.2404.05746

Allen, G. I., Gan, L., & Zheng, L. (2023). Interpretable Machine Learning for Discovery: Statistical Challenges and Opportunities. *Annual Review of Statistics and Its Application*, *11*(1), 97–121. doi: 10.1146/annurev-statistics-040120-030919

Aquila, V., Garfinkel, C. I., Newman, P., Oman, L., & Waugh, D. (2014). Modifications of the quasi-biennial oscillation by a geoengineering perturbation of the stratospheric aerosol layer. *Geophysical Research Letters*, *41*(5), 1738–1744. doi: 10.1002/2013gl058818

Baranowski, K., Faust, C., Eby, P., & Bharti, N. (2021). Quantifying the impacts of Australian bushfires on native forests and gray-headed flying foxes. *Global Ecology and Conservation*, *27*, e01566. doi: 10.1016/j.gecco.2021.e01566

Baño-Medina, J., Sengupta, A., Doyle, J. D., Reynolds, C. A., Watson-Parris, D., & Monache, L. D. (2025). Are AI weather models learning atmospheric physics? A sensitivity analysis of cyclone Xynthia. *npj Climate and Atmospheric Science*, *8*(1), 92. doi: 10.1038/s41612-025-00949-6

Bellman, R. E. (1957). *Dynamic programming.* Princeton, NJ: Princeton University Press. (Introduced the term “curse of dimensionality” in the preface: “All this may be subsumed under the heading ’the curse of dimensionality.’ Since this is a curse which has hung over the head of the physicist and astronomer for many a year...”)

Bhattacharjee, K., Naskar, N., Roy, S., & Das, S. (2020). A survey of cellular automata: types, dynamics, non-uniformity and applications. *Natural Computing*, *19*(2), 433–461. doi: 10.1007/s11047-018-9696-8

Bonkile, M. P., Awasthi, A., Lakshmi, C., Mukundan, V., & Aswin, V. S. (2018). A systematic literature review of Burgers’ equation with recent advances. *Pramana*, *90*(6), 69. doi: 10.1007/s12043-018-1559-4

Boussard, J., Nagda, C., Kaltenborn, J., Lange, C. E. E., Brouillard, P., Gurwicz, Y., … Rolnick, D. (2023). Towards Causal Representations of Climate Model Data. *arXiv*. doi: 10.48550/arxiv.2312.02858

Brouillard, P., Lachapelle, S., Kaltenborn, J., Gurwicz, Y., Sridhar, D., Drouin, A., … Rolnick, D. (2024). Causal Representation Learning in Temporal Data via Single-Parent Decoding. *arXiv*. doi: 10.48550/arxiv.2410.07013

Brown, H. Y., Wagman, B., Bull, D., Peterson, K., Hillman, B., Liu, X., … Lin, L. (2024). Validating a microphysical prognostic stratospheric aerosol implementation in E3SMv2 using observations after the Mount Pinatubo eruption. *Geoscientific Model Development*, *17*(13), 5087–5121. doi: 10.5194/gmd-17-5087-2024

Brunton, S. L., Proctor, J. L., & Kutz, J. N. (2016). Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(15), 3932-3937. doi: 10.1073/pnas.1517384113

Burgers, J. (1948). A Mathematical Model Illustrating the Theory of Turbulence. *Advances in Applied Mechanics*, *1*, 171–199. doi: 10.1016/s0065-2156(08)70100-5

Bühlmann, P., & Geer, S. v. d. (2011). Statistics for High-Dimensional Data, Methods, Theory and Applications. *Springer Series in Statistics*, 99–182. doi: 10.1007/978-3-642-20192-9_6

Capua, G. D., Kretschmer, M., Donner, R. V., Hurk, B. v. d., Vellore, R., Krishnan, R., & Coumou, D. (2019). Tropical and mid-latitude teleconnections interacting with the Indian summer monsoon rainfall: a theory-guided causal effect network approach. *Earth System Dynamics*, *11* (1), 17–34. doi: 10.5194/esd-11-17-2020

Capua, G. D., Runge, J., Donner, R. V., Hurk, B. v. d., Turner, A. G., Vellore, R., ... Coumou, D. (2020). Dominant patterns of interaction between the tropics and mid-latitudes in boreal summer: causal relationships and the role of timescales. *Weather and Climate Dynamics*, *1* (2), 519–539. doi: 10.5194/wcd-1-519-2020

Colombo, D., & Maathuis, M. H. (2014). Order-independent constraint-based causal structure learning. *Journal of Machine Learning Research*, *15* (1), 3741–3782.

Davis, W. L., Carlson, M. L., Tezaur, I. K., Bull, D. L., Peterson, K. J., & Swiler, L. P. (2025). Spatio-temporal multivariate cluster evolution analysis for detecting and tracking climate impacts. *Journal of Computational and Applied Mathematics*, *465*, 116583. doi: 10.1016/j.cam.2025.116583

Deng, Y., & Ebert-Uphoff, I. (2014). Weakening of atmospheric information flow in a warming climate in the Community Climate System Model. *Geophysical Research Letters*, *41* (1), 193–200. doi: 10.1002/2013gl058646

Diffenbaugh, N. S., Pal, J. S., Trapp, R. J., & Giorgi, F. (2005). Fine-scale processes regulate the response of extreme events to global climate change. *Proceedings of the National Academy of Sciences*, *102* (44), 15774–15778. doi: 10.1073/pnas.0506042102

Driscoll, D. A., Macdonald, K. J., Gibson, R. K., Doherty, T. S., Nimmo, D. G., Nolan, R. H., ... Phillips, R. D. (2024). Biodiversity impacts of the 2019–2020 Australian megafires. *Nature*, *635* (8040), 898–905. doi: 10.1038/s41586-024-08174-6

Dutton, E. G., & Christy, J. R. (1992). Solar radiative forcing at selected locations and evidence for global lower tropospheric cooling following the eruptions of El Chichón and Pinatubo. *Geophysical Research Letters*, *19* (23), 2313–2316. doi: 10.1029/92gl02495

Ebert-Uphoff, I., & Deng, Y. (2014). Causal Discovery from Spatio-Temporal Data with Applications to Climate Science. *2014 13th International Conference on Machine Learning and Applications*, 606-613. doi: 10.1109/icmla.2014.96

Ebert-Uphoff, I., & Deng, Y. (2012). A new type of climate network based on probabilistic graphical models: Results of boreal winter versus summer. *Geophysical Research Letters*, *39* (19). doi: 10.1029/2012gl053269

Fountalis, I., Dovrolis, C., Bracco, A., Dilkina, B., & Keilholz, S. (2018). δ-MAPS: from spatio-temporal data to a weighted and lagged network between functional domains. *Applied Network Science*, *3* (1), 21. doi: 10.1007/s41109-018-0078-z

Galytska, E., Weigel, K., Handorf, D., Jaiser, R., Köhler, R. H., Runge, J., & Eyring, V. (2022). Causal model evaluation of Arctic-midlatitude teleconnections in CMIP6. *Journal of Geophysical Research: Atmospheres*, *128* (17). doi: 10.1002/essoar.10512569.1

Glymour, C., & Scheines, R. (1986). Causal modeling with the TETRAD program. *Synthese*, *68* (1), 37–63. doi: 10.1007/bf00413966

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, *10*, 524. doi: 10.3389/fgene.2019.00524

Goerg, G., & Shalizi, C. (2013, 29 Apr–01 May). Mixed licors: A nonparametric algorithm for predictive state reconstruction. In C. M. Carvalho & P. Ravikumar (Eds.), *Proceedings of the sixteenth international conference on artificial intelligence and statistics* (Vol. 31, pp. 289–297). Scottsdale, Arizona, USA: PMLR. Retrieved from `https://proceedings.mlr.press/v31/goerg13a.html`

Golaz, J., Roekel, L. P. V., Zheng, X., Roberts, A. F., Wolfe, J. D., Lin, W., ... Bader, D. C. (2022). The DOE E3SM Model Version 2: Overview of the Physical Model and Initial Model Evaluation. *Journal of Advances in Modeling Earth Systems*, *14*(12). doi: 10.1029/2022ms003156

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica*, *37*(3), 424. (Granger Causality seminal paper) doi: 10.2307/1912791

Gray, L. J., Anstey, J. A., Kawatani, Y., Lu, H., Osprey, S., & Schenzinger, V. (2018). Surface impacts of the Quasi Biennial Oscillation. *Atmospheric Chemistry and Physics*, *18*(11), 8227–8247. doi: 10.5194/acp-18-8227-2018

Guo, S., Bluth, G. J. S., Rose, W. I., Watson, I. M., & Prata, A. J. (2004). Re-evaluation of SO2 release of the 15 June 1991 Pinatubo eruption using ultraviolet and infrared satellite sensors. *Geochemistry, Geophysics, Geosystems*, *5*(4). doi: 10.1029/2003gc000654

Guo, S., Rose, W. I., Bluth, G. J. S., & Watson, I. M. (2004). Particles in the great Pinatubo volcanic cloud of June 1991: The role of ice. *Geochemistry, Geophysics, Geosystems*, *5*(5). doi: 10.1029/2003gc000655

Hart, J., Gulian, M., Manickam, I., & Swiler, L. P. (2023). Solving High-Dimensional Inverse Problems with Auxiliary Uncertainty via Operator Learning with Limited Data. *Journal of Machine Learning for Modeling and Computing*, *4*(2), 105–133. doi: 10.1615/jmachlearnmodelcomput.2023048105

Higgins, T. B., Subramanian, A. C., Watson, P. A. G., & Sparrow, S. (2025). Changes to Atmospheric River Related Extremes Over the United States West Coast Under Anthropogenic Warming. *Geophysical Research Letters*, *52*(5). doi: 10.1029/2024gl112237

Hitchman, M. H., McKay, M., & Trepte, C. R. (1994). A climatology of stratospheric aerosol. *Journal of Geophysical Research: Atmospheres*, *99*(D10), 20689–20700. Retrieved from `https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/94JD01525` doi: 10.1029/94jd01525

Hollowed, J. P., Jablonowski, C., Brown, H. Y., Hillman, B. R., Bull, D. L., & Hart, J. L. (2024). Localized injections of interactive volcanic aerosols and their climate impacts in a simple general circulation model. *EGUsphere*, *2024*, 1–38. doi: 10.5194/egusphere-2024-335

Jones, D. B. A., Schneider, H. R., & McElroy, M. B. (1998). Effects of the quasi-biennial oscillation on the zonally averaged transport of tracers. *Journal of Geophysical Research: Atmospheres*, *103*(D10), 11235–11249. doi: 10.1029/98jd00682

Kalisch, M., & Bühlmann, P. (2007). Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *Journal of Machine Learning Research*, *8*, 613-636. Retrieved from `https://www.jmlr.org/papers/v8/kalisch07a.html`

Kamiński, M., Ding, M., Truccolo, W. A., & Bressler, S. L. (2001). Evaluating causal relations in neural systems: Granger causality, directed transfer function and statistical assessment of significance. *Biological Cybernetics*, *85*(2), 145–157. doi: 10.1007/s004220000235

Keellings, D., & Moradkhani, H. (2020). Spatiotemporal Evolution of Heat Wave Severity and Coverage Across the United States. *Geophysical Research Letters*, *47*(9). doi: 10.1029/2020gl087097

Kremser, S., Thomason, L. W., Hobe, M. v., Hermann, M., Deshler, T., Timmreck, C., ... Meland, B. (2016). Stratospheric aerosol—Observations, processes, and impact on climate. *Reviews of Geophysics*, *54*(2), 278–335. doi: 10.1002/2015rg000511

Krich, C., Runge, J., Miralles, D. G., Migliavacca, M., Perez-Priego, O., El-Madany, T., ... Mahecha, M. D. (2020). Estimating causal networks in biosphere–atmosphere interaction with the PCMCI approach. *Biogeosciences*, *17*(4), 1033–1061. doi: 10.5194/bg-17-1033-2020

Labitzke, K., & McCormick, M. P. (1992). Stratospheric temperature increases due to Pinatubo aerosols. *Geophysical Research Letters*, *19*(2), 207–210. doi: 10.1029/91gl02940

Li, Z., Kovachki, N., Azizzadenesheli, K., Liu, B., Bhattacharya, K., Stuart, A., & Anandkumar, A. (2020). Fourier Neural Operator for Parametric Partial Differential Equations. *arXiv*. doi: 10.48550/arxiv.2010.08895

Liu, Y., Niculescu-Mizil, A., Lozano, A., & Lu, Y. (2010). Learning Temporal Causal Graphs for Relational Time-Series Analysis. In *Proceedings of the 27th international conference on international conference on machine learning* (p. 687–694). Madison, WI, USA: Omnipress.

Marshall, L. R., Maters, E. C., Schmidt, A., Timmreck, C., Robock, A., & Toohey, M. (2022). Volcanic effects on climate: recent advances and future avenues. *Bulletin of Volcanology*, *84*(5), 54.

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, *405*(2), 442–451. Retrieved from https://www.sciencedirect.com/science/article/pii/0005279575901099 doi: 10.1016/0005-2795(75)90109-9

Miller, H. J. (2004). Tobler's First Law and Spatial Analysis. *Annals of the Association of American Geographers*, *94*(2), 284–289. doi: 10.1111/j.1467-8306.2004.09402005.x

Neto, E. C., Keller, M. P., Attie, A. D., & Yandell, B. S. (2010). Causal graphical models in systems genetics: A unified framework for joint inference of causal network and genetic architecture for correlated phenotypes. *The Annals of Applied Statistics*, *4*(1), 320–339. doi: 10.1214/09-aoas288

Nichol, J. (2024, July). *CaStLe Data Release for JGR MLC 2024.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.12701546 doi: 10.5281/zenodo.12701546

Nichol, J. J. (2025, May). *jjakenichol/castle: v0.1.0 - jgr-mlc publication release.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.15530557 doi: 10.5281/zenodo.15530557

Nichol, J. J., Peterson, M. G., Peterson, K. J., Fricke, G. M., & Moses, M. E. (2021, 10). Machine learning feature analysis illuminates disparity between E3SM climate models and observed climate change. *Journal of Computational and Applied Mathematics*, *395*, 113451. doi: 10.1016/j.cam.2021.113451

Nichol, J. J., Weylandt, M., Smith, M., & Swiler, L. (2023). *Benchmarking the PCMCI Causal Discovery Algorithm for Spatiotemporal Systems* (Tech. Rep.). Sandia National Laboratories. Retrieved from https://www.osti.gov/biblio/1991387

Nowack, P., Runge, J., Eyring, V., & Haigh, J. D. (2020). Causal networks for climate model evaluation and constrained projections. *Nature Communications 2020 11:1*, *11*(1), 1—11. Retrieved from http://www.nature.com/articles/s41467-020-15195-y doi: 10.1038/s41467-020-15195-y

Nukavarapu, N., Yang, J.-A., & Jankowska, M. M. (2023). Unsupervised Deep Learning Approach to Analyze Spatio-Temporal Change in Satellite Imagery. *IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, *00*, 2496–2499. doi: 10.1109/igarss52108.2023.10282519

O'Kane, T. J., Harries, D., & Collier, M. A. (2024). Bayesian Structure Learning for Climate Model Evaluation. *Journal of Advances in Modeling Earth Systems*, *16*(5). doi: 10.1029/2023ms004034

Palu, M. (2019). Coupling in complex systems as information transfer across time scales. *Philosophical Transactions of the Royal Society A*, *377*(2160), 20190094. doi: 10.1098/rsta.2019.0094

Pamfil, R., Sriwattanaworachai, N., Desai, S., Pilgerstorfer, P., Georgatzis, K., Beaumont, P., & Aragam, B. (2020). DYNOTEARS: Structure Learning from

Time-Series Data. *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, *108*, 1595–1605. Retrieved from https://proceedings.mlr.press/v108/pamfil20a.html

Parker, D. E., Wilson, H., Jones, P. D., Christy, J. R., & FOLLAND, C. K. (1996). The impact of Mount Pinatubo on world-wide temperatures. *International Journal of Climatology*, *16*(5), 487–497. doi: 10.1002/(sici)1097-0088(199605)16:5<487::aid-joc39>3.0.co;2-j

Parker, D. E., Wilson, H., Jones, P. D., Christy, J. R., & Folland, C. K. (1996). The impact of mount pinatubo on world-wide temperatures. *International Journal of Climatology*, *16*(5), 487-497. Retrieved from https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0088%28199605%2916%3A5%3C487%3A%3AAID-JOC39%3E3.0.CO%3B2-J doi: https://doi.org/10.1002/(SICI)1097-0088(199605)16:5<487::AID-JOC39>3.0.CO;2-J

Pathak, J., Subramanian, S., Harrington, P., Raja, S., Chattopadhyay, A., Mardani, M., . . . Anandkumar, A. (2022). FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. *arXiv*. doi: 10.48550/arxiv.2202.11214

Payne, A. E., Demory, M.-E., Leung, L. R., Ramos, A. M., Shields, C. A., Rutz, J. J., . . . Ralph, F. M. (2020). Responses and impacts of atmospheric rivers to climate change. *Nature Reviews Earth & Environment*, *1*(3), 143–157. doi: 10.1038/s43017-020-0030-5

Pearl, J. (1995). Causal Diagrams for Empirical Research. *Biometrika*, *82*(4), 669. doi: 10.2307/2337329

Pearl, J. (1998). Graphs, Causality, and Structural Equation Models. *Sociological Methods & Research*, *27*(2), 226–284. Retrieved from https://doi.org/10.1177/0049124198027002004 doi: 10.1177/0049124198027002004

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference.* Cambridge University Press. Retrieved from https://books.google.com/books?id=wnGU_TsW3BQC

Pearl, J., Glymour, M., & Jewell, N. (2016). *Causal Inference in Statistics: A Primer.* Wiley. Retrieved from https://books.google.com/books?id=L3G-CgAAQBAJ

Pearl, J., & Verma, T. S. (1992). A statistical semantics for causation. *Statistics and Computing*, *2*(2), 91–95. doi: 10.1007/bf01889587

Peters, J., Janzing, D., & Schlkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms.* Cambridge, Massachusetts: The MIT Press.

Pfleiderer, P., Schleussner, C.-F., Geiger, T., & Kretschmer, M. (2020). Robust predictors for seasonal Atlantic hurricane activity identified with causal effect networks. *Weather and Climate Dynamics*, *1*(2), 313–324. doi: 10.5194/wcd-1-313-2020

Polkova, I., Afargan-Gerstman, H., Domeisen, D. I. V., King, M. P., Ruggieri, P., Athanasiadis, P., . . . Baehr, J. (2021). Predictors and prediction skill for marine cold-air outbreaks over the Barents Sea. *Quarterly Journal of the Royal Meteorological Society*, *147*(738), 2638–2656. doi: 10.1002/qj.4038

Raghu, V. K., Ramsey, J. D., Morris, A., Manatakis, D. V., Sprites, P., Chrysanthis, P. K., . . . Benos, P. V. (2018). Comparison of strategies for scalable causal discovery of latent variable models from mixed data. *International Journal of Data Science and Analytics*, *6*(1), 33–45. doi: 10.1007/s41060-018-0104-3

Ramsey, J. D. (2014). A Scalable Conditional Independence Test for Nonlinear, Non-Gaussian Data. *arXiv*, *abs/1401.5031*. Retrieved from http://arxiv.org/abs/1401.5031 doi: 10.48550/arxiv.1401.5031

Reichenbach, H. (1956). *The Direction of Time* (M. Reichenbach, Ed.). Dover Publications.

Robock, A. (2000). Volcanic eruptions and climate. *Reviews of Geophysics*, *38*(2), 191-219. Retrieved from https://agupubs.onlinelibrary.wiley.com/doi/

abs/10.1029/1998RG000054 doi: https://doi.org/10.1029/1998RG000054

Rubenstein, P. K., Bongers, S., Schölkopf, B., & Mooij, J. M. (2018). From Deterministic ODEs to Dynamic Structural Causal Models. In *Uai'18: Proceedings of the twenty-ninth conference on uncertainty in artificial intelligence.* AUAI Press. Retrieved from http://auai.org/uai2018/proceedings/papers/43.pdf doi: 10.48550/arxiv.1608.08028

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701. doi: 10.1037/h0037350

Runge, J. (2018). Causal network reconstruction from time series: From theoretical assumptions to practical estimation. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *28*(7), 075310. doi: 10.1063/1.5025050

Runge, J., Bathiany, S., Bollt, E., Camps-Valls, G., Coumou, D., Deyle, E., ... Zscheischler, J. (2019). Inferring causation from time series in Earth system sciences. *Nature Communications*, *10*(2553). doi: 10.1038/s41467-019-10105-3

Runge, J., Gerhardus, A., Varando, G., Eyring, V., & Camps-Valls, G. (2023). Causal inference for time series. *Nature Reviews Earth & Environment*, *4*(7), 487–505. doi: 10.1038/s43017-023-00431-y

Runge, J., Nowack, P., Kretschmer, M., Flaxman, S., & Sejdinovic, D. (2019). Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, *5*(11), 4996—5023. Retrieved from http://advances.sciencemag.org/ doi: 10.1126/sciadv.aau4996

Runge, J., Petoukhov, V., Donges, J. F., Hlinka, J., Jajcay, N., Vejmelka, M., ... Kurths, J. (2015). Identifying causal gateways and mediators in complex spatio-temporal systems. *Nature Communications*, *6*(1), 8502. doi: 10.1038/ncomms9502

Saetia, S., Yoshimura, N., & Koike, Y. (2021). Constructing Brain Connectivity Model Using Causal Network Reconstruction Approach. *Frontiers in Neuroinformatics*, *15*, 619557. doi: 10.3389/fninf.2021.619557

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward Causal Representation Learning. *Proceedings of the IEEE*, *109*(5), 612–634. doi: 10.1109/jproc.2021.3058954

Sheth, P., Shah, R., Sabo, J., Candan, K. S., & Liu, H. (2022). STCD: A Spatio-Temporal Causal Discovery Framework for Hydrological Systems. *2022 IEEE International Conference on Big Data (Big Data)*, *00*, 5578–5583. doi: 10.1109/bigdata55660.2022.10020845

Shimizu, S., Hoyer, P. O., Hyvarinen, A., & Kerminen, A. (2006). A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, *7*(72), 2003-2030. Retrieved from https://www.jmlr.org/papers/volume7/shimizu06a/shimizu06a.pdf

Sjolte, J., Adolphi, F., Guðlaugsdòttir, H., & Muscheler, R. (2021). Major Differences in Regional Climate Impact Between High- and Low-Latitude Volcanic Eruptions. *Geophysical Research Letters*, *48*(8). doi: 10.1029/2020gl092017

Soden, B. J., Wetherald, R. T., Stenchikov, G. L., & Robock, A. (2002). Global Cooling After the Eruption of Mount Pinatubo: A Test of Climate Feedback by Water Vapor. *Science*, *296*(5568), 727–730. doi: 10.1126/science.296.5568.727

Spirtes, P., & Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, *9*(1), 62–72. doi: 10.1177/089443939100900106

Spirtes, P., Glymour, C., & Scheines, R. (1993). Causation, Prediction, and Search. *Lecture Notes in Statistics*. doi: 10.1007/978-1-4612-2748-9

Strang, G. (2016). *Introduction to Linear Algebra* (Fifth ed.). Wellesley, MA: Wellesley-Cambridge Press.

Sugihara, G., May, R., Ye, H., Hsieh, C.-h., Deyle, E., Fogarty, M., & Munch, S.

(2012). Detecting Causality in Complex Ecosystems. *Science*, *338*(6106), 496–500. Retrieved from https://www.science.org/doi/abs/10.1126/science.1227079 doi: 10.1126/science.1227079

Thomas, M. A., Giorgetta, M. A., Timmreck, C., Graf, H.-F., & Stenchikov, G. (2009). Simulation of the climate impact of Mt. Pinatubo eruption using ECHAM5 – Part 2: Sensitivity to the phase of the QBO and ENSO. *Atmospheric Chemistry and Physics*, *9*(9), 3001–3009. doi: 10.5194/acp-9-3001-2009

Tibau, X.-A., Reimers, C., Gerhardus, A., Denzler, J., Eyring, V., & Runge, J. (2022). A spatiotemporal stochastic climate model for benchmarking causal discovery methods for teleconnections. *Environmental Data Science*, *1*, e12. doi: 10.1017/eds.2022.11

Timmreck, C. (2012). Modeling the climatic effects of large explosive volcanic eruptions. *Wiley Interdisciplinary Reviews: Climate Change*, *3*(6), 545–564.

Trenberth, K. E., & Dai, A. (2007). Effects of Mount Pinatubo volcanic eruption on the hydrological cycle as an analog of geoengineering. *Geophysical Research Letters*, *34*(15). doi: 10.1029/2007GL030524

Tsonis, A. A., Deyle, E. R., Ye, H., & Sugihara, G. (2017). Convergent Cross Mapping: Theory and an Example. *Advances in Nonlinear Geosciences*, 587–600. doi: 10.1007/978-3-319-58895-7_27

Walker, R. T. (2022). GEOGRAPHY, VON THÜNEN, AND TOBLER'S FIRST LAW: TRACING THE EVOLUTION OF A CONCEPT. *Geographical Review*, *112*(4), 591–607. doi: 10.1080/00167428.2021.1906670

Weylandt, M., & Swiler, L. P. (2024). Beyond pca: Additional dimension reduction techniques to consider in the development of climate fingerprints. *Journal of Climate*, *To appear*. doi: 10.1175/JCLI-D-23-0267.1

Zhang, X., Zhao, X.-M., He, K., Lu, L., Cao, Y., Liu, J., ... Chen, L. (2011). Inferring gene regulatory networks from gene expression data by path consistency algorithm based on conditional mutual information. *Bioinformatics*, *28*(1), 98–104. doi: 10.1093/bioinformatics/btr626

Zhang, Z., Li, G., Cai, Y., Cheng, X., Sun, Y., Zhao, J., ... An, Z. (2022). Millennial-Scale Monsoon Variability Modulated by Low-Latitude Insolation During the Last Glaciation. *Geophysical Research Letters*, *49*(1). doi: 10.1029/2021gl096773

Zhao, H., Kitsios, V., O'Kane, T. J., & Bonilla, E. V. (2024). Bayesian Factorised Granger-Causal Graphs For Multivariate Time-series Data. *arXiv*. doi: 10.48550/arxiv.2402.03614

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). DAGs with NO TEARS: Continuous Optimization for Structure Learning. In *Proceedings of the 32nd international conference on neural information processing systems* (p. 9492–9503). Red Hook, NY, USA: Curran Associates Inc.

Zhu, J. Y., Zhang, C., Zhang, H., Zhi, S., Li, V. O., Han, J., & Zheng, Y. (2016). pg-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data. *IEEE Transactions on Big Data*, *4*(4), 571–585. doi: 10.1109/tbdata.2017.2723899