

Causal Decomposition of Complex Systems

℘

Prediction of Chaos using Machine Learning

Kausale Zerlegung von komplexen Systemen

℘

Vorhersage von Chaos mittels maschinellen Lernens

A Dissertation at the
Faculty of Physics of the
Ludwig Maximilian University Munich

Eine Dissertation an der
Fakultät für Physik der
Ludwig-Maximilians-Universität München

submitted by *vorgelegt von*

Haochun Ma

on the *am*

30th November 2023

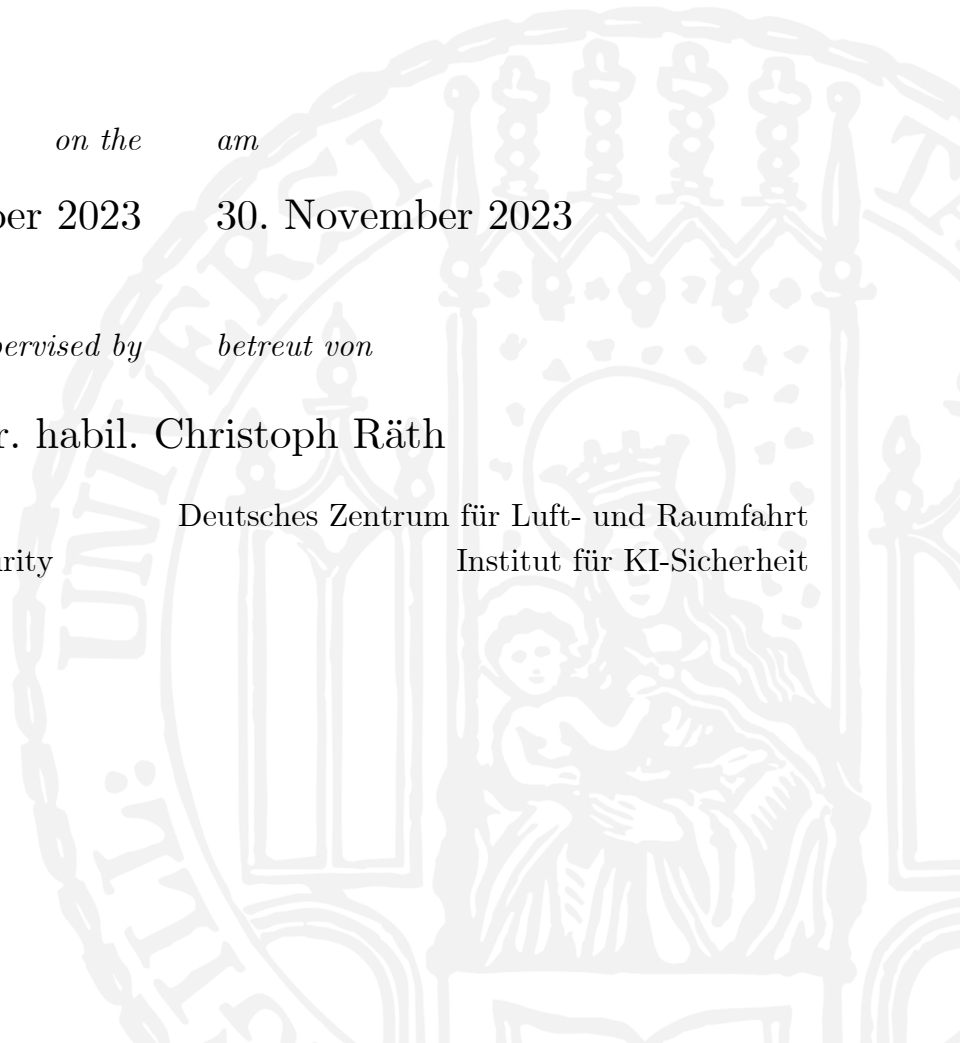
30. November 2023

supervised by *betreut von*

PD Dr. habil. Christoph Räth

German Aerospace Center
Institute for AI Safety & Security

Deutsches Zentrum für Luft- und Raumfahrt
Institut für KI-Sicherheit



1st Referee
2nd Referee
Doctoral Examination

PD Dr. habil. Christoph R ath
Prof. Dr. Thomas Kuhr
11th January 2024

To those who made this dissertation possible

First and foremost, I would like to express my deepest gratitude to *Christoph R ath*, who is, quite simply, the best supervisor anyone could ask for. Your never-ending support, extensive expertise, and oozing passion for our field of research were the main driving force behind this dissertation — next to the countless meetings and discussions in the Aumeister beer garden, which did not end even after the bell rang.

Among the participants in these surprisingly productive meetings were *Davide Prosperino* and *Alexander Haluszczynski*, who contributed greatly to the research presented in this dissertation. I am especially grateful to both of you for proofreading my manuscripts when, blunted by countless readings, I could no longer recognize the most obvious errors.

Before diving into the research, which is the fun part of an academic career, comes the hard and often painful part of solving problem sheets and writing exams. Without my dear friends *Gordian Edenhofer*, *Jakob Roth*, *Johannes B urger*, *Lukas Nakamura*, and *Paul Ockenfu *, I would not have made it through my studies, as their precise note-taking, perfect homework solutions, and (allowed) cheat sheets were essential. I am convinced, however, that our hour-long discussions, cookouts, and highly subsidized concert visits were of even greater importance.

Words cannot express the gratitude that I have for the two most important people in my life: my *Mum* and *Dad*. Having you as my parents and being blessed with your unconditional love and support is all I need. So this dissertation is for you.

Thank you all so much

Zusammenfassung

Wir leben in einem *komplexen* System. Daher ist es unerlässlich, über Techniken zur Analyse und zum Verständnis seiner verschleierte Dynamik zu verfügen, um die Entscheidungsfindung zu verbessern. Ziel dieser Dissertation ist es, einen Beitrag zur Forschung zu leisten, die unsere Möglichkeiten erweitert, diese komplexen Systeme für uns weniger intransparent zu machen.

Zunächst wird aufgezeigt, welche Auswirkungen es auf praktische Anwendungen hat, wenn *Nicht-linearität* — ein oft vernachlässigter Faktor bei *kausaler Inferenz* — berücksichtigt wird. Daher untersuchen wir die kausalen Beziehungen innerhalb dieser Systeme und beleuchten insbesondere die Unterscheidung zwischen linearen und nichtlinearen Kausalitätsfaktoren. Nachdem wir die erforderlichen Methoden entwickelt haben, wenden wir sie auf einen realen Anwendungsfall an und zeigen, dass leichte Anpassungen bestimmter Finanzmarktmodelle durch die Auflösung des *Korrelations-Kausalitäts-Fehlschlusses* zu erheblichen Vorteilen führen können.

Sobald die linearen und nichtlinearen Kausalzusammenhänge bekannt sind, können wir aus der zugrunde liegenden Kausalitätsstruktur die *Differentialgleichungen* ableiten, um die *Interpretierbarkeit* von Modellierungen und Vorhersagen zu verbessern. Durch die Feinjustierung der Parameter dieser Gleichungen durch das Phänomen der *Synchronisierung von Chaos* können wir sicherstellen, dass sie die Daten optimal darstellen.

Allerdings lassen sich nicht alle komplexen Systeme durch Differentialgleichungen adäquat beschreiben. Daher bietet die Anwendung von Techniken des *maschinellen Lernens* wie *Reservoir Computing* bei der Vorhersage chaotischer Systeme erhebliche datenbasierte Vorteile. Obwohl ihre Architektur relativ einfach ist, ist die Gewährleistung einer vollständigen Interpretierbarkeit und *Hardware-Realisierung* immer noch von einer erhöhten Effizienz und reduzierten Datenanforderungen abhängig. In dieser Dissertation werden einige der notwendigen Änderungen an der traditionellen Architektur vorgestellt, um physikalisches Reservoir Computing näher an die Realisierung zu bringen.

Abstract

We live in a *complex* system. Therefore, it is essential to possess techniques to analyze and comprehend its intricate dynamics in order to improve decision making. The objective of this dissertation is to contribute to the research that enhances our ability to make these complex systems less intransparent to us.

Firstly, we illustrate the impact on practical applications when *nonlinearity* — an often disregarded factor in *causal inference* — is taken into account. Therefore, we investigate the causal relationships within these systems, particularly shedding light on the distinction between linear and nonlinear drivers of causality. After developing the necessary methods, we apply them to a real-world use case and demonstrate that making slight adjustments to certain financial market frameworks can result in considerable advantages because of the resolution of the *correlation-causation fallacy*.

Subsequently, once the linear and nonlinear causal connections are understood, we can derive *governing equations* from the underlying causality structure to enhance the *interpretability* of models and predictions. By fine-tuning the parameters of these equations through the phenomenon of *synchronization of chaos*, we can ensure that they optimally represent the data.

Nevertheless, not all complex systems can be accurately described by governing equations. Therefore, the implementation of *machine learning* techniques like *reservoir computing* in predicting chaotic systems offers significant data-driven advantages. While their architecture is relatively simple, ensuring full interpretability and *hardware realizations* still relies on increased efficiency and reduced data requirements. This dissertation presents some of the necessary modifications to the traditional reservoir computing architecture to bring physical reservoir computing closer to realization.

Contents

Dedica

Zusammenfassung

Abstract

1	Introduction	1
2	Identifying Linear and Nonlinear Causality Drivers	15
2.1	Background and Motivation	16
2.2	Causal Inference and Decomposition	18
2.3	Financial Data and Frameworks	26
2.4	Framework Validation and Application Results	30
3	Deriving Governing Equations using Causality and Synchronization	41
3.1	Background and Motivation	42
3.2	Derivation of Equation Terms from Causality	44
3.3	Estimating Equation Parameters using Synchronization	45
3.4	Algorithm Validation and Application Results	51
4	Predicting Chaos with Binary and Minimal Reservoir Computing	59
4.1	Background and Motivation	60
4.2	Prediction of Dynamical Systems	61
4.3	Evaluating Predictions of Chaotic Systems	68
4.4	Block-Diagonal Reservoirs	71
4.5	Minimal Reservoir Computing	73
4.6	Prediction Results and Parameter Robustness	77

Summary and Outlook

Synthetic Systems

Publications

- Identifying Causality Drivers and Deriving Governing Equations of Nonlinear Complex Systems
- Efficient Forecasting of Chaotic Systems with Block-Diagonal and Binary Reservoir Computing
- A Novel Approach to Minimal Reservoir Computing
- Linear and Nonlinear Causality in Financial Markets

Bibliography

Chapter 1

Introduction

We are in an era where an unprecedented amount of data is being produced, processed, and used in decision making. This presents us with a number of challenges and opportunities when it comes to using this vast amount of data to build models and make predictions. This can be challenging because reality rarely adheres to linear dynamics. Both human-made systems and nature exhibit complexity, nonlinearity, and even chaos. While simplifications and linearizations may be sufficient for basic modeling tasks, real-world applications require a deeper understanding of the full dynamics involved.

Physicists have traditionally been at the forefront of this investigation, developing the tools necessary to study the complex relationships that govern our world. The widespread adoption of machine learning techniques has made it possible to capture and predict the full intricate nature of nonlinear dynamical systems, which is now a tangible reality. While effective, these methods conceal their underlying processes by operating in unexplainable, high-dimensional spaces, leaving users to rely on trust rather than understanding. Yet, transparency is critical in fields as diverse as climate modeling, turbulent air flows, epidemics, and financial markets — areas where explainability is as essential as accuracy.

The convergence of physics, mathematics, and computational innovation has laid the groundwork for novel methods to analyze nonlinearities and chaotic behavior present in various scientific and technological fields. This dissertation aims to merge traditional principles with modern computational techniques to reveal the intrinsic dynamics of complex systems. By integrating physics with state-of-the-art machine learning, we will investigate methods for measuring linear and nonlinear causality, deriving governing equations, and predicting the chaotic behavior of complex systems.

Complex Systems and Chaos

Chaos: When the present determines the future, but the approximate present does not approximately determine the future.

On a piece of paper
Edward Lorenz

In scientific investigation, few concepts have expanded knowledge and sparked imagination like complex systems and chaos. These concepts not only exist as theoretical musings, but also establish the foundation of our world with insights that transcend various disciplines and domains. Complex systems comprise interrelated networks of elements that exhibit intricate behaviors and structures. They are recognized for their dynamic nature in which interactions between components can result in hard-to-predict phenomena that cannot be explained by an understanding of the individual components alone [1]. These systems challenge established notions of cause and effect, from the coordinated flight of birds to the patterns of financial markets and the functionality of the human brain. Chaos theory, an aspect of the study of complex systems, investigates how small alterations in initial conditions may lead to vastly differing outcomes. An emblematic demonstration of this theory is the famous butterfly effect [2], a metaphor suggesting that a butterfly flapping its wings in Brazil could potentially spawn a tornado in Texas. Chaos indicates a fundamental characteristic of mathematical and physical systems that holds significant repercussions for predictability and control.

Mathematical Properties of Chaos

By understanding chaos mathematically, researchers gain valuable insights into the behavior of complex systems. They can predict outcomes within established boundaries and uncover fundamental order amidst seemingly chaotic phenomena. It is important to note that chaos and randomness are distinct, as chaos arises from deterministic dynamics that are governed by mathematical principles. These principles are instrumental in defining and evaluating chaotic systems [1]. Although there is no universally agreed-upon mathematical definition of chaos, the following common principles are fundamental to this dissertation:

- *Sensitivity to Initial Conditions*: small shifts in parameters can drastically alter the behavior of a system, causing bifurcations and unpredictable transitions. Positive *Lyapunov* exponents confirm this erratic nature as they indicate an exponential divergence of trajectories, even from very close starting points. Together, these characteristics underscore the inherent challenge in predicting and understanding chaotic dynamics, highlighting their instability and complex nature [3].
- *Topological Transitivity*: this property guarantees that for any two open sets in the space, trajectories from one will eventually intersect the other, reflecting the system's inherent mixing nature and contributing to its sensitive dependence on initial conditions [1].
- *Synchronization*: this phenomenon arises when two or more initially uncorrelated chaotic systems exhibit coordinated behavior. This coordination can take the form of complete, phase, or generalized synchronization, leading to a functional relationship between the states of the systems. The mechanism driving this phenomenon involves coupling and mutual adaptation [4].

The Lorenz Attractor

The history of understanding chaotic systems has a significant presence in the field of physics. In 1687, Newton's solution of the two-body problem marked a significant milestone [5], but the subsequent extension to the three-body problem revealed the complex dependence on initial conditions, leading to chaotic dynamics. This insight, as elaborated by Henry Poincaré in the 1890s [6], led to the development of alternative strategies, such as approximate techniques and topological methodologies, including the *Poincaré map*. Nevertheless, it was not until the latter part of the 20th century that chaos theory gained significant attention and recognition, with contributions from Edward Lorenz, Mitchell Feigenbaum, and James Yorke.

In 1963, Edward Lorenz published a groundbreaking scientific paper entitled *Deterministic Nonperiodic Flow* [7]. This paper presented Lorenz's findings on the sensitive dependence of initial conditions within the Lorenz system, ultimately challenging the previously held belief of predictability within complex systems. Furthermore, Lorenz's work introduced the concept of chaos, which has since inspired a vast array of subsequent research within the field.

The mathematical theory underlying the Lorenz system revolves around its nonlinear dynamics and the emergence of chaos. The three differential equations describe the evolution of three variables: x , y , and z . These variables represent the convective flow, temperature variation, and vertical temperature variation, respectively. The equations involve parameters that govern the system's behavior, such as the *Prandtl number* σ , the *Rayleigh number* ρ , and a system-specific constant β :

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z.\end{aligned}$$

The solutions to the Lorenz equations exhibit several distinctive properties and have had a profound impact on our understanding of chaotic behavior, the limitations of deterministic models, and the sensitivity to initial conditions. Even small variations in the initial values of x , y , and z can lead to vastly different trajectories over time, as illustrated in Figure 1.1.

This sensitivity is a hallmark of chaotic behavior and exemplifies the *butterfly effect*. Additionally, the solutions to the Lorenz system exhibit a *strange* attractor, which is a fractal geometric structure that captures the long-term behavior of the system [8]. The Lorenz attractor is a globally attracting, non-periodic orbit characterized by its complex shape resembling butterfly wings or the number eight [1]. The system's mathematical simplicity has made it a widely used example in textbooks, research papers, and computer simulations to illustrate key concepts of chaos theory. The Lorenz equations have been instrumental in the development of numerical methods, bifurcation analysis, and the study of strange attractors.

Lorenz's research uncovered the limitations of deterministic weather prediction models, emphasizing the unpredictability of complex atmospheric systems. The recognition of chaos and sensitivity to initial conditions has had far-reaching influences on meteorology, climate science, and general comprehension of complex systems. Therefore, the implications of the Lorenz system transcend the field of mathematics and physics. It has led scientists to adopt stochastic models and ensemble forecasting techniques to account for the inherent uncertainties in complex systems. The advent of computers and advancements in computational techniques facilitated the exploration and visualization of chaotic systems, leading to applications in various scientific disciplines, and even art. Today, the Lorenz attractor has achieved iconic status and is often used in popular culture as a visual representation of chaos and complexity.

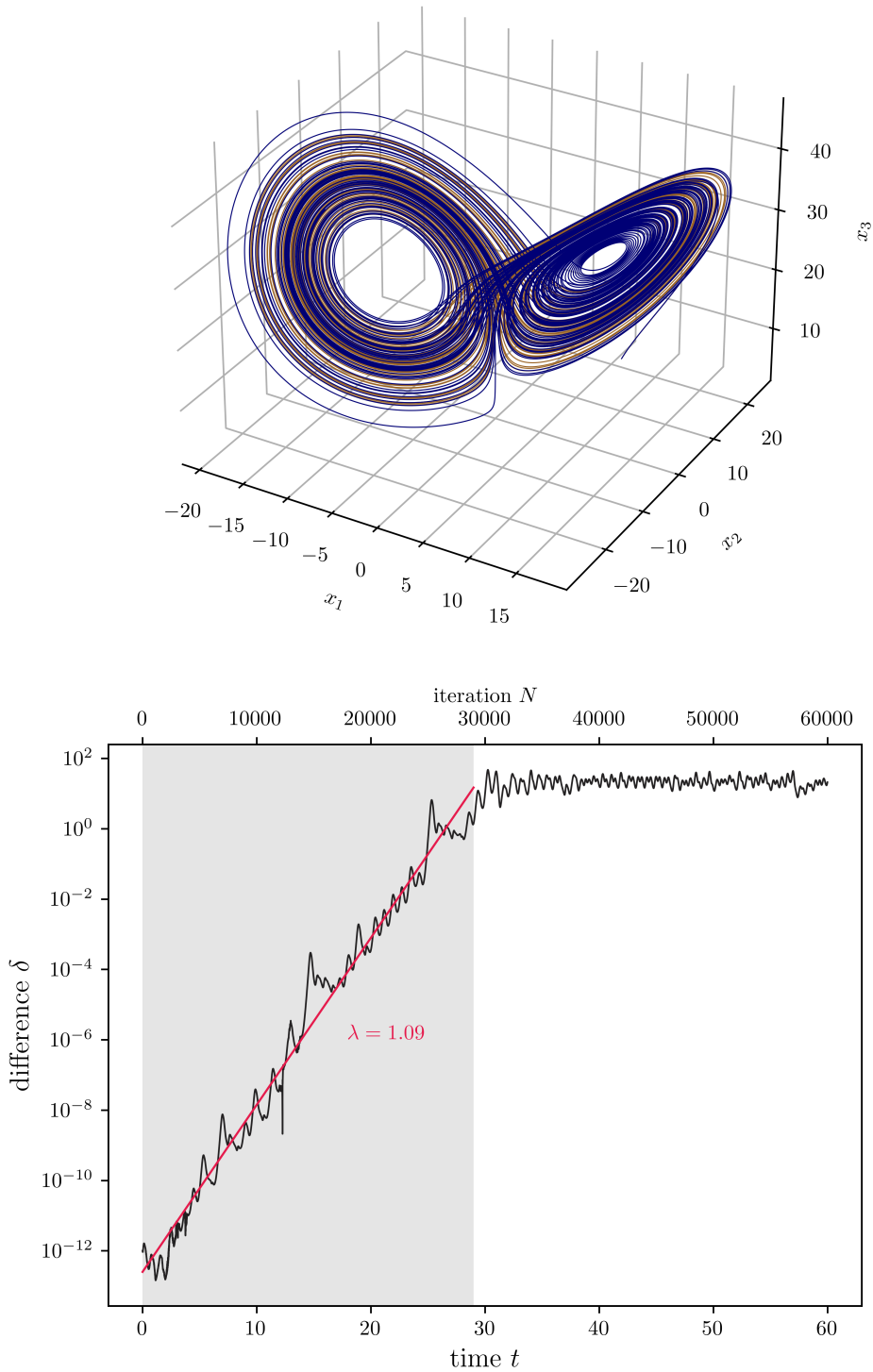


Figure 1.1: Lorenz Attractor and Chaos. The upper graph illustrates two separate trajectories on the Lorenz attractor. One can observe why the term *strange* is fitting upon examining the pattern. Both trajectories have the same parameters and are integrated over 10^6 steps with a step size of 10^{-3} . The blue trajectory begins from the initial condition $(1, 1, 1)^\top$, while the brown trajectory starts from the initial condition $(1 + 10^{-12}, 1, 1)^\top$. Both colors' clarity in the plot confirms that these trajectories follow different routes on the attractor. The lower graph exhibits a logarithmic scale view of the norm of the discrepancy between these two trajectories. A clear linear relationship is evident in the initial phase, which is highlighted in gray. This linearity enables the determination of the largest Lyapunov exponent λ through the linear interpolation depicted in red. Adapted from Prosperino [9].

Causal Inference and Decomposition

Everyone who confuses correlation with
causation eventually ends up dead.

An Internet Meme
Unknown

*Understanding how simple rules and interactions can drive complex and unpredictable behavior is crucial, as modeling, simulating, and analyzing complex systems are indispensable tools in our rapidly interconnected world. Therefore, understanding the causal connections within a system is of great importance. The history of causal inference dates back to the development of scientific thinking and the quest to comprehend the relationships between cause and effect. Scientists throughout history have confronted the intricacies of causality, leading to the emergence of various concepts and theories [10]. In the context of classical Newtonian physics, causality was interpreted as the simultaneous interaction of *actio et reactio* [5]. However, Einstein's theory of general relativity expanded the concept of causality, connecting events through the light cone of spacetime and adding temporal and spatial dimensions [11]. The introduction of quantum mechanics further complicated the notion of causality, introducing probabilistic behavior and non-determinism, rendering the idea of a strict cause-and-effect relationship inconceivable [12]. In today's data-driven world, it is imperative to distinguish between cause and correlation and comprehend the complex relationship between variables over time.*

Evolution of Causal Inference

The evolutionary path of causal inference mirrors the conceptual evolution of causality. This is exemplified by Granger's model, which captures analogous patterns through time-lagged time series regression. The model aligns with our conventional understanding of causality, depicting a temporally shifted series of events. Nonetheless, Schreiber [13] resolved a key drawback of *Granger Causality* [14], which solely measures linear interdependencies. He utilized information theory and probability-based metrics to gauge the decline in unpredictability between paired time series using *Transfer Entropy*. As research on chaotic nonlinear systems advanced, including the previously mentioned Lorenz attractor, causal inference techniques began to integrate the reconstruction of the dynamic interplay between interconnected variables. One of the most advanced and innovative techniques emerging from the so-called *state space reconstruction* methods is *Convergent Cross Mapping* [15], which is based on *Takens' Theorem* and the transitive relationships inherent in dynamic system topology.

In the early stages of causal inference, computational resource limitations were often a bottleneck for researchers to leverage complex models and methods. However, advancements in computational power expanded the horizons of possibilities for causal inference. The arrival of powerful computing machinery, parallel processing, and optimized algorithms paved the way for more sophisticated and effective causal analyses [10]. As we consider the future, the interplay between computational resources and causal inference holds tremendous potential for groundbreaking discoveries. The advent of quantum computing, big data analytics, and artificial intelligence presents new opportunities and challenges, paving the way for further understanding of causality [16].

Decomposing Linear and Nonlinear Causality

One crucial aspect of contemporary causal analysis is disentangling the linear and nonlinear properties, a nuanced yet vital approach that enhances our understanding of complex systems, which constitutes the core of this dissertation. The use of *Fourier Transform* surrogates provides powerful methods for evaluating and interpreting such causal relationships. The methodology relies on the *Discrete Fourier Transform* of the time series, which separates linear properties into amplitudes and nonlinear properties into phases:

$$X[k] = \sum_{n=0}^{N-1} x[n] \cdot e^{-i\frac{2\pi}{N}kn}.$$

In this expression, $x[n]$ denotes the time series in the time domain and $X[k]$ represents its counterpart in the frequency domain. The frequency corresponding to a given k is $\frac{k}{NT}$ Hz, where T is the sampling interval and N is the total number of samples in the time series.

Randomizing the phases of the Fourier transform with uniformly distributed numbers selectively destroys nonlinear features while leaving the linear characteristics unaffected [17]. This elegant technique allows us not only to distinguish between linear and nonlinear behavior, but also to probe deeper into the causal structure. It provides opportunities for advanced analysis, allowing us to quantify nonlinear causal links and formulate governing equations based on the foundational causal interdependencies.

Deriving Governing Equations

As seen for the Lorenz system, understanding the governing differential equations of a dynamical system is paramount for precise modeling and prediction, as it provides a mathematical representation of the system's inherent dynamics. Expressed in forms as:

$$\frac{d\mathbf{x}}{dt} = \mathbf{F}(\mathbf{x}, t),$$

the equations encapsulate how state variables evolve over time, with \mathbf{x} representing the state vector, t denoting time, and $\mathbf{F}(\mathbf{x}, t)$ specifying the rate of change. By employing numerical methods such as the *Runge-Kutta integration* [18] for solving the equations, one can approximate the system's future states, even when analytical solutions are intractable. Subsequently, the governing equations facilitate the identification of equilibrium points and the assessment of stability through techniques like linearization and the calculation of eigenvalues [19]. Additionally, they enable sensitivity analysis, allowing for the determination of how variations in parameters affect system behavior, which is crucial for optimization and control. Furthermore, having a clear mathematical formulation of the system lays the groundwork for integrating advanced computational techniques, including machine learning, to enhance predictive accuracy and handle large-scale, complex systems.

In this dissertation, we employ an integrative approach that connects the analysis and inference of causality with the derivation of governing equations within complex nonlinear systems. This innovative mathematical technique enables us to create a clear justification for deriving the corresponding differential equations based solely on causalities. Additionally, we use synchronization of chaos to precisely calibrate the parameters of the equations. Our methodology captures the essential features of the system while providing transparency and interpretability — often lacking in modern computational models. Hence our research provides evidence of the ongoing improvement in decoding complex systems.

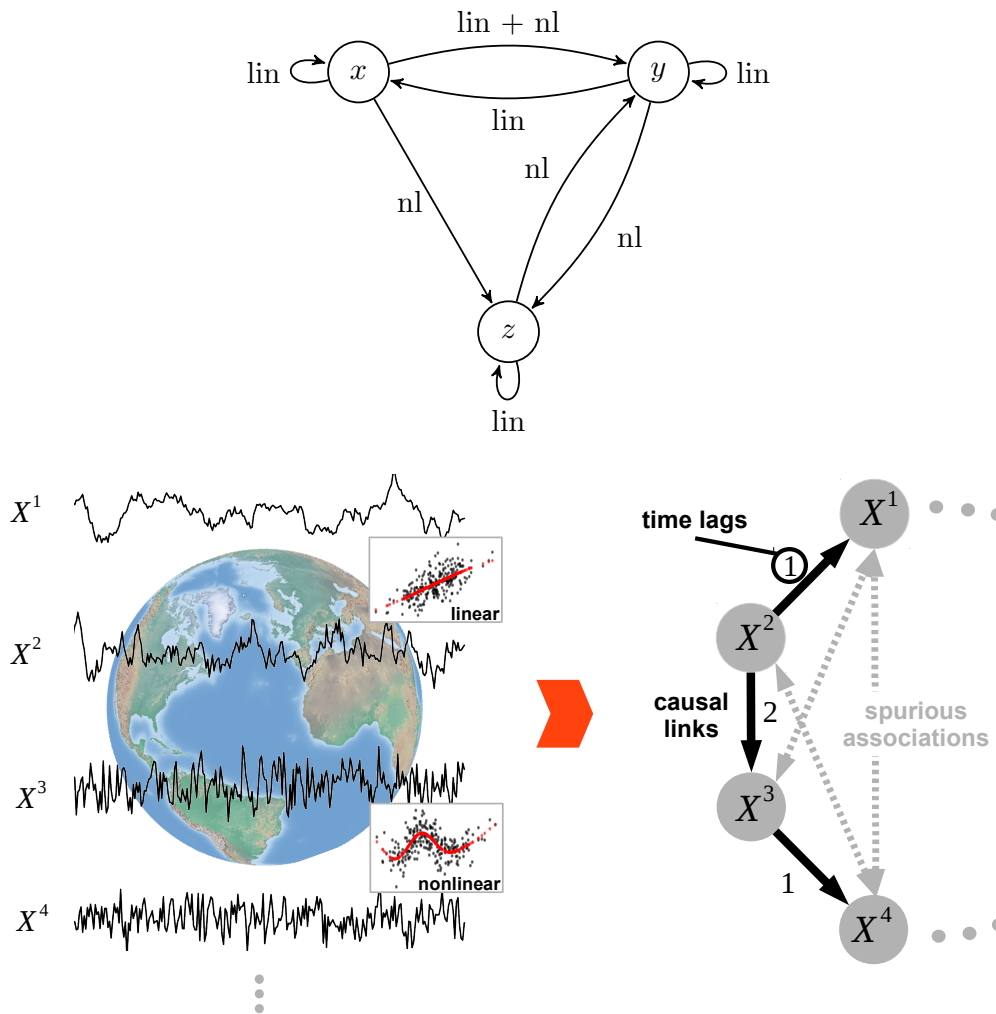


Figure 1.2: Causal Inference. The upper graph displays the causality pictogram depicting the linear and nonlinear dependencies between the state variables of the Lorenz system. The bottom graph provides an illustration of a real-world equivalent to the causal inference problem. For instance, consider a large-scale time series dataset in the left panel originating from a complex system like the Earth's. The goal is to reconstruct the causal relationships illustrated in the right panel, giving equal attention to both linear and nonlinear dependencies, along with their corresponding time lags (as indicated by the link labels). Differentiating legitimate causal associations from false ones, which can occur due to common drivers (e.g. $X^1 \leftarrow X^2 \rightarrow X^3$) or indirect transitive pathways (e.g. $X^2 \rightarrow X^3 \rightarrow X^4$), poses a challenge. Causal inference aims to uncover hidden links between variables, resulting in the creation of causal networks that are less dense than mere correlation networks. Adapted from Runge, Nowack, Kretschmer, et al. [20].

Prediction of Dynamical Systems

I never predict anything, and I never will.

Legendary Football Interview
Paul Gascoigne

When governing equations provide detailed knowledge about a system's behavior, such as in many physical or mechanical systems, numerical methods can be used for forecasting [21]. These models rely heavily on the system's rigorous mathematical underpinnings and offer a reliable approach for comprehending and predicting system dynamics [22]. They serve as a link between theoretical physics and real-world applications, translating equations into practical insights. However, the terrain of dynamical systems is far from uniform, and in many cases the underlying equations may be unidentified, only partially understood, or too computationally expensive to be feasibly simulated. This gap between our imperfect understanding and the complicated nature of mathematical models presents a challenge that traditional methods cannot easily overcome. This is where machine learning models excel, particularly in opaque terrain. Unlike knowledge-based models, machine learning approaches to forecasting time series do not require an intimate understanding of the system's governing equations. Machine learning relies solely on data, drawing inferences and making predictions through computational algorithms that learn from the patterns and structures within the data itself [23]. This data-driven approach offers new opportunities for prediction in scenarios where standard methods fail, while also facilitating navigation of complex systems with elusive governing equations. Nonetheless, the balance between understanding and computational efficiency is a critical consideration.

Machine Learning and Physics

Time series data that arises from physical systems is inherently nonlinear and subject to a multitude of interacting factors, presenting a considerable challenge for prediction. *Machine Learning* algorithms have become powerful tools in the field of time series prediction [24]. The adaptability, robustness, and flexibility of deep learning algorithms have paved the way for predicting complex dynamic systems that were once unfeasible to simulate using traditional methods [25]. Among these architectures, *Recurrent Neural Networks* and *Long Short-Term Memory* networks have gained recognition for their ability to capture temporal dependencies [26]. Their natural aptitude for learning from sequences, retaining previous states, and anticipating forthcoming states renders them particularly capable in situations where time dependence is a critical factor [27].

Although *black-box* techniques have made significant strides, certain challenges still impede their full potential. In domains like climate modeling, energy systems, and finance, where comprehending the model's decision-making process is as crucial as its prediction accuracy, interpretability is vital. *Physics-informed machine learning* offers a promising paradigm that combines the benefits of data-driven methods with domain-specific physical knowledge [28]. In this hybrid approach, machine learning architectures incorporate the governing equations of physical processes directly [29]. This yields greater predictive accuracy and interpretability, particularly in situations with limited data — a frequent issue in physical sciences and engineering [30]. Additionally, the combination of both approaches can reveal new relationships that may be missed by each individual approach, resulting in innovative insights [31].

Reservoir Computing

In the ever-expanding world of machine learning and computational modeling, *Reservoir Computing* is a prominent paradigm. It bridges the gap between intricate computations and practical implementations [32]. Traditional recurrent neural networks require extensive training for all connections, whereas reservoir computing simplifies the learning process by training only the connections that lead out of a fixed, random neural network known as the *reservoir* [33]. This approach has recently attracted the attention of researchers and practitioners alike due to its simplicity, efficiency, and broad applicability, especially in the analysis of temporal data [34]. Furthermore, the simple architecture and limited number of hyperparameters facilitate interpretability, prompting extensive research on the interpretation and comprehension of reservoir computing’s learning process and how individual architecture components affect predictions [35].

However, what distinguishes reservoir computing and holds promise for advancement is its compatibility with hardware implementation. Reservoir computing’s fixed internal connections make it highly appropriate for physical realization [36], allowing reservoir computing’s potential to extend beyond standard software applications. One of the most promising areas for hardware implementation is in optoelectronic and photonic systems [37]. These systems utilize the unique characteristics of light to create high-speed and energy-efficient reservoirs, which can process information at a scale and speed that is impossible with just electronic means. *Neuromorphic* engineering is showcasing new opportunities for embedding reservoir computing into silicon that mimics the highly interconnected network of neurons and synapses in the brain [38]. This avenue can promote more energy-efficient and biologically plausible hardware that sheds light on both artificial intelligence and neuroscience. Additionally, the recent investigation of advanced components like memristors and quantum devices introduces novel prospects in hardware-accelerated reservoir computing [39].

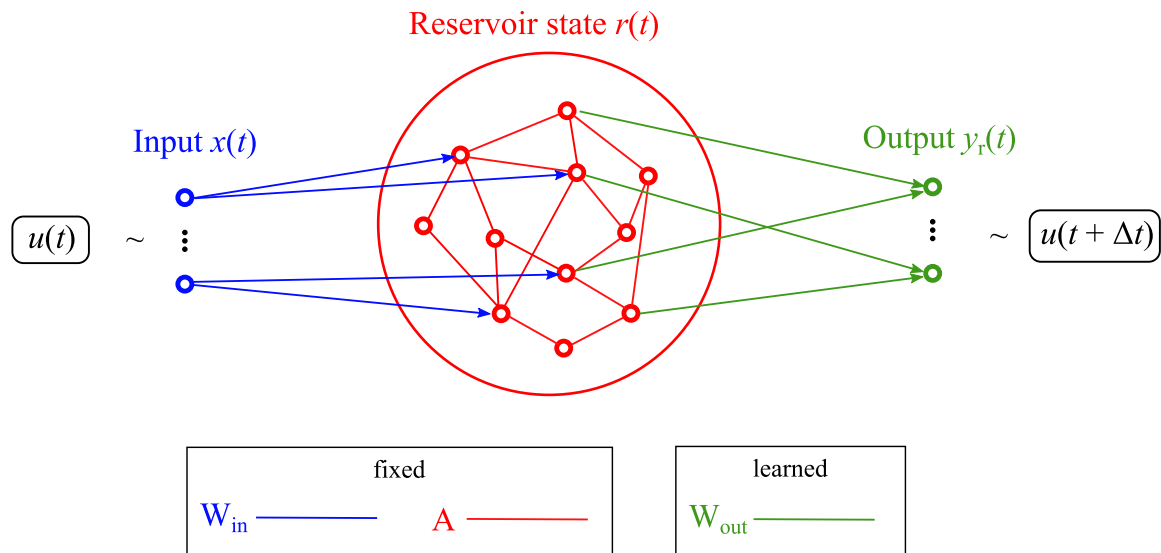


Figure 1.3: Reservoir Computing. This Figure outlines the fundamental architecture of reservoir computing. The learning efficiency is based on the simplicity of the architecture, which is mainly fixed. The input weights (blue) and the reservoir (red) are randomly created and fixed throughout the learning process. Merely the output weights (green) are optimized using a linear regression, making the process very fast. Adapted from Duncan [40].

Chapter 2

Identifying Linear and Nonlinear Causality Drivers

H. Ma, A. Haluszczyński,
D. Prosperino & C. R ath
“Identifying causality drivers and
deriving governing equations of
nonlinear complex systems”
*Chaos: An Interdisciplinary
Journal of Nonlinear Science*
vol. 32, no. 10, 2022

H. Ma, D. Prosperino,
A. Haluszczyński & C. R ath
“Linear and nonlinear causality
in financial markets”
*Submitted to
Chaos: An Interdisciplinary
Journal of Nonlinear Science*
TBD, 2023

Understanding cause-and-effect relationships presents a significant challenge for developing analytical and predictive models in various scientific fields. While techniques for inferring causality are consistently evolving, adequately addressing the identification of its drivers remains a major obstacle. This issue is particularly critical when dealing with complex systems, where determining whether causality emerges from linear or nonlinear properties proves extremely valuable. Before implementing our developed methods in real-world complex systems, we validate them on synthetic chaotic systems and demonstrate the significance of nonlinear features in causality. In the financial sector, researchers and practitioners must identify and measure interdependencies among financial instruments. However, conventional techniques such as Pearson correlation exhibit constrained descriptive aptitude and only capture linear dependencies [41]. Therefore, we present a comprehensive approach that includes both linear and nonlinear causalities. We detect significant nonlinear causality in stock indices in Germany and the United States. While correlation may approximate linear causality, it fails to account for nonlinear factors, leading to an underestimation of causality. Our research also highlights the potential use of causality in generating market signals, implementing pair trading, and managing portfolio risk.

2.1 Background and Motivation

Causality, one of the fundamental principles of scientific inquiry, has been extensively studied across many generations and disciplines. Over time, understandings of causality have developed alongside the growth in complexity of physical theories. Although causal inference mainly aims to measure causality, investigating its properties and drivers has been of secondary importance to research. Separating linear and nonlinear causal features is not only an academic exercise but also has significant practical implications [42]. Incorrectly attributing a nonlinear relationship as linear can lead to misguided conclusions, poor predictions, and potentially disastrous real-world decisions. Conversely, recognizing nonlinearity can reveal deeper insights into the underlying mechanisms of a system, opening the door to more effective interventions and controls.

Initial advancements in this field were achieved by Paluš, Albrecht, & Dvořák [43], resulting in the development of a diagnostic test that was specifically designed to identify nonlinear dynamic relationships in time series using *Mutual Information*. Haluszczyński, Laut, Modest, et al. [42] then took a different approach, utilizing *Fourier Transform* surrogates to distinguish linear and nonlinear components of mutual information. This approach was utilized to identify nonlinear correlations within financial data. Additionally, Hlinka, Hartman, Vejmelka, et al. [44] conducted further quantification of the impact of nonlinearity on connectivity, particularly in the context of climate data.

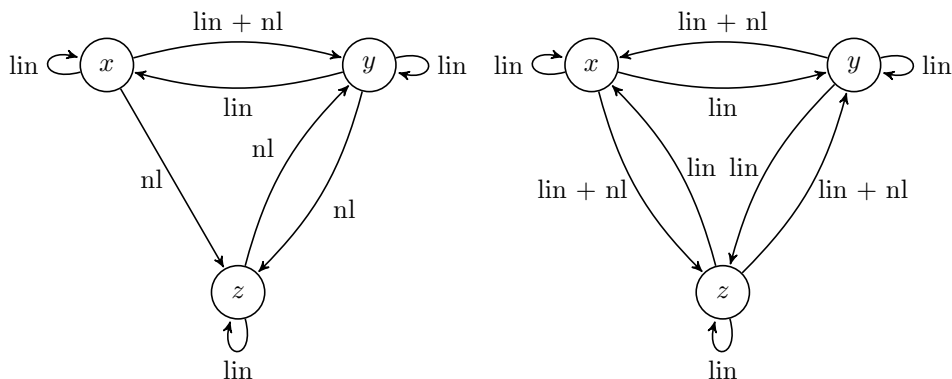


Figure 2.1: Causality Pictograms. The linear (lin) and nonlinear (nl) causal links of the Lorenz (left) and Halvorsen (right) systems are represented by the labeled arrows. Their governing equations are defined in the Equations 51 and 53, respectively. The central goal of our framework is to accurately identify these links and then express their magnitude in a reliable manner.

Understanding the interdependence of financial assets is crucial in several financial sectors, particularly in assessing portfolio-related risks [45]. As a consequence, industry practitioners have been closely monitoring the development of co-dependency metrics while the field of econophysics is receiving growing attention in the physics community, providing a fresh outlook on traditional financial approaches [46]. This new outlook utilizes statistical physics tools, including signal processing, agent-based market frameworks, and random matrix theory [47].

Predominantly, the co-dependencies of financial instruments are characterized by the linear co-dependency metrics of their return time series. However, there is a growing body of research that highlights the nonlinear characteristics of these series [48]. In particular, Mantegna & Stanley [49] demonstrated the power-law scaling dynamics of the probability distributions of financial indices, while Ghashghaie, Breymann, Peinke, et al. [50] identified turbulent cascades in foreign exchange markets. Such findings call into question the adequacy of linear dependence metrics.

A pressing issue is the continued reliance on Pearson correlation [51] as a proxy for causality due to the complexities of determining causality in dynamic systems. Granger’s famous study in the 1960s [14] specifically addressed the *correlation-causality fallacy* and led to the development of more advanced tools for causal inference — nevertheless, the use of Pearson correlation remains popular due to its simplicity in calculation and interpretation as seen in Equation 2.1.

However, even very simple dynamics, such as the *coupled difference* which is described in Equation 56, can lead to misinterpretations regarding the quantification of co-dependence. Due to the chaotic nature of this system [52], it exhibits so-called *mirage correlations*, which means that variables may be positively correlated for long periods but can spontaneously become anti-correlated or non-correlated. This can lead to problems when fitting models or inferring causality from observational data [15] as shown in the Figure 2.2 below:

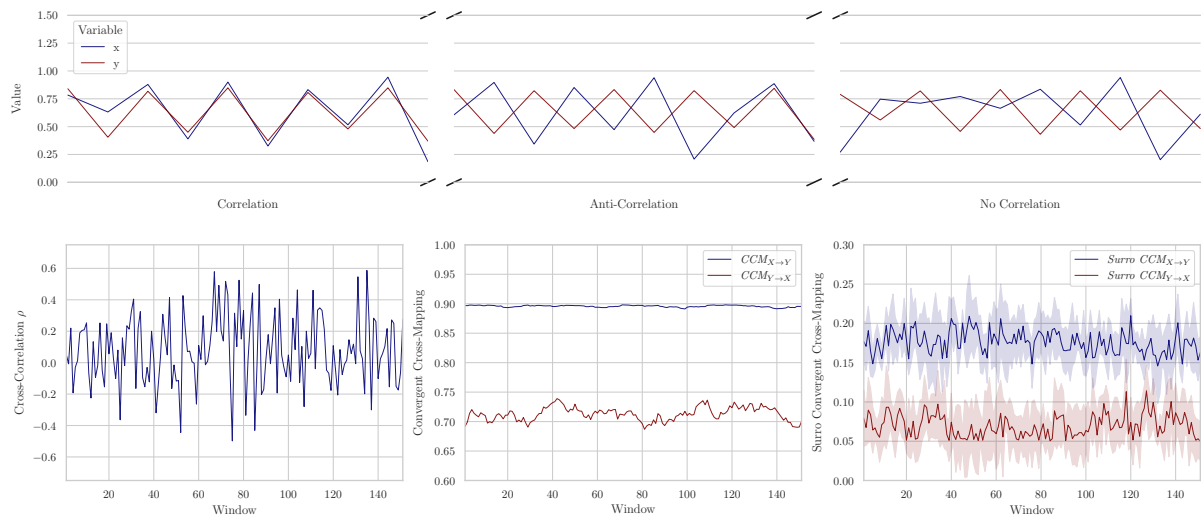


Figure 2.2: Mirage Correlations and Causality. The top row shows different regimes of the coupled difference system defined in Equation 56. It appears that the variables are correlated in the first regime, anti-correlated in the second, and lose all coherence in the third. The bottom row shows rolling correlation (left), causality (center), and linear causality (right). Causality is measured using *Convergent Cross Mapping*. While correlation fluctuates between periods of positivity, negativity, and zero correlation, the causality remains steady in both directions over time. The same is applicable to linear causality. After comparing the measurements with the governing equations, it is evident that causality offers a better representation of the co-dependence between the two state variables, and is more stable and accurate than correlation.

This bias poses a potential danger when practitioners rely solely on correlation to assess portfolio risk. To mitigate this risk, we propose a straightforward solution: incorporate measures of causality into the processes of market signal inference, pair trading, and portfolio construction. By doing so, we aim to effectively address this bias and improve the accuracy and robustness of these financial practices.

2.2 Causal Inference and Decomposition

The concept of causality is deeply rooted in our way of thinking and is a fundamental principle of physics. In the following, we discuss several methods in order to infer causality within complex systems. It is a domain where statistical techniques meet philosophical interpretation, striving to unveil the underlying causal relationships that govern observed phenomena [53]. The search for comprehension of complex relationships has inspired the creation of numerous techniques aimed at unraveling causality. Among these, three prominent categories — *Granger Causality*, *Transfer Entropy*, and *Convergent Cross Mapping* — have come to the forefront of contemporary research [10].

- *Granger Causality* measures the extent to which the past values of one time series variable can predict the future values of another variable [14]. It is based on the idea that a cause should precede its effect in time. Granger Causality is widely used due to its simplicity and interpretability, but it is limited to detecting only linear dependencies between variables.
- *Transfer Entropy* is an information-theoretic measure that quantifies the amount of information transferred between time series variables. It captures the nonlinear dependencies and can detect causal relationships even in the presence of noise. Transfer Entropy can be interpreted as an extension of Granger causality and was shown by Barnett, Barrett, & Seth [54] to be equivalent for Gaussian random variables.
- *Convergent Cross Mapping*, a state-space reconstruction method, aims to reconstruct the underlying dynamics of a system based on embedded time series data [15]. It leverages the topological structure of the system and its attractors to infer causal relationships between state variables.

However, it should be noted that our framework applies to any method that can detect nonlinear causality.

Pearson Correlation

Before delving into the methods of causal inference, we first introduce the *Pearson Correlation* [41]. The coefficient is named after Karl Pearson, who developed it from a related concept that Francis Galton introduced during the late 19th century. It was a significant development in statistics as it allowed for the quantification of the strength and direction of a linear relationship between two continuous variables, and has since become a standard tool in experimental design and data analysis within numerous scientific disciplines [51].

It serves as a benchmark in our research as it continues to be widely used in the financial industry due to its ease of calculation and interpretability. This statistical measure quantifies the strength and direction of the linear relationship between two variables. It is computed as follows:

$$\rho(\mathbf{x}, \mathbf{y}) \equiv \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2} \sqrt{\sum_{t=1}^T (y_t - \bar{y})^2}}, \quad (2.1)$$

where x_t denotes the time series' value at time t and $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$ signifies its expected mean. The correlation is normalized and bounded to the interval $[-1, 1]$ and thus allows direct comparisons across pairwise correlations. It is used in finance to evaluate the correlation between the returns of different financial instruments.

However, it is important to note that the Pearson correlation only measures linear relationships. Therefore, it might not be a good indicator of association if the relationship between the variables is nonlinear, or if the data does not meet the assumptions of normality and homoscedasticity [51].

Granger Causality

Named after Clive Granger, who introduced the concept in the 1960s, *Granger Causality* (GC) provides a quantitative framework to determine whether the past values of one variable provide valuable information for predicting future values of another variable [14]. It is based on the premise that if a variable X *Granger causes* another variable Y , the past values of X should contain useful information for predicting the future values of Y , beyond what is already captured by the lagged values of Y itself. In other words, X has a causal influence on Y , if including X 's history in a predictive model improves the forecast accuracy of Y compared to using only the past values of Y .

The process of estimating Granger causality comprises fitting autoregressive models and comparing the forecast accuracy of two models in competition: one model that includes the lagged values of potential causal variables, and another that contains solely the lagged values of the response variable. The preferred method for estimating GC is the *Vector Autoregression* (VAR) model [55]. In a VAR model, each variable is regressed on its own lagged values and lagged values of all other variables, capturing dependencies among them and facilitating causality evaluation. The GC test compares the residual variances of two models, one with the candidate causal variable(s) and the other without. A significant reduction in residual variance indicates the presence of GC.

Mathematically, the causality from X to Y can be tested by comparing the following two VAR models with p lags:

$$Y_t = \sum_{i=1}^p \beta_{1,i} Y_{t-i} + \epsilon_{1,t}, \quad (2.2)$$

$$Y_t = \sum_{i=1}^p \beta_{1,i} Y_{t-i} + \sum_{i=1}^p \beta_{2,i} X_{t-i} + \epsilon_{2,t}, \quad (2.3)$$

where $\epsilon_{1,t}$ and $\epsilon_{2,t}$ are the error terms of the first and second model, respectively. The null hypothesis is that the coefficients $\beta_{2,i}$ are jointly equal to zero, indicating no Granger causality from X to Y . Typically, the hypothesis can be tested using statistical tests such as the F -test [56]. To render the measure both continuous and normalized, we employ a specific normalization technique, expressed by

$$GC_{X \rightarrow Y} = 1 - \min \left\{ \left(\frac{RSS_2}{RSS_1} \right)^2, 1 \right\} \in [0, 1], \quad (2.4)$$

where RSS_i denotes the *Residual Sum of Squares* (RSS) corresponding to the two distinct models [57]. This equation ensures that the measure lies within the bounded interval $[0, 1]$, creating a consistent and standardized quantification.

Limitations and Considerations

GC, while a valuable tool, has some limitations that researchers must consider. First, it detects only linear causal relationships and may fail to capture nonlinear dependencies. Nonlinear interactions can lead to false negatives or positives in the analysis, as linear models may not adequately represent the underlying dynamics. Second, GC does not imply a direct cause-and-effect relationship but rather measures predictability [54]. Other factors, such as omitted variables or common drivers, can influence both the predictor and the response variable, leading to spurious results. Therefore, caution should be exercised in interpreting GC as definitive causal evidence [58].

Advancements in GC have focused on addressing its limitations and extending its applicability. Researchers have explored extensions of GC to incorporate nonlinear relationships by using nonlinear autoregressive models or kernel-based approaches. These advancements allow for capturing nonlinear interactions and detecting more complex causal relationships [59].

Transfer Entropy

Transfer Entropy (TE) is a powerful information-theoretic measure that has gained popularity in the field of causal inference, particularly in the analysis of time series data. After the demonstration of the equivalence between GC and TE, specifically for Gaussian variables as shown by Barnett, Barrett, & Seth [54], the metric introduced by Schreiber [13] has come to be broadly recognized as the information-theoretical extension of GC. It provides a way to quantify the directed flow of information between variables, which allows assessing causal relationships in a probabilistic framework. Mathematically, the TE from X to Y is defined as:

$$TE_{X \rightarrow Y} = \sum_{x_t, y_t, x_{t-k}} P(x_t, y_t, x_{t-k}) \log \left(\frac{P(y_t | y_{t-1}, x_{t-k})}{P(y_t | y_{t-1})} \right), \quad (2.5)$$

where $P(x_t, y_t, x_{t-k})$ represents the joint probability distribution of X_t , Y_t , and the past values of X (X_{t-k}). $P(y_t | y_{t-1}, x_{t-k})$ and $P(y_t | y_{t-1})$ denote the conditional probability distributions of Y_t given its past and the past of Y and X , respectively. The logarithm in the formula reflects the information gain. TE can also be expressed using joint and marginal entropies, which avoids directly referencing conditional entropy:

$$TE_{X \rightarrow Y} = H(Y_{t+1}, Y_t) + H(Y_t, X_t) - H(Y_{t+1}, Y_t, X_t) - H(Y_t), \quad (2.6)$$

where $H(Y_{t+1}, Y_t)$, $H(Y_t, X_t)$, $H(Y_{t+1}, Y_t, X_t)$, and $H(Y_t)$ are the joint and marginal entropies of the respective variables. To facilitate comparison between different estimations of TE, we employ the subsequent normalization:

$$TE_{X \rightarrow Y} = \frac{H(Y_{t+1}, Y_t) + H(Y_t, X_t) - H(Y_{t+1}, Y_t, X_t) - H(Y_t)}{\sqrt{H(Y_{t+1}, Y_t) \cdot H(X_{t+1}, X_t)}} \in [0, 1]. \quad (2.7)$$

This normalization approach stems from our understanding of TE as an asymmetric causal measure. This interpretation aligns with the concept of covariance, which, when rescaled, results in the normalized form as the Pearson correlation described in Equation 2.1. Thus, the above normalization ensures that TE can be meaningfully compared across different instances or applications, adhering to our theoretical framework.

Limitations and Considerations

Estimating TE from data entails the complex task of estimating underlying probability distributions, a challenge that intensifies with high-dimensional time series. Methods such as the nearest neighbor technique, kernel-based estimators, and Bayesian approaches have been suggested to address this issue [53]. Although histograms with equally distributed bins are commonly employed to estimate densities, Mynter [60] revealed that this strategy might introduce biases, as the estimation process is highly reliant on specific partitioning details. Consequently, securing a robust estimator is not straightforward. In our research, we found that using equally spaced bins produces satisfactory results. This bin configuration has been empirically validated by Baur & R ath [61], who employed it in constructing generalized local states in reservoir computing.

It is important to note that the application of TE can occasionally produce false causalities, depending on the conditioning dimension. Interpreting causal connections, identifying true nonlinear relationships, and the sensitivity of TE to various model parameters can be challenging. Additionally, further convolution of the measure’s applicability can be caused by its normalization, susceptibility to noise, and the presupposition of *Markovian* processes [62].

Convergent Cross Mapping

Convergent Cross Mapping (CCM) is a widely-used technique for causal inference in complex dynamical systems [15]. Its purpose is to uncover causal connections between variables by reconstructing their underlying dynamics. CCM is based on the premise that embedded variables with causal links will display similar dynamic behaviors, which is known as *shadowing*. At the core of this approach is Takens' theorem, a cornerstone result in dynamical systems theory. Floris Takens presented this theorem in 1981, establishing the mathematical basis for reconstructing a dynamical system using solely one state variable's observations [63]. With this approach, one can discern the geometric and dynamical characteristics of an attractor from a one-dimensional time series. This is incredibly valuable in practical applications where full state measurements are not feasible.

Theorem 1 (Takens' Theorem) *Consider a dynamical system with an attractor A embedded in \mathbb{R}^m , and let $f : \mathbb{R}^m \rightarrow \mathbb{R}^m$ be the system's evolution function. We observe the system through a function $h : \mathbb{R}^m \rightarrow \mathbb{R}$ that maps states of the system to real numbers. Given a time series $\{x_t\}$ of observations obtained from the system, the embedding of the system in a space of dimension $\kappa = 2m + 1$ can be done using delay coordinates:*

$$\mathbf{y}_t = (x_t, x_{t-\tau}, \dots, x_{t-(2m)\tau}),$$

where τ is the time delay. The theorem now states that, under certain generic conditions, an embedding map $\Phi : A \rightarrow \mathbb{R}^{2m+1}$ is given by:

$$\Phi(x) = (h(x), h(f(x)), \dots, h(f^{2m}(x))).$$

Thus, Φ is a diffeomorphism onto its image, and the attractor's topology in the state space is preserved in the embedded space.

After selecting suitable values for the embedding dimension and time delay, one can analyze the fundamental dynamics, forecast future behavior, and reveal concealed connections within the system. Thus, the CCM algorithm can be summarized as follows:

Algorithm 1 Convergent Cross Mapping

- 1: **Time Delay Embedding.** Embed the time series data of X and Y into higher-dimensional spaces using the embedding dimension κ and time delay τ .
 - 2: **Library Construction.** Create a library of vectors from the reconstructed state space \mathbf{X} , denoted as $\mathcal{L}_{\mathbf{X}}$, and a library of vectors from the reconstructed state space \mathbf{Y} , denoted as $\mathcal{L}_{\mathbf{Y}}$.
 - 3: **Nearest Neighbor Selection.** For each vector $\mathbf{X}(i)$ in the shadow manifold $\mathcal{M}_{\mathbf{X}}$, find its nearest neighbor in $\mathcal{M}_{\mathbf{Y}}$, denoted as $\mathbf{Y}(j)$. Similarly, for each vector $\mathbf{Y}(k)$ in $\mathcal{M}_{\mathbf{Y}}$, find its nearest neighbor in $\mathcal{M}_{\mathbf{X}}$, denoted as $\mathbf{X}(l)$.
 - 4: **Cross Mapping.** Assess the predictability of X based on Y by comparing the distances between the vector pairs $\mathbf{X}(i)$ and $\mathbf{Y}(j)$, and the vector pairs $\mathbf{Y}(k)$ and $\mathbf{X}(l)$. A statistical measure, such as the Pearson correlation ρ , can be used to quantify the predictability.
 - 5: **Convergence Analysis.** Repeat the cross mapping procedure for different library lengths. Evaluate the correlation as a function of the number of points used and assess the convergence of the results. The convergence of the cross mapping indicates the presence of a causal relationship between X and Y .
-

Convergence Analysis

In the typical application of CCM, convergence necessitates visual inspection. Consequently, we devise a more sophisticated method that uses rolling windows. For a vector of correlations ρ with a length of n , we compute the standard deviation in each window. To be considered convergent, the standard deviation must consistently decrease and eventually fall below a pre-defined threshold θ . If convergence is achieved, the average of the last s values is computed to reduce the impact of anomalies. On the other hand, if convergence is not attained, the causality measure in CCM is assigned a value of zero. This method is mathematically expressed as:

$$CCM_{X \rightarrow Y} \equiv \begin{cases} \frac{1}{n} \sum_{i=1}^s \rho_{n-s+i} & \text{if } \rho \text{ converges} \\ 0 & \text{otherwise} \end{cases} \in [-1, 1] . \quad (2.8)$$

This automated process facilitates the evaluation of CCM causality for various connections within a system at a reasonable speed. To normalize the measure to the interval $[0, 1]$ and render it comparable with other causal inference methods, the correlation distance, denoted as $d = \sqrt{2(1 - \rho)}$, can be employed.

Limitations and Considerations

CCM's ability to identify causal relationships within time series data is impacted by various factors. The outcomes can be altered by the presence of noise or missing values in the data [64]. Additionally, the intricate process of selecting suitable embedding dimensions κ and time delays τ depends on the dataset's particular characteristics [65]. For instance, the most suitable value for τ can be determined by identifying the initial local minimum in the mutual information regarding τ . Furthermore, the *False Nearest Neighbor* algorithm can assist in locating the tiniest embedding size that maintains the attractor's structure, ensuring that adjacent points in the original time series remain adjacent in the embedded version [66].

However, challenges exist within CCM. False positives and negatives may arise, particularly when dealing with obscured common drivers or confounding variables. The assessment of convergence in CCM is a complex task, and errors in this assessment can lead to misleading conclusions [67]. Moreover, the requirement for deterministic dynamics can sometimes limit the applicability of CCM [68]. Efforts to enhance the functionality of CCM have focused on improving its efficiency, reliability, and versatility [69]. Innovations encompass algorithms that decrease the computational burden, allowing CCM to analyze more complex and larger datasets. Enhancements in the adaptive selection of ideal embedding dimensions and time delays, as well as advanced techniques for noise reduction and missing data management, contribute to its robustness [64]. Recent research has facilitated the application of CCM to short time series, expanding its potential use in various domains [70].

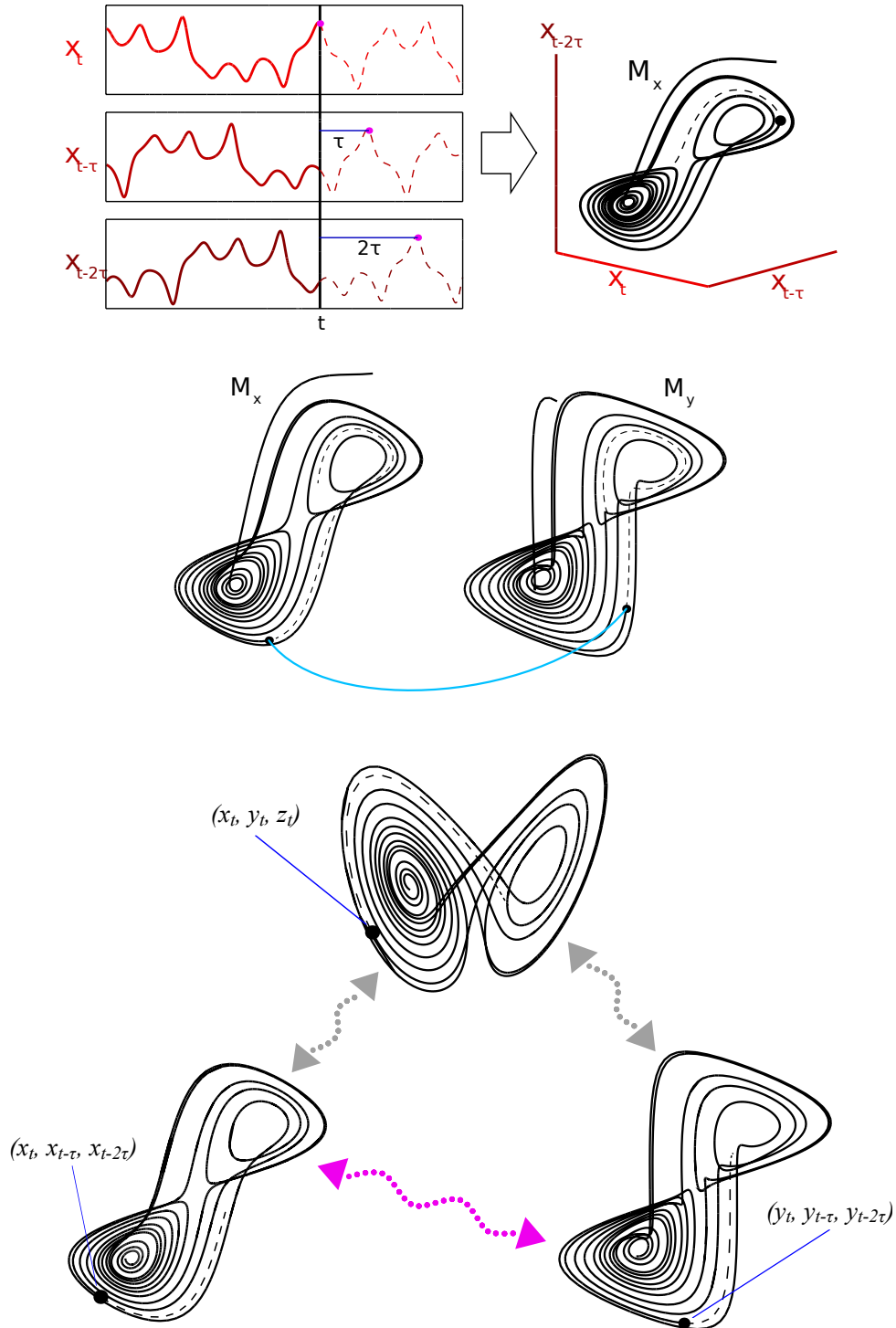


Figure 2.3: Convergent Cross Mapping. The graphs depict different steps of the CCM inference of the Lorenz system. The top graph represents the initial phase where the x -coordinate is embedded in a two-dimensional space using an embedding dimension of $\kappa = 2$ and a time delay of τ . This embedding creates a reconstructed state space that captures the underlying dynamics of the original system. In the middle graph, the focus shifts to the shadow manifolds of coordinates X and Y . Here, the process of selecting the nearest neighbor is illustrated. The bottom graph demonstrates the cross mapping between the reconstructions of the Lorenz attractor. This is where the cross mapping technique is applied to the reconstructed state space of the coordinates, linking the dynamics of one variable to the other. Adapted from *rEDM: An R package for Empirical Dynamic Modeling and Convergent Cross Mapping* [71].

Fourier Transform Surrogates

In our investigation of the causal structure of time series systems, we utilize *Fourier Transform* (FT) surrogates to separate the impact of linear and nonlinear drivers. FT surrogates are a widely recognized tool for generating surrogate data that preserves the linear properties of the original time series, including its power spectrum and amplitude distribution, while eliminating nonlinear properties [17]. This methodology has been widely used in nonlinear data analysis, especially to examine nonlinearity hypotheses. The construction of FT surrogates has been detailed in previous literature [72]:

Algorithm 2 Fourier Transform Surrogates

- 1: **Fourier Transform.** Given a real-valued time series $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, compute its Fourier transform $\mathbf{F}(\mathbf{x})$ using the *Fast Fourier Transform* (FFT) algorithm [73].

$$\mathbf{F}(\mathbf{x}) = FFT(\mathbf{x}).$$

- 2: **Phase Randomization.** Preserve the amplitudes but randomize the phases of the Fourier coefficients. This can be done by multiplying the complex Fourier coefficients by a random phase factor $e^{i\phi}$, where ϕ is uniformly distributed over the interval $[0, 2\pi]$. The phase-randomized Fourier Transform $\mathbf{F}'(\mathbf{x})$ is given by:

$$\mathbf{F}'_k = |\mathbf{F}_k| \cdot e^{i\phi_k}, \quad \phi_k \in [0, 2\pi].$$

- 3: **Inverse Fourier Transform.** Compute the inverse FT of the phase-randomized coefficients to obtain the surrogate time series $\tilde{\mathbf{x}}$:

$$\tilde{\mathbf{x}} = IFFT(\mathbf{F}'(\mathbf{x})).$$

By keeping the amplitudes of the original data and only randomizing the phases, the resulting surrogates maintain the power spectral density of the original time series but break the higher-order statistical dependencies.

This method is often used to test the null hypothesis that the data is linearly coupled, as any significant deviation in a nonlinear measure between the original time series and the FT surrogates can be an indicator of nonlinearity.

Linear and Nonlinear Measures

In order to evaluate how much of a causal measure is attributed to linear or nonlinear effects, we adopt a specific approach that involves the calculation of measures on surrogate time series. The surrogate of time series \mathbf{x} , when subjected to the random phases of realization k , is denoted as $\tilde{\mathbf{x}}^{(k)}$. Within this research, we focus on bivariate measures $\psi(\mathbf{x}, \mathbf{y})$, which quantify the relationship between two time series. To enhance the reliability of our findings, we average metrics derived from surrogate time series over various instances K of random phases. The corresponding surrogate or linear measure is defined as the average over K surrogate realizations of both time series:

$$\tilde{\psi}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\tilde{\mathbf{x}}^{(k)}, \tilde{\mathbf{y}}^{(k)}). \quad (2.9)$$

Here, the superscript k indicates adding identical random phases to both time series in a single realization. This method ensures that phase differences are undisturbed, maintaining specific properties such as the Pearson correlation [74].

To further examine the nonlinear aspects of the measure, we compute the discrepancy between the primary measure and its linear surrogate counterpart:

$$\psi^{nl} \equiv \psi - \tilde{\psi}. \quad (2.10)$$

However, to avoid potential spurious effects leading to negative nonlinearities, we recommend implementing the following measure:

$$\max \left\{ 0, \psi - \tilde{\psi} \right\}. \quad (2.11)$$

This method enables us to differentiate the nonlinear aspect of the measurement in a comprehensible way. Moreover, we establish the cross-measure by merely substituting the first time series:

$$\psi^{cross}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\tilde{\mathbf{x}}^{(k)}, \mathbf{y}), \quad (2.12)$$

and analogously define the reverse as the anti-measure:

$$\psi^{anti}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\mathbf{x}, \tilde{\mathbf{y}}^{(k)}). \quad (2.13)$$

The intuition behind the cross- and anti-measure is to analyze the influence of the linear part of x on y under the measure ψ and vice versa.

Nested Measures

When shifting to dynamic analysis using rolling windows, the measure transforms into a vector, ψ , which offers new possibilities for assessing nonlinearity. This shift permits the exploration of co-dependence between two measurements with the aid of a third, nested measure ψ_{nest} .

$$\psi_{nest} \equiv \rho(\psi_1, \psi_2). \quad (2.14)$$

Particularly, the Pearson correlation coefficient ρ can be utilized to examine the correlation between the original measure and its corresponding surrogate, stated as:

$$\rho(\psi, \tilde{\psi}). \quad (2.15)$$

This method also allows for the depiction of the determination coefficient using the Pearson correlation, as referenced in [75]:

$$R^2 = \rho^2 \in [0, 1]. \quad (2.16)$$

This equation allows us to measure the contribution of linear influences, specifically the portion of the variance in the measure ψ that can be predicted by the surrogate measure $\tilde{\psi}$. Any remaining variance arises from nonlinear characteristics:

$$1 - \rho^2(\psi, \tilde{\psi}). \quad (2.17)$$

Furthermore, there is an exploration application of the correlation-causality fallacy [76]. This application entails calculating the proportion of causality explained by correlation:

$$\rho^2(\psi, \rho), \quad (2.18)$$

serving as a gauge of the causal relationship that can be explained by correlation.

2.3 Financial Data and Frameworks

After developing methodologies to quantify linear and nonlinear causality, our focus shifts to their practical application. Here, we utilize datasets from the German and U.S. stock exchanges and implement two financial frameworks that depend on the interdependence of financial instruments. By incorporating causal inference and decomposition, we demonstrate the straightforward integration of these methods and their contribution to improved performance outcomes. Our approach highlights the ease of integrating causality into current financial models, enabling informed decision-making in the financial industry and potentially leading to substantial improvements in strategy outcomes and risk management.

Financial Data

For our real-world analysis, we select a subset of stocks from the DAX and Dow-Jones indices, which represent the 30 most capitalized and thus most influential companies in Germany and the U.S., respectively. Starting on January 19, 1973, our data consists of the daily closing prices of all stocks included in the index through April 20, 2022, to provide a consistent universe of stocks over the entire period. This results in a total of $N_{DAX} = 11$ and $N_{DJ} = 17$ time series with 12785 data points. Note that the survival bias [77] is negligible for our analysis. To ensure stationary time series, we convert the stock prices p_t into logarithmic returns using:

$$x_t = \log p_t - \log p_{t-1}. \quad (2.19)$$

The data’s time span is sufficient to analyze significant market occurrences, commencing with the worldwide economic recession in the beginning of the 1980s and encompassing *Black Monday* on October 19, 1987, when global stock markets collapsed for the first time after World War II. From 1997 to 2001, excessive speculations occurred in the markets, and numerous technology companies were overvalued, leading to the *dotcom bubble* [78]. The market experienced significant price declines in July and September 2002 due to the bubble burst. Our data also covers the 2007/2008 subprime mortgage crisis, which led to a market decline from its all-time high in October 2007 before ultimately crashing following the collapse of *Lehman Brothers* on September 15, 2008. Between 2015 and 2016, investors sold equities globally due to a combination of slowing GDP growth in China and the Greek debt default. The dataset also encompasses the event known as *Volmageddon* on February 5, 2018, when a sizable sell-off in the American stock market caused a surge in implied market volatility [79]. Lastly, the data incorporates the effects of the COVID-19 pandemic, which, among other consequences, instigated an unexpected worldwide stock market collapse on February 20, 2020. Our time period also includes a number of important global political events. These include the fall of the *Berlin Wall* on November 9, 1989, triggering the collapse of the Soviet Union, the attacks of September 11, 2001, and the Russian invasion of Ukraine on February 24, 2022.

Rolling Windows

To obtain dynamically evolving results, we divide the data into overlapping rolling windows [80] and compute our measures for each interval following the approach of Haluszczyński, Laut, Modest, et al. [42]. We use a sliding window of $T_w = 1000$ trading days, which corresponds to about four years of data. The gap or step between successive intervals is set to $\delta T = 20$ trading days, roughly equivalent to one month. Thus, the w -th interval is displayed as:

$$\mathbf{x}^{(w)} = (x_{1+(w-1)\cdot\delta T}, \dots, x_{T_w+(w-1)\cdot\delta T}), \quad (2.20)$$

which gives a total of $w = 594$ overlapping windows. A (causality) measure $\psi(\mathbf{x}, \mathbf{y}) \mapsto \mathbb{R}$, which maps two time series to a real number, is thus transformed into a vector.

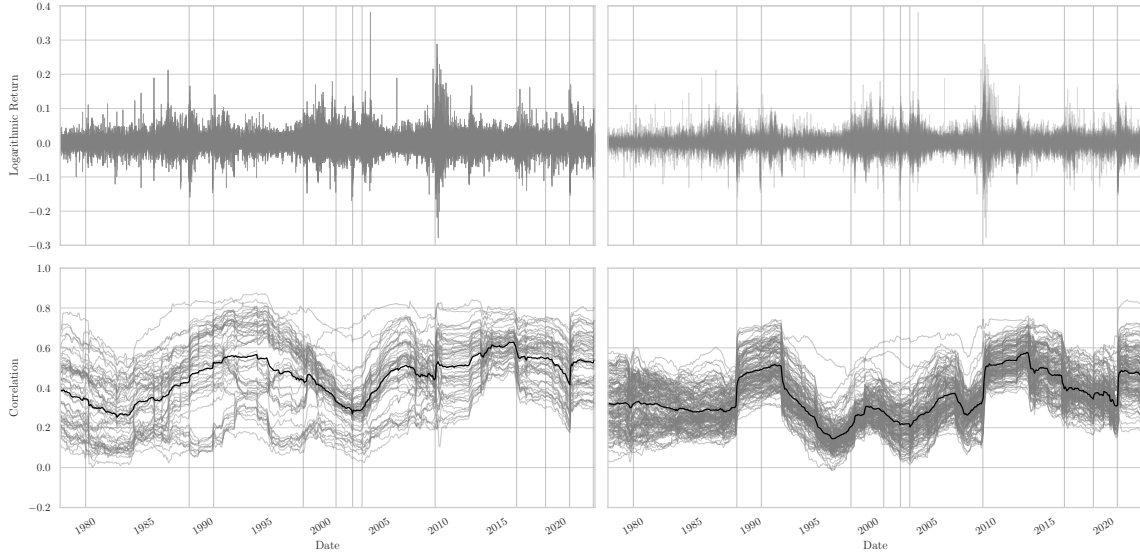


Figure 2.4: Historical Stock Returns and Correlation. The top row shows the logarithmic returns of the historical stock data of the German DAX (left) and the U.S. Dow-Jones (right) index, respectively. Each line represents the logarithmic return of one stock over time. The bottom row shows the pairwise correlations between the stocks. Each line displays the rolling correlation between two stocks over time, with the black line depicting the average correlation across all stocks. The vertical lines represent important economic or political events.

Pair Trading

Pair trading is a popular and widely used strategy in quantitative finance that seeks to capitalize on relative price movements between two highly correlated assets [81]. The strategy is based on the concept of mean reversion, which holds that the prices of assets that are historically correlated tend to revert to their average historical relationship over time. When they deviate from this correlation (i.e., one stock goes up while the other goes down or vice versa), we take a *long* position in the underperforming stock and a *short* position in the outperforming stock, expecting them to return to their historical correlation [82]. Going long on a position means expecting rising returns, whereas going short on a position means expecting falling returns. Thus, a basic form of the strategy involves the following steps:

Algorithm 3 Pair Trading

- 1: **Correlation Calculation.** We calculate the rolling historical and the short-term correlation between two stocks.
- 2: **Signal Generation.** When the current correlation ρ_t deviates from its historical mean $\bar{\rho}_{hist}$ by a certain threshold, a trading signal is generated. A common approach is to use the *z-score* of the difference, which measures the number of standard deviations by which the current correlation deviates from its historical mean:

$$z_t = \frac{\rho_t - \bar{\rho}_{hist}}{\sigma_{\rho_{hist}}}, \quad (2.21)$$

where $\bar{\rho}_{hist}$ and $\sigma_{\rho_{hist}}$ denote the mean and standard deviation of the historical correlation, respectively.

- 3: **Trade Execution.** When the z-score surpasses a preset threshold value (e.g., above a positive threshold for a long trade or below a negative threshold for a short trade), a trade is triggered. Buying an underperforming asset while simultaneously shorting an overperforming asset is referred to as a long trade.
-

Portfolio Optimization

In finance, the *Markowitz Portfolio Theory* (MPT) serves as a fundamental concept for investors and financial analysts [83]. Created by Harry Markowitz in 1952, the theory transformed the approach to portfolio creation. Its basic tenet posits that rational investors aim to maximize potential returns while minimizing risks. The evaluation of an asset's risk and return should include the context of the entire portfolio, as opposed to evaluation in isolation. Before delving into portfolio optimization technicalities, we first introduce fundamental portfolio metrics:

- *Expected Return*: the expected return of a portfolio $E(R_p)$ is calculated by taking a weighted sum of the expected returns of its individual assets.

$$E(R_p) = \sum_{i=1}^n w_i \cdot E(R_i), \quad (2.22)$$

where w_i is the weight of asset i in the portfolio, and $E(R_i)$ denotes the expected return of asset i . Although historical returns do not guarantee future performance, it is common to use a historical mean as an estimation for the expected returns [82].

- *Variance*: the portfolio variance σ_p^2 measures the risk of a portfolio, taking into account individual asset variances and their correlations. The formula for portfolio variance is:

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i \cdot w_j \cdot \sigma_i \cdot \sigma_j \cdot \rho_{ij}, \quad (2.23)$$

where w_i and w_j indicate the weights of assets i and j in the portfolio, and σ_{ij} indicates the covariance between assets i and j . Correlation can be substituted with a causality measure ψ , or if the normalized measure ψ falls within the range of $[0, 1]$, the sign of the correlation can be utilized:

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i \cdot w_j \cdot \sigma_i \cdot \sigma_j \cdot \psi_{ij} \cdot \text{sgn}(\rho_{ij}), \quad (2.24)$$

where $\text{sgn}(\cdot)$ denotes the *signum* function.

- *Sharpe Ratio*: the *Sharpe Ratio* S is a measure to calculate the risk-adjusted return of an investment portfolio [84]. Developed by William F. Sharpe, this metric helps investors understand the return of an investment compared to its risk:

$$S = \frac{E(R_p - R_f)}{\sigma_p}, \quad (2.25)$$

where R_f is the risk-free rate of return. The Sharpe ratio proves particularly valuable because it offers a straightforward measurement of the excess return in relation to risk.

- *Value-at-Risk*: a commonly used method to assess the risk of historical portfolio performance is the *Value-at-Risk* (VaR). It calculates the possible loss in value of an investment or portfolio within a set time frame at a specific α confidence level [85]. A $1-\alpha$ VaR of x implies that there is a probability of α that the portfolio will incur losses greater than x . Unlike portfolio variance, VaR measures tail risk without assuming a normal distribution, making it particularly important for risk management purposes. We utilize the default α value of 1%.

Two portfolios of significant importance in MPT are the *Minimum Risk* and *Maximum Sharpe Ratio* Portfolios, each serving a distinct purpose in the investment strategy aligned with the investor’s risk tolerance. The Minimum Risk Portfolio aims to preserve capital and maintain stability, while the Maximum Sharpe Ratio Portfolio seeks to optimize the trade-off between risk and return. Together, these principles underpin modern portfolio theory, aiding investors in making knowledgeable choices that align with their financial objectives and risk tolerances.

- *Minimum Risk*: this portfolio is crucial for risk-averse investors seeking to minimize their risk exposure and earn satisfactory investment returns. For those who prioritize security over higher gains, this portfolio comprises assets that, when combined, yield the least possible variance based on historical data. This optimization problem’s solution yields the (long-only) weights of the assets included in the portfolio:

$$\begin{aligned} & \text{Minimize} && \sigma_p^2 \\ & \text{Subject to} && E(R_p) = \text{target return} \\ & && \sum_{i=1}^n w_i = 1 \\ & && w_i \geq 0 \quad \text{for all } i. \end{aligned}$$

It serves as a reference point for assessing risk in other portfolios. If a portfolio has a greater anticipated return while maintaining the same level of risk as the Minimum Risk portfolio, it may be deemed more efficient.

- *Maximum Sharpe Ratio*: also known as the *Tangency Portfolio*, this portfolio is crucial for investors seeking the most efficient return for the level of risk they are willing to accept. It is considered optimal in a risk-adjusted sense because it maximizes the excess return for every unit of risk taken and serves as a guide for investors to allocate their capital in a way that compensates them most generously for the risks they endure:

$$\begin{aligned} & \text{Maximize} && S \\ & \text{Subject to} && \sum_{i=1}^n w_i = 1 \\ & && w_i \geq 0 \quad \text{for all } i. \end{aligned}$$

The Maximum Sharpe Ratio is often used as a reference for the market portfolio in the *Capital Asset Pricing Model* (CAPM), under the assumption that all investors will choose a mix of the risk-free asset and the market portfolio according to their risk preference [86].

We present a straightforward method for incorporating causality measures into portfolio construction using these two portfolios. We regularly optimize the portfolio’s weightings to align with current market conditions. This is accomplished by implementing the rolling causality measures mentioned previously. Consequently, the benefits of using causality measures as the co-dependency metric for the portfolio can be assessed, taking into account its impact on both performance and risk management.

2.4 Framework Validation and Application Results

After establishing the necessary methods for our framework, we present the results of our analysis in the following Section. As was motivated by Figure 2.2 above, we observe that for complex and chaotic systems it is difficult to measure the co-dependence of variables using correlation, since they can exhibit different regimes of positive, negative, and no correlation, even though they are governed by exactly the same governing equations. The rolling window analysis of the correlation is unrobust and changes significantly over time, illustrating the need for a different measure to reliably gauge co-dependence. Causality measures, such as TE and CCM, are a valuable technique for measuring causality in both directions and offer consistent results over time. Moreover, FT surrogates allow for the separation of causality into linear and nonlinear contributions, facilitating comprehension of the complex co-dependence. Our analysis and major findings are divided into two parts:

- *Synthetic Systems*: we first apply this approach to synthetic systems and demonstrate that a considerable amount of nonlinearity drives the causality in the Lorenz and Halvorsen systems. Although GC can exclusively detect linear causality, TE and CCM suggest that nonlinear properties significantly determine causality. Our findings reveal that, in the Lorenz and Halvorsen systems, the contribution of nonlinearity remains independent of the strength of the nonlinear coupling.
- *Financial Markets*: after validating our methods on synthetic systems, we demonstrate their application to financial markets, revealing noteworthy TE and CCM resulting from nonlinear features in both the German and U.S. stock markets. Additionally, we observe the presence of the correlation-causality fallacy since the Pearson correlation frequently serves as a viable proxy for linear causality in the financial industry. However, investors may underestimate the causality within their portfolio and potentially overlook portfolio risk by neglecting nonlinear causality. We demonstrate significant benefits of pair trading and portfolio optimization when replacing correlation with causality.

Decomposing Causality in Synthetic Systems

Our examination of the Lorenz and Halvorsen systems, which are described in Equations 50 and 52, reveals that causality is primarily driven by nonlinear properties. Figure 2.5 illustrates this observation, where the *box plots* [87] demonstrate that all surrogate-based causalities measured by TE and CCM are significantly lower than the original causality. This discrepancy arises because the surrogate time series only retain linear properties, while nonlinear effects are eliminated.

Consequently, we observe that a substantial portion of TE and CCM can be attributed to nonlinear properties. As anticipated, we confirm that GC accurately measures linear causality, as both the original and surrogate GC are on the same scale. Any slight deviations are due to the inaccuracies in the linear regression required for GC calculation. In addition, we analyze anti- and cross-causalities, measuring the causal flow from the linear properties of one time series to both the linear and nonlinear properties of another. We find that these causalities vanish for all three inference methods, further indicating that the causal flows are primarily dominated by nonlinearity.

To validate that our method only measures linear and nonlinear causality when the governing equations are fully linear and nonlinear, we conduct the analysis for two dummy models given in Equations 54 and 55. Figure 2.6 demonstrates that our methods are valid. The fully linear model predominantly exhibits linear causality, as GC is significant, and the original and surrogate TE and CCM have similar strengths. For the fully nonlinear model, we observe the opposite case, where GC is low, and the surrogate TE and CCM are significantly lower than the original TE and CCM.

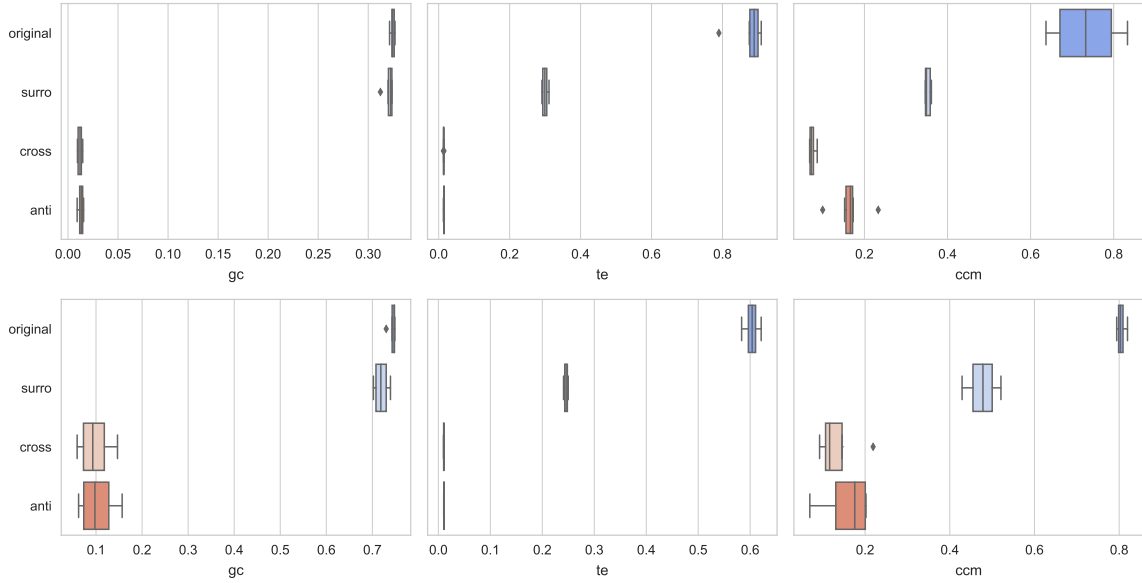


Figure 2.5: Causality Box Plots of Chaotic Systems. For both the Lorenz (upper row) and Halvorsen (lower row) systems, the mean of the original, surrogate, cross, and anti-matrices were calculated for GC, TE, and CCM (left to right). The sample size was 50 simulations using the standard configuration. Surrogate causalities were averaged over 10 surrogate realizations. Lozenge symbols were utilized to mark outliers based on the *Interquartile Range* (IQR) [88].

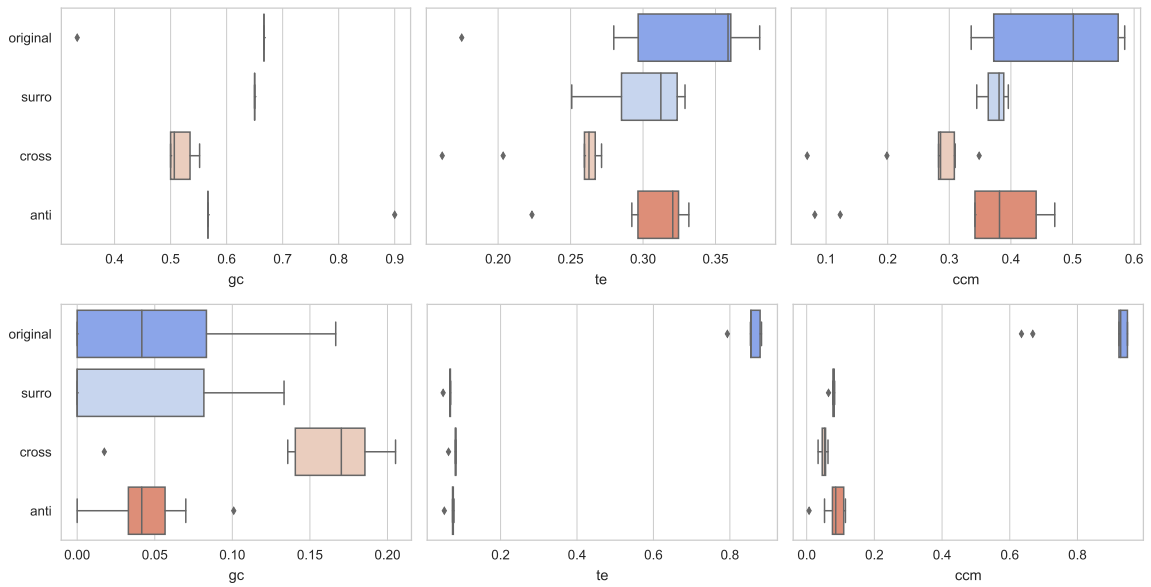


Figure 2.6: Causality Box Plots of Dummy Systems. The systems analyzed in this Figure are the fully linear (top row) and nonlinear (bottom row) systems. The configuration is analogous to Figure 2.5.

Variation of Nonlinearity Strength

To investigate whether causality can be consistently attributed to nonlinearity, we add parameters to the Lorenz and Halvorsen systems to vary the strength of the nonlinear terms in the governing equations. The modified governing equations are outlined in Equations 51 and 53.

We analyze variations in the degree of nonlinearity in the Lorenz and Halvorsen attractors. While both systems diverge for nonlinearity degrees that are less than or equal to 0, the upper bounds are arbitrarily chosen as we do not observe any significant changes to the attractor form. We conclude that the level of nonlinearity only affects the scale of the attractors.

This behavior translates directly to the causality analysis, as shown in Figure 2.7 for the Lorenz system. The original causality is notably greater than the surrogate causality for both TE and CCM, across all degrees of nonlinearity. Additionally, the grids display no discernible gradient, which suggests that causality is not influenced by the degree of nonlinearity. We find similar results for the Halvorsen system.

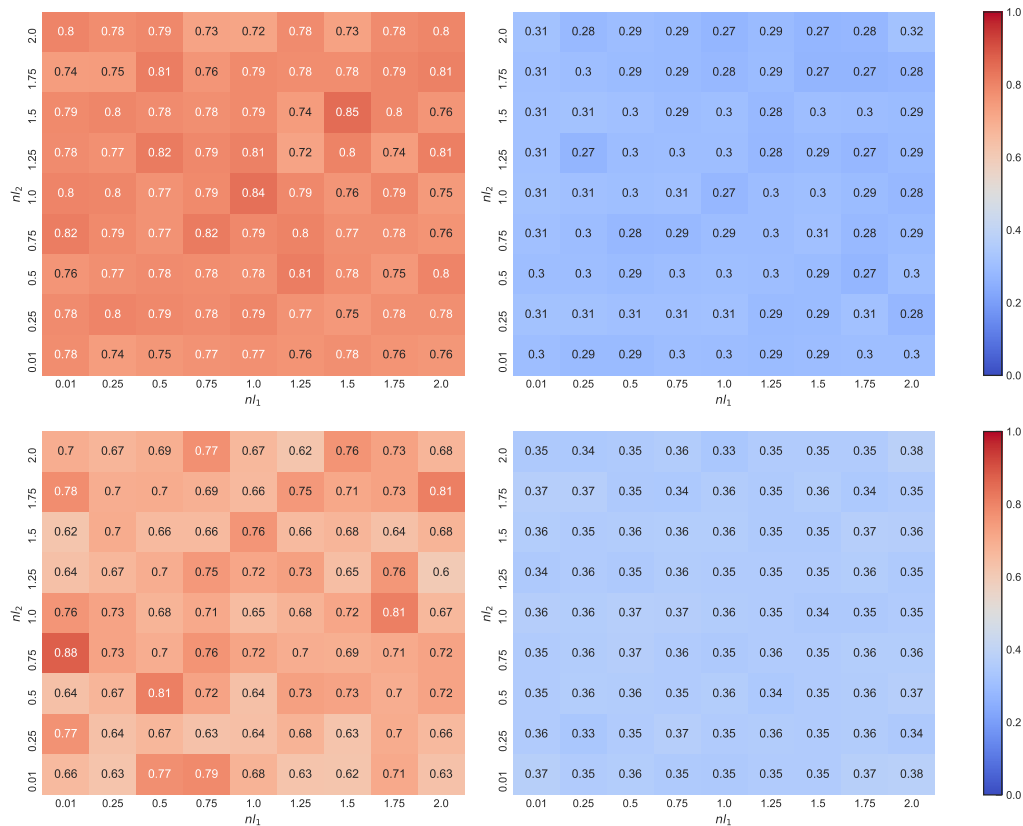


Figure 2.7: Causality for Different Nonlinearity Strengths. This Figure displays TE (top row) and CCM (bottom row) of the Lorenz attractor for different degrees of nonlinearity. We analyze the causalities for changes of the additional nonlinearity parameters between 0.01 and 2 respectively. The left grid illustrates the original causality, whereas the right grid presents surrogate causality. All grid entries are the average of 50 simulations, and the surrogate-causalities are averaged over 10 surrogate realizations.

Furthermore, we find that TE and CCM are effective in identifying nonlinearity but not measuring the strength of linear and nonlinear causal connections if the underlying dynamics are too similar. This discovery is of consequence for Chapter 3, where we discuss how to derive the governing equations from the underlying causal structure. Since the strength of the causal connection cannot be reliably detected, an algorithm based on the synchronization of chaos is used to calibrate the parameters of the governing equations.

Historical Causality in Financial Markets

After establishing the methods, we apply them to financial data from the German and U.S. stock markets. The data and the rolling correlations are visually depicted in Figure 2.4 above. Notably, these correlations undergo significant shifts during and after pivotal economic and political events. This phenomenon can be attributed to the changing behavior of investors and other market participants in response to these impactful occurrences. Furthermore, this effect extends to our investigation of causality measures, as demonstrated in Figures 2.8 and 2.8. These Figures reveal that linear and nonlinear causality measures, such as TE and CCM, exhibit analogous responses to these events.

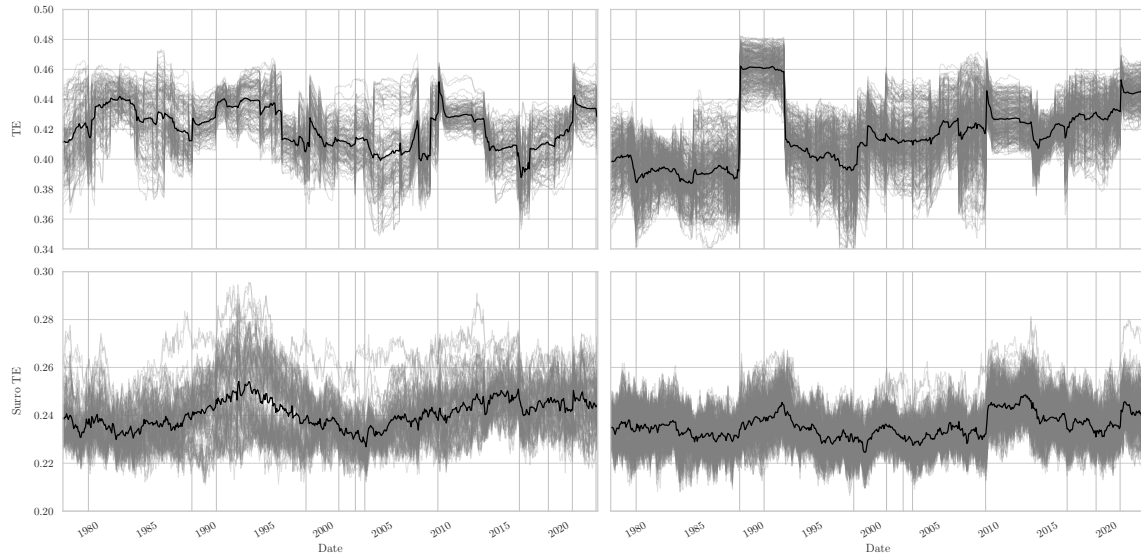


Figure 2.8: Transfer Entropy. This Figure shows the historical TE of stocks within the German DAX (left) and the U.S. Dow-Jones (right) indices, respectively. Each gray line displays the rolling TE between two stocks over time, with the black line depicting the average correlation across all stocks. The bottom row illustrates the corresponding surrogate TE averaged over 50 realizations. The vertical lines represent important economic or political events.

When examining TE, it is evident that TE displays sharp fluctuations in response to events while surrogate TE remains relatively stable and less reactive. In contrast, surrogate CCM exhibits a stronger response than regular CCM, with significant jumps similar to the observed correlation patterns. During these events, there were significant increases in correlation, TE, and surrogate CCM, particularly among U.S. stocks. Three major events that demonstrate this behavior are Black Monday in 1987, the global financial crisis in 2009, and the COVID-19 pandemic in 2020.

These observations indicate that the events caused significant changes in the market structure, which is understandable given their profound influence on the global economy. A notable finding is that TE demonstrates more substantial fluctuations than surrogate TE during these events, whereas the opposite is seen for CCM. This implies that the stock market's linear dynamics were significantly influenced, potentially because investors simultaneously adjusted their stock positions to respond to the crashes.

To determine the extent of nonlinear contributions to our causality measures, we analyze how much of the causality can be explained by its surrogate. Therefore, we employ the nonlinearity measure as defined in Equation 2.17. We evaluate the extent to which the linear properties account for the variation in the original causality using the squared Pearson correlation. Figure 2.10 illustrates the development of nonlinear causality over time, indicating that nonlinear TE and CCM demonstrate comparable yet not identical patterns. Both measures indicate increased nonlinearity during the period between the burst of the dotcom bubble and the start of the global financial crisis. However, we noticed periods of lower nonlinearity before and after this time frame.

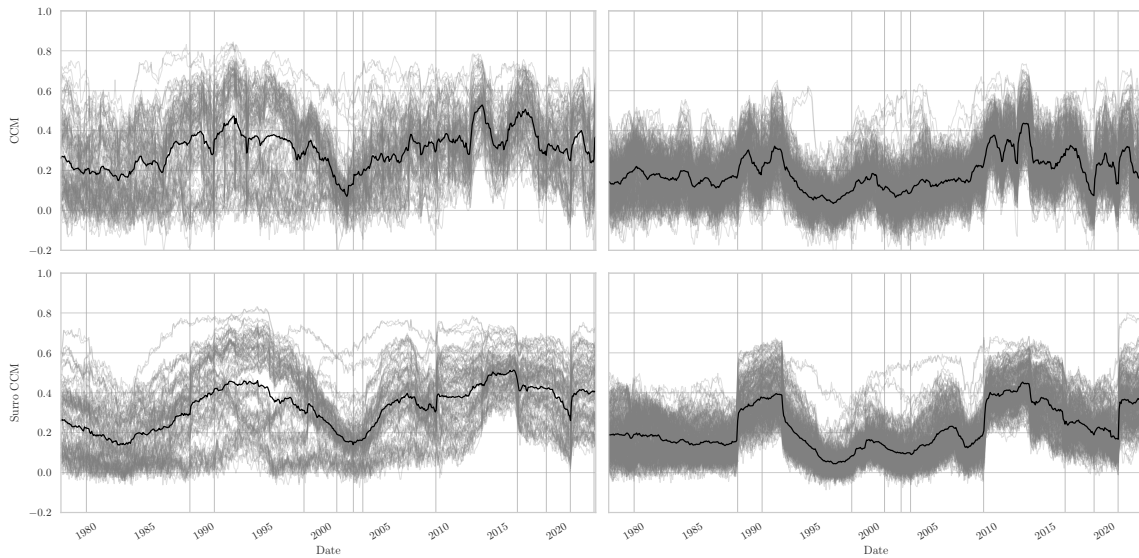


Figure 2.9: Convergent Cross Mapping. The configuration is analogous to Figure 2.8.

These two significant economic events need to be evaluated differently as the dotcom bubble caused more nonlinearity in its aftermath. In contrast, the global financial crisis, initiated by the American housing market crisis, brought about a period of more linear market behavior. The CCM illustrates this behavior drastically, with jumps exceeding 20%. In conclusion, our analysis indicates that nonlinear causality presents a beneficial resource for anticipating and evaluating financial effects, contingent upon continuous monitoring and evaluation within the context of changing market dynamics.

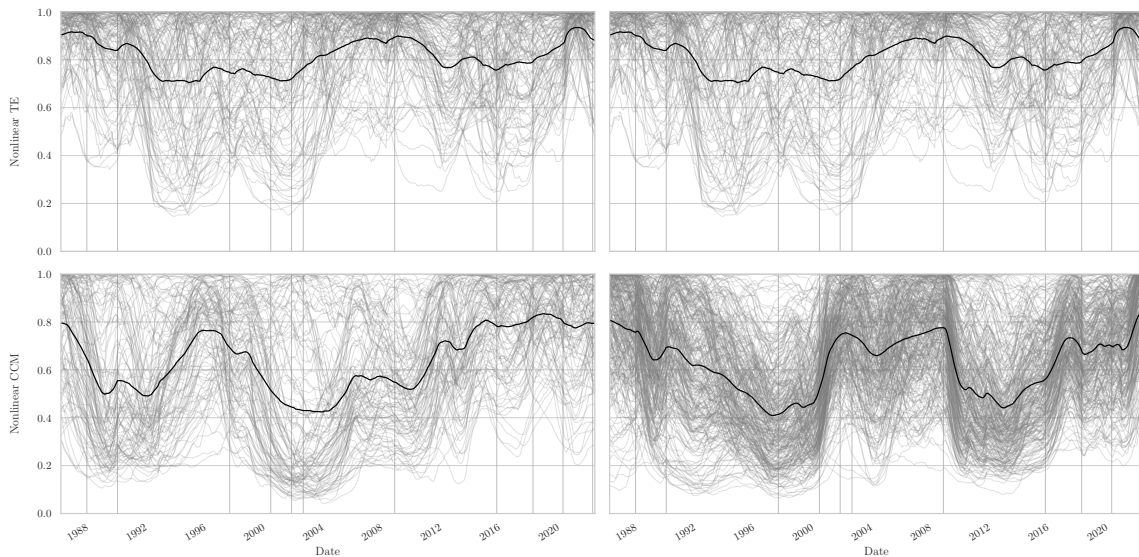


Figure 2.10: Nonlinear Causality. The top row displays the historical nonlinear TE of stocks in the German DAX (on the left) and the U.S. Dow-Jones (on the right) indices. The bottom row presents the nonlinear CCM. The configuration is analogous to Figure 2.8

Correlation-Causality Fallacy

Upon examination of Figure 2.11, it becomes evident that both the original and surrogate TE exhibit a moderate correlation. Notably, there is an intriguing exception during the period spanning from approximately 1990 to 2002 in the U.S. stock market, where a substantial portion, approximately 75%, of TE can be attributed to correlation. This spike coincided with the rise and eventual burst of the dotcom bubble, suggesting that it might have served as an indicator of abnormal market behavior during this period.

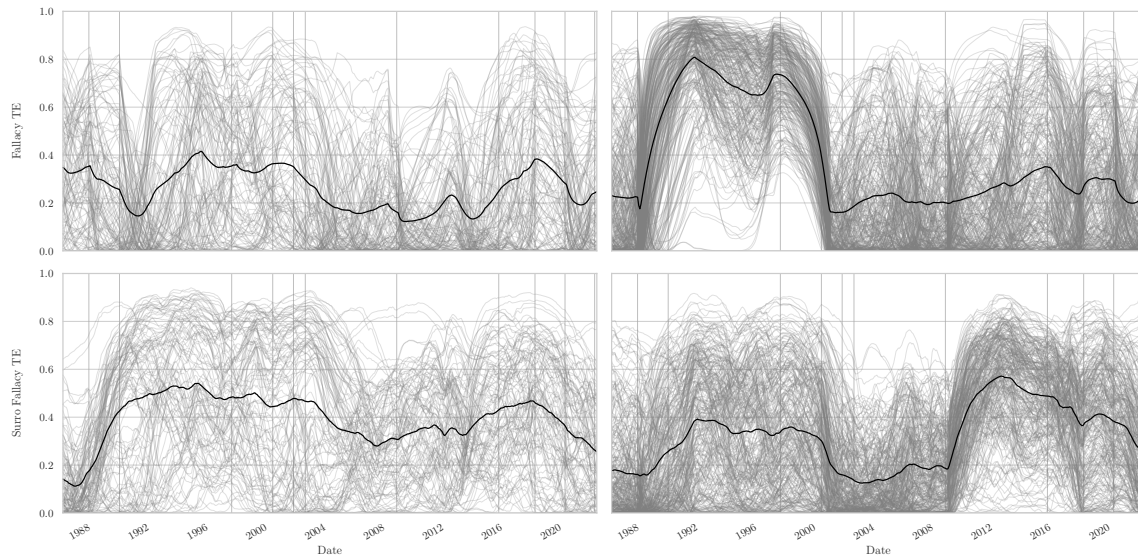


Figure 2.11: Fallacy Transfer Entropy. The configuration is analogous to Figure 2.8.

One of the most significant findings from this analysis is the observation that fallacy surrogate CCM is remarkably high, around 90%, in both the German and U.S. stock indices, as depicted in Figure 2.12. This suggests that correlation effectively acted as a suitable proxy for linear causality for the majority of the past few decades. However, in periods where this fallacy diminishes, such as the aftermath of the dotcom bubble in 2002 and the onset of the global financial crisis in 2008, relying solely on correlation as a measure of co-dependence significantly underestimates portfolio risk, as nonlinear effects cannot be disregarded.

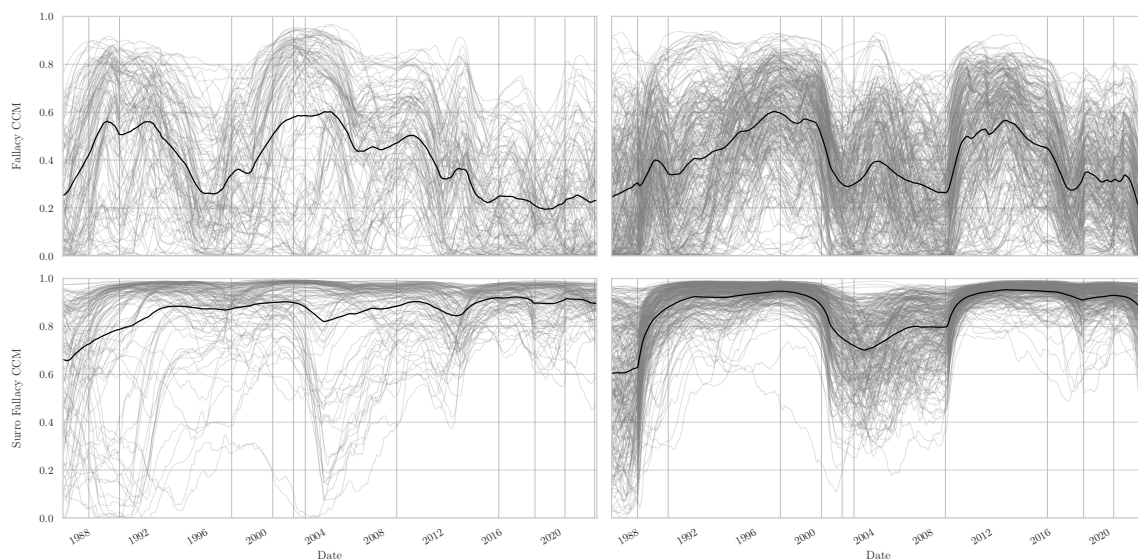


Figure 2.12: Fallacy Convergent Cross Mapping. The configuration is analogous to Figure 2.11.

Pair Trading and Portfolio Optimization

To effectively apply causality measures in practical financial scenarios, we introduce two popular financial frameworks that rely on the interdependence between assets. Our first concept explores pair trading, a logical choice based on the premise that two assets typically revert to a default correlation, and deviations from this norm can be profitable.

In Figure 2.13, we illustrate the seamless integration of causality measures using two chemical industry stocks from Germany, namely Bayer and BASF. Notably, although the co-dependence measures' evolution differences are relatively similar, the subtle distinctions over time have a significant impact on trading performance. Of particular interest is the finding that using surrogate CCM as a trading strategy yields significantly higher returns, approximately six times more than using correlation, despite the measures' apparent similarity. Moreover, both TE and CCM perform better than correlation, while surrogate TE underperforms and even results in negative returns. This clear example highlights the potential of employing a causality-based pair trading strategy.

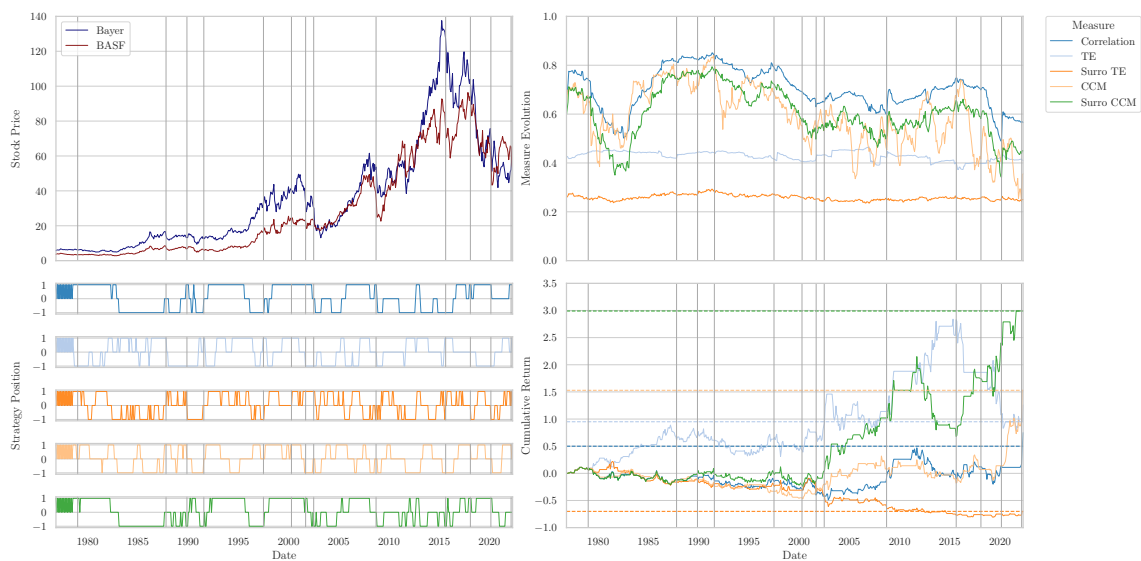


Figure 2.13: Pair Trading. The stock prices of two companies from the DAX (Bayer and BASF) are displayed in the top left graph. The top right graph presents the co-dependence measures over time, with each color corresponding to a specific co-dependence measure that is included in the legend on the right-hand side. The bottom left graph illustrates the strategy positions over time, with long position in Bayer and short position in BASF indicated by 1, the opposite indicated by -1 , and no investment indicated by 0. The graph in the lower right corner illustrates the cumulative return achieved by the strategy over time. The dotted horizontal lines mark the strategy's most recent cumulative return value. The vertical lines indicate notable economic or political events.

As previously highlighted, relying solely on correlation can potentially lead to an underestimation of risk, a perilous scenario when managing a portfolio. In Figure 2.14, we employ stocks from the U.S. Dow-Jones index and minimize risk by dynamically optimizing the portfolio weights on a monthly basis. It becomes evident that the allocations of a portfolio using correlation and CCM exhibit visible disparities over time. This divergence is reflected in the portfolio's downside returns and overall performance. Notably, we observe that a portfolio employing surrogate TE, CCM, and surrogate CCM achieves a superior 1% VaR while slightly enhancing portfolio performance.

Similarly, in the context of optimizing the Sharpe ratio, as depicted in Figure 2.15, the inclusion of causality measures results in a more favorable risk-return profile. When optimizing the stocks of the German DAX index, we note a reduction in portfolio standard deviation and an increase in portfolio value over time, particularly when employing original and surrogate CCM.

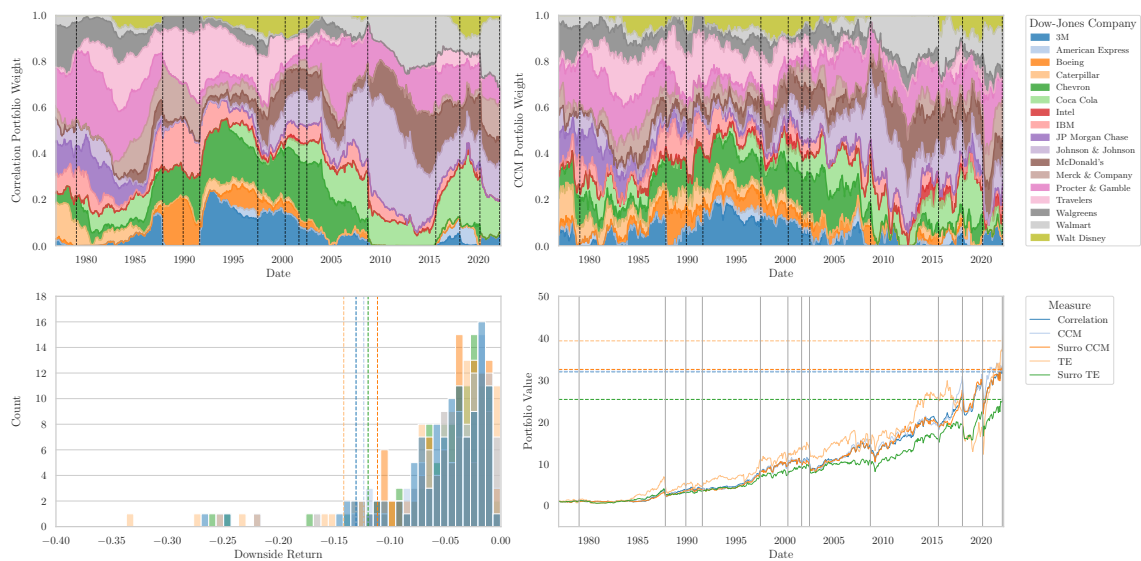


Figure 2.14: Minimum Risk Portfolio Optimization. The top row displays the optimized Minimum Risk Portfolio weights over time using both the correlation (on the left) and CCM (on the right) as co-dependence measures. Each colored area represents a stock from the Dow-Jones, which is mapped in the legend to the right. The dotted vertical lines depict significant economic or political events. In the bottom row, the left graph illustrates the distributions of the downside returns when using different co-dependence measures. The vertical lines depict the VaR at $\alpha = 1\%$ level. The graph to the right displays the portfolio's value over time. The vertical lines denote significant economic or political occurrences. The dotted horizontal lines denote the portfolio's most recent value. Each color corresponds to a particular co-dependence measure, which is mapped in the right-hand side legend.

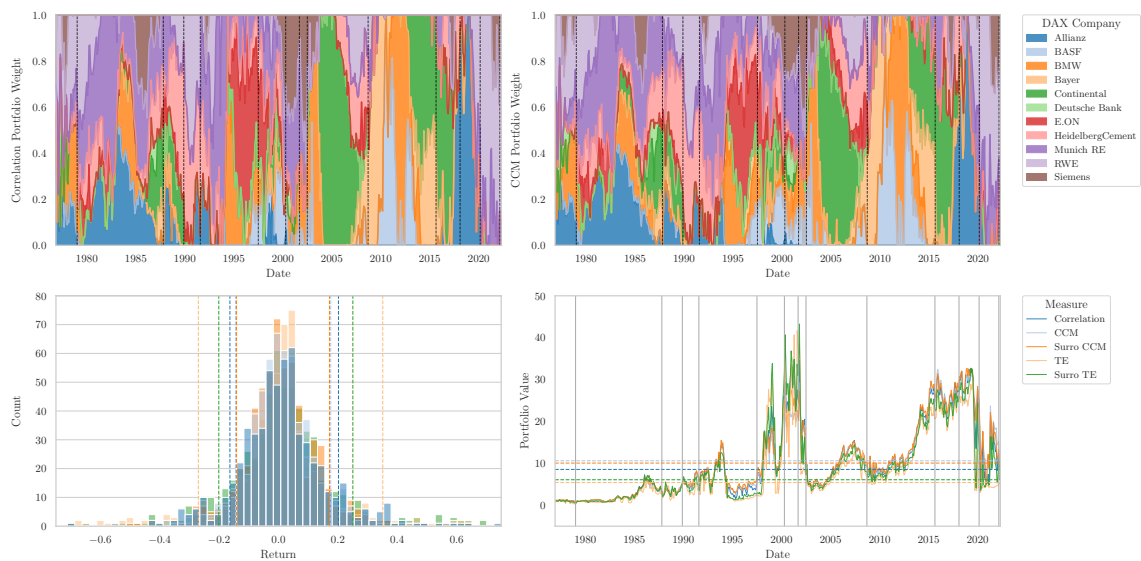


Figure 2.15: Maximum Sharpe Ratio Portfolio Optimization. The configuration is analogous to Figure 2.14. The bottom left plot displays the standard deviation of the returns instead of the VaR.

Chapter 3

Deriving Governing Equations using Causality and Synchronization

H. Ma, A. Haluszczynski,
D. Prosperino & C. R ath
“Identifying causality drivers and
deriving governing equations of
nonlinear complex systems”
*Chaos: An Interdisciplinary
Journal of Nonlinear Science*
vol. 32, no. 10, 2022

D. Prosperino, H. Ma & C. R ath
“A generalized machine learning method
for estimating parameters of complex
systems using synchronization”
*Submitted to
Chaos: An Interdisciplinary
Journal of Nonlinear Science*
TBD, 2023

In recent times, the abundance of data and advanced computer resources have stimulated interdisciplinary efforts to extract governing equations directly from data. This challenge is relevant to various fields, such as fluid dynamics, biological systems, and financial industries, with the aim of uncovering fundamental, mathematical relationships embedded in unprocessed information. Throughout history, mathematical models in the physical sciences have been expressed as differential equations. These equations are essential for conveying the fundamental relationships, drawing on both empirically-based findings and theoretical frameworks which often require significant simplification for practical application. Thanks to recent technological and computational advances in data collection and analysis, we can now extract equations directly from observed data. However, current techniques can operate in complex dimensions, rendering them as black-box processes with obscured inner workings. In contrast, we introduce a transparent and innovative framework that translates established causal mechanisms into mathematical equations, effectively depicting the underlying data. Through the integration of machine learning, we determine the optimal equation parameters that accurately reflect the data based on our initial assumptions. The combination of previous expertise and empirical evidence, reinforced by synchronization of chaos, enables us to effectively employ gradient descent algorithms. Our approach demonstrates robustness to noise and to a wide range of synthetic systems, and accurately identifies the appropriate parameters regardless of the complexity of the system.

3.1 Background and Motivation

The rapid expansion of computational resources has led to an unparalleled surge in data production and processing. Various methods exist for utilizing this data to generate forecasts and models. However, many approaches mask their process in a high-dimensional space, resulting in opaqueness. Instead, an alternative approach involves deducing the governing equations from a provided time series. This yields a model that is comprehensible and can be easily explained. Models like these could be beneficial in fields where both accuracy and explainability hold equal importance.

The path towards deriving governing equations from data has experienced profound evolution since the initial steps in the 1990s. Early techniques were primarily rooted in applying the *Flow Method*, as demonstrated by researchers such as Breeden & Hübler [89] and Eisenhammer, Hübler, Packard, et al. [90]. Since then, the last few decades have witnessed significant expansion in the study of these methodologies, particularly in the realm of nonlinear dynamical systems. Notably, Brunton, Proctor, & Kutz [91] marked a significant advancement by introducing *Sparse Identification of Nonlinear Dynamics* on the Lorenz attractor. This was further complemented by novel approaches like *Automated Inference of Dynamics* [21] and diverse machine learning techniques [92]. However, the number of potential terms grows exponentially for data that is high-dimensional and nonlinear.

Therefore, we introduce a methodology that utilizes causal inference to differentiate between linear and nonlinear effects and identify significant variables. Following our framework from Chapter 2, we can identify and distinguish between linear and nonlinear causal connections among system variables. Therefore, for a time series of N dimensions $\mathcal{S} = \{x_1, \dots, x_N\}$, we can now calculate the causality matrix $\psi(x, y)$:

$$\Psi(\mathcal{S}) \equiv \begin{pmatrix} \psi(x_1, x_1) & \dots & \psi(x_1, x_N) \\ \psi(x_2, x_1) & \dots & \psi(x_2, x_N) \\ \vdots & \ddots & \vdots \\ \psi(x_N, x_1) & \dots & \psi(x_N, x_N) \end{pmatrix},$$

which fully describes the causal links between the system variables. It resembles an *adjacency matrix* that represents finite graphs — hence the entries $\Psi_{i,j}$ quantify the causal flow from x_i to x_j . Similarly, we can compute the corresponding surrogate-based causality matrices.

We assume that the time series originates from a deterministic dynamic system, where a finite sample is sufficient to identify its causal structure. By separating linear and nonlinear causalities, the terms of the governing equations become separately identifiable. Thus, we argue that the causal structure can be fully described by a linear matrix differential equation and a nonlinear component:

$$\frac{d\mathbf{x}}{dt} = \left(\frac{d\mathbf{x}}{dt}\right)_{lin} + \left(\frac{d\mathbf{x}}{dt}\right)_{nl} = \Psi^{lin}\mathbf{x} + \Psi^{nl} \odot \mathbf{x}^n,$$

where \odot denotes the rationale for deriving the nonlinear terms, which will be explained later, and the superscript n denotes an n -dimensional *Cartesian* product. Next, the objective is to optimize a set of dynamic equations by determining the coefficients that accurately represent the data.

The recent expansion of deep learning technologies has led to the development of a variety of advanced first-order optimizers. Therefore, we focus exclusively on gradient descent algorithms. The first attempt using first-order optimizations was presented by Mariño & Míguez [93], but it is limited to the Lorenz system and lacks general applicability. While Abarbanel, Creveling, Farsian, et al. [94] further examines this topic within the field of data assimilation, there is still a dearth of comprehensive analysis regarding synchronization and stability considerations.

In this Chapter, we expand upon the algorithm created by Mariño & Míguez [93] and implement it to a variety of chaotic systems. Our approach proves to be highly robust against external noise and uncertainty regarding the underlying equations. We then evaluate the efficacy of our algorithm by reconstructing the Lorenz equations from multiple sets of data.

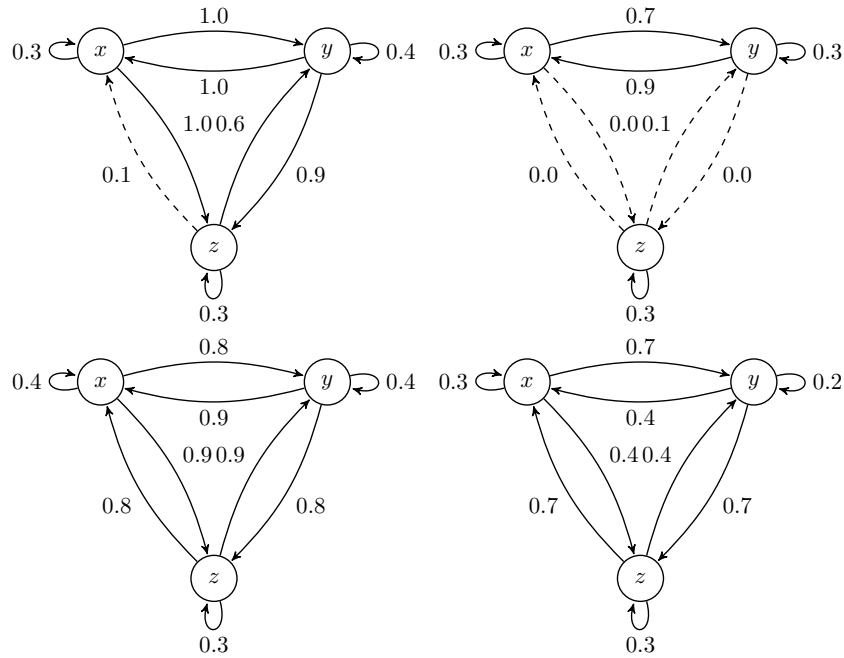


Figure 3.1: Causality Graphs. This Figure displays the causal graphs of the Lorenz (top row) and Halvorsen (bottom row) systems. The CCM between the state variables is depicted on the left, while the surrogate CCM is depicted on the right. The dashed lines indicate a lack of significant causality. These graphs allow for a clear and effortless extraction of the governing equations' parameters.

3.2 Derivation of Equation Terms from Causality

As previously motivated, in order to obtain the governing equations, it is necessary to create both linear and nonlinear causality matrices of the system beforehand. The surrogate matrix $\Psi_{i,j}^{surro}$ represents the linear causal flow between the system variables, while the cross-matrix $\Psi_{i,j}^{cross}$ represents the linear self-loops, which are causal flows of a variable to itself:

$$\Psi^{lin} = \delta_{i,j} \Psi_{i,j}^{cross} + (1 - \delta_{i,j}) \Psi_{i,j}^{surro}. \quad (3.1)$$

Next, the nonlinear causality matrix Ψ^{nl} is calculated by utilizing the original and surrogate matrices. This can be accomplished using the techniques delineated in Section 2.2.

To enhance the robustness of the derivation, we eliminate causalities below a predetermined threshold of θ^{lin} and attribute them to errors in causal inference. For nonlinear terms, we select a higher threshold of θ^{nl} to compensate for inaccuracies from two causal inferences. For simplicity, all nonlinear terms are assumed to be of order $n = 2$. The terms of the equations are derived using the following algorithm:

Algorithm 4 Derivation of Equation Terms

- 1: **Linear Terms.** The linear terms of the governing equations can be easily extracted via:

$$\left(\frac{dx_j}{dt}\right)_{lin} = \sum_i^N \Theta(\Psi_{i,j}^{lin} - \theta^{lin}) x_i, \quad (3.2)$$

where Θ denotes the *Heaviside* function.

- 2: **Quadratic Nonlinear Term.** If in one column x_j of Ψ^{nl} only one entry $x_i \neq x_j$ exceeds the threshold θ^{nl} , then the nonlinear term entering the equation is:

$$\left(\frac{dx_j}{dt}\right)_{nl} = \Theta(\Psi_{i,j}^{nl} - \theta^{nl}) x_i^2, \quad (3.3)$$

since we assume that the entire nonlinear causal flow of the system must be accumulated in the variable x_i .

- 3: **Mixed Nonlinear Terms.** If multiple entries $\{x_k, x_{k+1}, \dots, x_l\}$ in Ψ^{nl} exceed the threshold, then all pair combinations enter the equation:

$$\left(\frac{dx_j}{dt}\right)_{nl} = \sum_{i=k}^n \sum_{j \leq i}^l \Theta(\Psi_{i,j}^{nl} + \Psi_{j,i}^{nl} - \theta^{nl}) x_i x_j,$$

since we argue that the nonlinear causal flow must be split between all possible pairs.

- 4: **Merging Linear and Nonlinear Terms.** Then, we merge the linear and nonlinear parts of the derivatives to construct the complete governing equations:

$$\left(\frac{dx_j}{dt}\right) = \left(\frac{dx_j}{dt}\right)_{lin} + \left(\frac{dx_j}{dt}\right)_{nl}. \quad (3.4)$$

To complete the algorithm, coefficients for the individual term can be assigned and calibrated to the data — therefore, we discuss numerous state-of-the-art gradient-descent-based algorithms in the next Section.

3.3 Estimating Equation Parameters using Synchronization

After identifying the terms of our governing equations, we must calibrate the parameters to accurately represent the underlying data. For this part of our algorithm, we adhere to the fundamental concept proposed by Mariño & Míguez [93]. We begin with two systems: a primary system and a secondary system. The primary system holds the unknown parameters. The secondary system comprises our prior knowledge of the governing equations' structure and an initial estimate of the parameters to calculate. We then link the secondary system to the primary one and adjust its parameters until it synchronizes to the primary system. Once this happens, the secondary system provides a reliable estimation for the primary system. Accordingly, we determine the optimal parameter set that accurately describes the primary system.

Synchronization of Coupled Systems

In their study, Eroglu, Lamb, & Pereira [95] presented a method for coupling two chaotic systems of the same type, resulting in their trajectories converging to a common one. Here, we extend their approach to synchronize a secondary system *to* a primary one, whereby the primary system's trajectory remains unaltered, while the secondary system's trajectory is pushed to emulate the primary one. In this case, it is necessary to maintain the trajectory of the primary system without any alterations since using discrete data points as the primary system, which cannot be modified, is one of its applications.

We define two arbitrary N -dimensional systems $\underline{x}, \underline{y} \in \mathbb{R}^N$. We propose augmenting the system \underline{y} with the coupling $\alpha \mathbf{H}(\underline{x} - \underline{y})$. As a result, \underline{x} becomes the primary, driving system, while \underline{y} becomes the secondary, driven system. This results in the following dynamical description:

$$\dot{\underline{x}} = F(\underline{x}) \quad (3.5a)$$

$$\dot{\underline{y}} = F(\underline{y}) + \alpha \mathbf{H}(\underline{x} - \underline{y}) . \quad (3.5b)$$

Here, $F: \mathbb{R}^N \rightarrow \mathbb{R}^N$ describes the evolution of the dynamical system and the matrix $\mathbf{H}: \mathbb{R}^N \rightarrow \mathbb{R}^N$ describes the coupling of the secondary system to the primary one. We require $\mathbf{H}(\underline{0}) = \underline{0}$, meaning that for synchronized systems the coupling vanishes as soon as both systems are on the same trajectory [95]. The parameter $\alpha \in \mathbb{R}^+$ is named *coupling strength*.

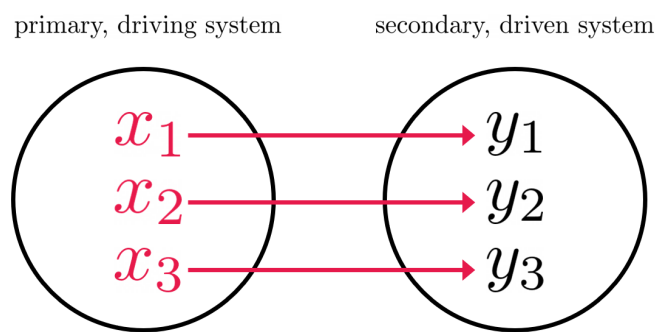


Figure 3.2: Unidirectional External Coupling. This Figure depicts the coupling scheme of a primary driving system and a secondary driven system for unidirectional external coupling, using the identity matrix as the coupling matrix \mathbf{H} . This representation describes an arbitrary three-dimensional chaotic system. Adapted from Prosperino [9].

Following the argument presented by Eroglu, Lamb, & Pereira [95], we can demonstrate that if the coupling strength is large enough, the systems described by Equations 3.5 synchronize. For the purposes of this discussion, we assume identity coupling, where $\mathbf{H} = \mathbf{1}$. This simplifies the coupling term in Equation 3.5b to:

$$\alpha \mathbf{H}(\underline{x} - \underline{y}) = \alpha (\underline{x} - \underline{y}) . \quad (3.6)$$

Additionally, we define a difference variable $\underline{z} = \underline{x} - \underline{y}$. Utilizing Equations 3.5, we can describe the evolution of the new variable \underline{z} by:

$$\begin{aligned}\dot{\underline{z}} &= \dot{\underline{x}} - \dot{\underline{y}} \\ &= F(\underline{x}) - F(\underline{y}) - \alpha \underline{z}.\end{aligned}\tag{3.7}$$

Now we must determine the coupling strength α at which the two systems will synchronize. To do this, we can use a *Taylor series expansion* [9] to estimate a critical coupling α_c . This yields the following condition:

$$\alpha \geq \alpha_c = \lambda,\tag{3.8}$$

where λ is the largest Lyapunov exponent of the system. Figure 3.3 demonstrates the numerical evidence for the Lorenz system. It clearly displays a significant reduction in synchronization loss between the two systems. This synchronization loss E_s is defined as the average deviation from synchronization between the two systems:

$$E_s = \frac{1}{N} \sum_{i=1}^N \|\underline{x}_i - \underline{y}_i\|.\tag{3.9}$$

For our optimization algorithm, this indicates that we may set the coupling strength α to a relatively high value.

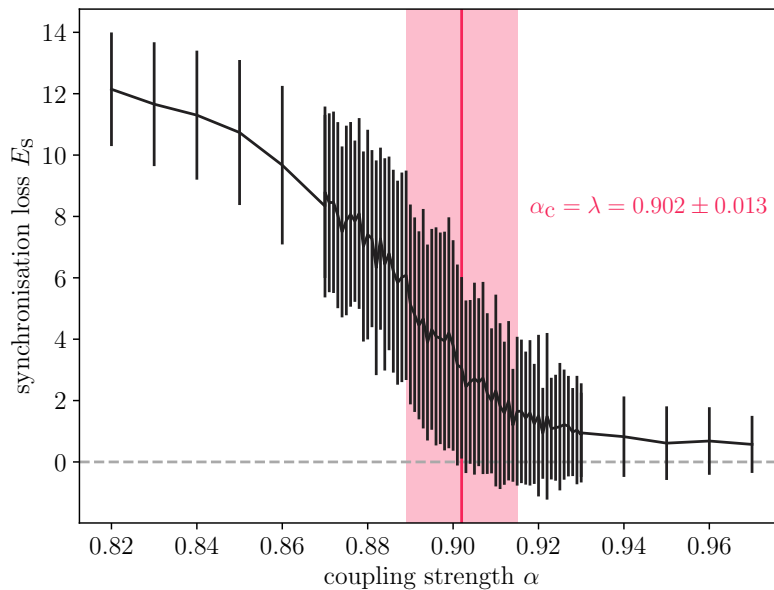


Figure 3.3: Critical Coupling Strength and Synchronization Loss. This Figure demonstrates a significant reduction in the synchronization loss E_s when the coupling strength exceeds the largest Lyapunov exponent for the Lorenz system. The average and standard deviation are calculated based on 100 iterations of each coupling strength.

Loss Function and Convex Surface

If two systems share the same initial condition and derivative, they will evolve identically. Therefore, based on the initial data point, the initial state of the secondary system is known by construction. The states of each system are not only dependent on the parameters, but also depend on the employed integration method. By optimizing on the derivative, we can eliminate an intermediary step while applying the chain rule. Thus, the absolute states of a system are not required, and the loss function does not depend on the absolute states $\{\underline{x}\}$ and $\{\underline{y}\}$, but instead on their derivatives.

Any mathematical norm can serve as the loss function \mathcal{L} , but in practice, the *Mean-Squared Error* (MSE) is a valuable measure of error. Instead of using an online optimization approach as presented by Mariño & Míguez [93], our algorithm utilizes a predetermined number of time steps for optimization. We refer to this fixed number of steps as the *evaluation length* l_e . Therefore, we compute the loss by adding up the MSE across each step in the evaluation length. At each time step i , a loss of ℓ_i occurs which are then summed to yield the overall loss, \mathcal{L} , across all steps:

$$\ell_i = \left(\dot{\underline{x}}_i - \dot{\underline{y}}_i \right)^2 \quad (3.10a)$$

$$\mathcal{L} = \frac{1}{l_e} \sum_{i=1}^{l_e} \ell_i. \quad (3.10b)$$

Experimentally, we find that coupling yields a convex loss surface, as depicted in Figure 3.4. There, we compute the loss \mathcal{L} for two coupled Lorenz systems and two uncoupled Lorenz systems. The primary system used standard parameters, while the secondary system was held constant at $\beta = \frac{8}{3}$ and various choices for parameters σ and ρ were swept through. It is not feasible to attempt optimization for determining the minimum on the surface of uncoupled systems, due to the absence of a clear direction towards the minimum of the loss. However, the loss surface of the coupled systems has a convex shape, which means that the application of gradient descent algorithms is justified and promising.

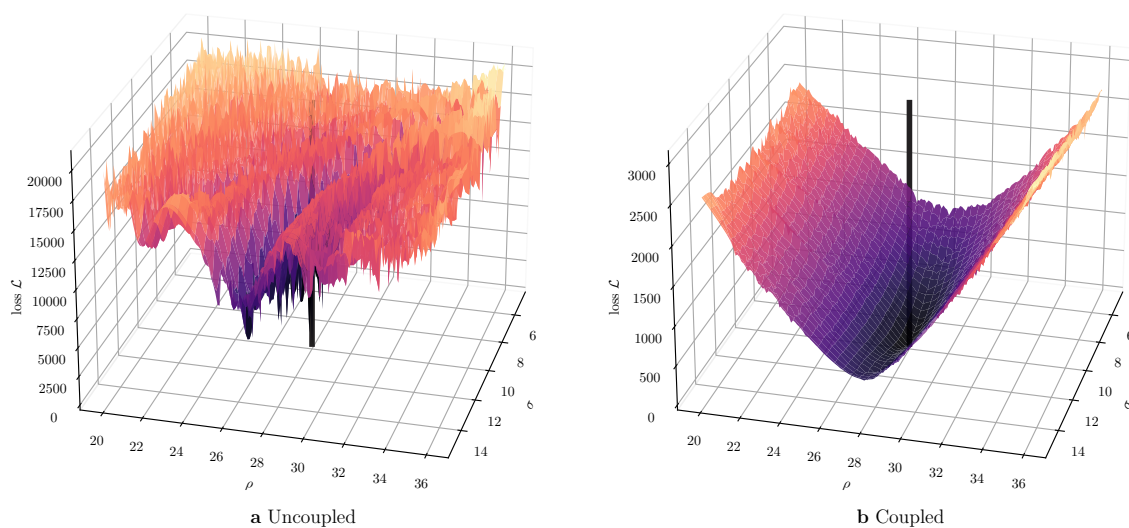


Figure 3.4: Loss Surfaces. This Figure illustrates the loss function \mathcal{L} for two uncoupled Lorenz systems (left) and two coupled Lorenz systems (right). The loss is calculated using an evaluation length of $l_e = 10^3$ and the systems are coupled with a strength of $\alpha = 10$. The primary system's parameters are represented by the black line. The right surface exhibits a convex shape, allowing for an optimizer based on gradient descent to operate effectively.

Update Step using Gradient Descent

Since the coupling transforms a non-convex loss surface into a convex one, we can calculate the loss function gradient with respect to the parameters $\underline{\theta}$ and take steps in the negative gradient direction. This iterative approach leads to a decrease in the loss function for the set of parameters with each step, resulting in a more precise description of the primary system.

The loss function ℓ is not directly dependent on the parameters $\underline{\theta}$ — rather, the evolution of the secondary system \underline{y} is the key factor. As a result, when calculating the derivative of the loss function with respect to a specific parameter θ in the parameter vector $\underline{\theta}$, the chain rule must be applied. Additionally, Equation 3.10b indicates a summation over every coordinate n within an N -dimensional system, given that vector-valued systems are being addressed. Generally, we can formulate the subsequent equation for the j -th entry g_j of the gradient \underline{g} by deriving Equation 3.10a:

$$g_j = \frac{\partial \mathcal{L}}{\partial \theta_j} = \frac{1}{l_e} \sum_{i=1}^{l_e} \frac{\partial \ell_i}{\partial \theta_j} \quad (3.11a)$$

$$\frac{\partial \ell}{\partial \theta_j} = -2 \sum_{n=1}^N (\dot{x}_n - \dot{y}_n) \frac{\partial \dot{y}_n}{\partial \theta_j}. \quad (3.11b)$$

For the purposes of readability, we exclude the i index from equation 3.11b. However, it is important to note that this equation represents the loss term of a single sample. Note that if a certain dimension n does not rely on the parameter θ_i , this term in the sum over the coordinates will be 0. Performing the calculation for the Lorenz system leads to the following expression for the gradients of its parameters $\underline{\theta} = (\sigma, \rho, \beta)^\top$:

$$\frac{\partial \ell}{\partial \sigma} = -2 (\dot{x}_1 - \dot{y}_1) (y_2 - y_1) \quad (3.12a)$$

$$\frac{\partial \ell}{\partial \rho} = -2 (\dot{x}_2 - \dot{y}_2) y_1 \quad (3.12b)$$

$$\frac{\partial \ell}{\partial \beta} = +2 (\dot{x}_3 - \dot{y}_3) y_3. \quad (3.12c)$$

After calculating the gradient, it is necessary to update the system parameters in the direction that minimizes the loss function. Utilizing the gradient-descent algorithm is the most straightforward approach, in which one subtracts the direction with a negative slope from the parameters. This method has been previously employed by Mariño & Míguez [93]:

$$\underline{\theta}_k = \underline{\theta}_{k-1} - \underline{\eta} \odot \underline{g}_k, \quad (3.13)$$

where \odot symbolizes the *Hadamard product* between two vectors. The index signifies the k -th step of optimization, and we perform the optimization process until all parameters have converged. The vector $\underline{\eta}$ represents the learning rate, which determines the size of the update step after each optimization step k . Because certain parameters may be more sensitive than others in complex systems, we select a distinct learning rate for each parameter. Heuristically, we find that adjusting the learning rate of each parameter, such that the initial update step results in a change of order 10^{-1} , yields good performance.

After iterating until the loss is small enough, we evaluate the error of the parameters using the mean absolute error E_θ of the fitted parameters $\hat{\underline{\theta}}$ with respect to the real parameters $\underline{\theta}$ described by:

$$E_\theta = \frac{1}{\dim \underline{\theta}} \sum_{i=1}^{\dim \underline{\theta}} |\theta_i - \hat{\theta}_i|,$$

which allows us to understand the numerical precision of our optimization.

Modern Optimizers using Adaptive Moments

A contemporary optimizer, known as *Adam* and named after adaptive moment estimation, was presented by Kingma & Ba [96]. It calculates an estimation of the gradient's first and second moments, which are achieved by exponentially weighting past gradients (first moment) and past squared gradients (second moment). The vector \underline{m} holds the first moment, whereas the vector \underline{v} holds the second moment. With the gradient at optimization step k denoted as \underline{g}_k , the moment estimations are updated via:

$$\begin{aligned}\underline{m}_k &= \beta_1 \underline{m}_{k-1} + (1 - \beta_1) \underline{g}_k \\ \underline{v}_k &= \beta_2 \underline{v}_{k-1} + (1 - \beta_2) \left(\underline{g}_k \odot \underline{g}_k \right).\end{aligned}$$

Both vectors, \underline{m} and \underline{v} , are initialized with $\underline{0}$. This initialization introduces a bias towards $\underline{0}$. Therefore, they suggest the following bias-corrected estimate:

$$\begin{aligned}\hat{\underline{m}}_k &= \frac{1}{1 - \beta_1^k} \underline{m}_k \\ \hat{\underline{v}}_k &= \frac{1}{1 - \beta_2^k} \underline{v}_k,\end{aligned}$$

where the superscript k represents taking the value to the k -th power, resulting in the update step:

$$\underline{\theta}_k = \underline{\theta}_{k-1} - \underline{\eta} \odot \hat{\underline{m}}_k \oslash \left(\sqrt{\hat{\underline{v}}_k} + \varepsilon \right), \quad (3.14)$$

where \oslash represents element-wise division, and the square root is taken over each element of $\hat{\underline{v}}_k$. The step size for each parameter $\underline{\eta}$ is determined using the method discussed in Section 3.3. We implement the suggested parameters, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-8}$, and observe successful optimization with the convergence of the parameters.

An expansion to the Adam optimizer that we will present in this Section is the *AMSGrad* optimizer by Reddi, Kale, & Kumar [97]. The authors discovered situations in which the Adam optimizer fails to converge to the optimal solution. This is due to the consideration of past squared gradients' moving average, which is one of Adam's primary characteristics. The problem arises when the optimizer takes excessively large steps under certain circumstances. If the second moment estimate \hat{v}_k^j for a parameter j at step k becomes too small, it can cause problems. To address this, they calculated the element-wise maximum value of the new and previous estimates. This prevents the step size from increasing for each parameter:

$$\hat{v}_k^j = \max \left(\hat{v}_{k-1}^j, v_k^j \right) \quad \forall j \in \{1, \dots, \dim \underline{\theta}\}.$$

Additionally, the optimizer is simplified by discarding the bias correction terms. As a result, the complete update step can be calculated using:

$$\begin{aligned}\underline{m}_k &= \beta_1 \underline{m}_{k-1} + (1 - \beta_1) \underline{g}_k \\ \underline{v}_k &= \beta_2 \underline{v}_{k-1} + (1 - \beta_2) \left(\underline{g}_k \odot \underline{g}_k \right) \\ \hat{\underline{v}}_k &= \max \left(\hat{\underline{v}}_{k-1}, \underline{v}_k \right) \\ \underline{\theta}_k &= \underline{\theta}_{k-1} - \underline{\eta} \odot \underline{m}_k \oslash \sqrt{\hat{\underline{v}}_k}.\end{aligned} \quad (3.15)$$

The maximum function in the upper equation is applied element-wise for each variance estimation of each parameter. Again, we use the recommended values for the remaining coefficients:

$\beta_1 = 0.99$, $\beta_2 = 0.999$. Unlike the original article, we use a different reference learning rate $\underline{\eta}$ for each parameter as derived in Section 3.3.

We compare the performance of each optimization algorithm using the Lorenz system as the primary system. The system has an initial state of $(5, 5, 5)^\top$ and a time step of $dt = 10^{-3}$. To ensure a fair comparison, the hyperparameters for each optimization algorithm remain constant with an evaluation length of $l_e = 10^3$ and a coupling strength of $\alpha = 10^3$. An initial guess of 1 is used for each parameter that requires estimation in the secondary system. The Figure below displays the results.

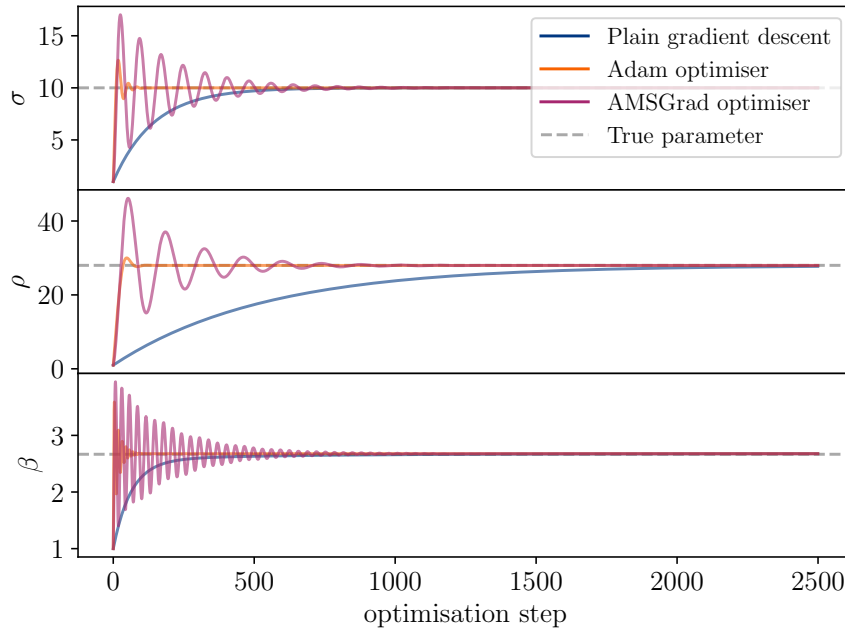


Figure 3.5: Coefficient Errors. This Figure shows the coefficient errors for different optimizers applied on the Lorenz system. We observe that the modern optimizers, Adam and AMSGrad, outperform the plain gradient-descent.

All optimization algorithms reach the correct solution, according to the results. However, sophisticated optimizers show faster convergence than plain gradient descent. Furthermore, we observe that the AMSGrad optimizer oscillates around the true parameters while the Adam optimizer shows no such oscillations. Our analysis does not reveal any adverse impacts on the optimization process resulting from the oscillations. Therefore, we will use the AMSGrad optimizer going forward, given its superior convergence rate in many cases compared to the Adam optimizer [96].

3.4 Algorithm Validation and Application Results

After presenting our algorithm for deriving governing equations from causality, we validate its effectiveness on synthetic systems and present some key results when the algorithm is applied to a set of synthetic chaotic systems. We find that our algorithm can accurately identify the correct equation terms based on causality. The loss surface is transformed into a convex one through coupling, allowing us to utilize modern machine learning techniques to calibrate equation parameters. Our results demonstrate the algorithm’s superior performance and robustness, even in the face of noise and potential errors in equation terms.

Translating Causality Structures to Equation Terms

Based on the framework developed in Chapter 2, this Section aims to examine the structure of both linear and nonlinear causality for the Lorenz and Halvorsen systems. The results provide a basis for deriving governing equations. Figure 3.6 illustrates that the x and y pair is mainly impacted by linear properties, resulting in the surrogate causality of GC and CCM being overshadowed, with both directions having an equal contribution. The surrogate TE indicates that there is mainly linear causality flowing from x to y , which accounts for approximately 41%. This finding agrees with the governing equations, where the equation for x has a linear contribution from y , whereas the equation for y has linear and nonlinear contributions from x . The rest of the causality is distributed evenly across the other flows. In contrast, the causal structure of the Halvorsen attractor illustrates that every flow contributes equally to causality across all inference techniques and types, as displayed in Figure 3.6. This pattern aligns with the circulant nature of the governing equations.

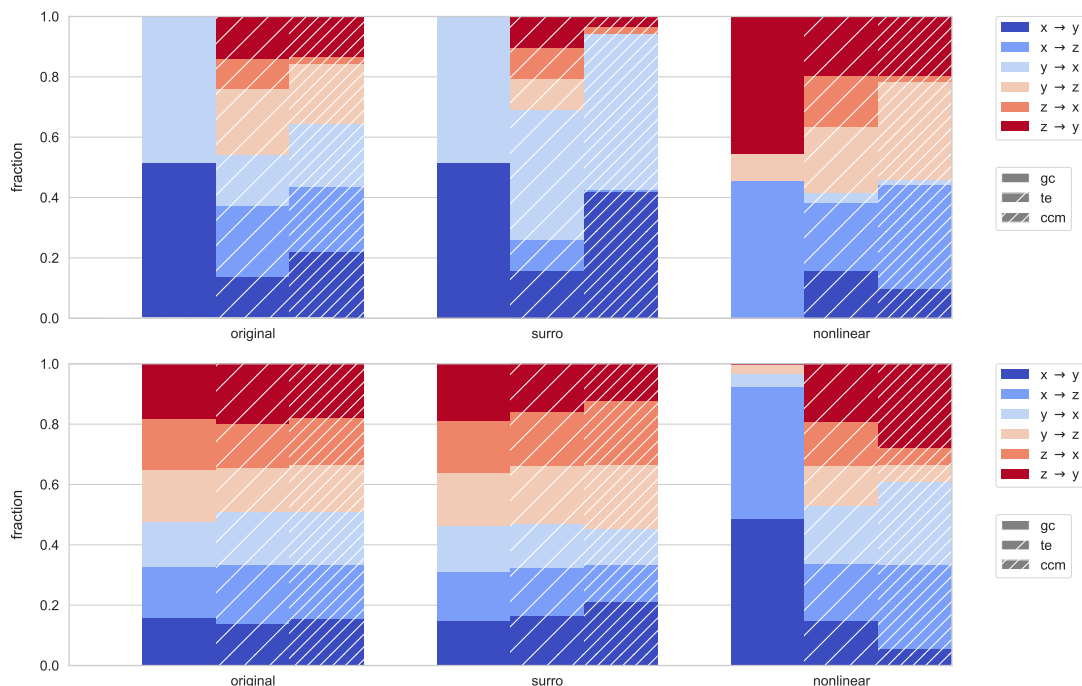


Figure 3.6: Causality Decomposition. This Figure displays the causal decomposition of the Lorenz (top row) and Halvorsen (bottom row) systems. We calculate the original, surrogate, and nonlinear causality for GC, TE, and CCM, respectively. To obtain each causal link’s contribution to the overall system causality, we divide the link’s causality by the system’s. The color map displays the contribution of individual causal flows to the total causality, while the different inference techniques are indicated by white stripes. The individual fractions are averaged across 50 simulations using the standard configuration. The surrogate-based causalities are averaged across 10 realizations.

To verify the rationale for deriving governing equations, we apply it to the Lorenz and Halvorsen systems, accompanied by their corresponding CCM-causal graphs shown in Figure 3.1. Here are the equations derived specifically for the Lorenz system:

$$\begin{aligned}\frac{dx}{dt} &= y - x \\ \frac{dy}{dt} &= x - xz - y \\ \frac{dz}{dt} &= xy - z,\end{aligned}\tag{3.16}$$

while the derived equations for the Halvorsen system are given by:

$$\begin{aligned}\frac{dx}{dt} &= x - y - z - y^2 \\ \frac{dy}{dt} &= y - z - x - z^2 \\ \frac{dz}{dt} &= z - x - y - x^2.\end{aligned}\tag{3.17}$$

After comparing our rationale to the true governing equations, we determine that it precisely replicates the terms in the Lorenz and Halvorsen systems given in Equations 50 and 52. Our results hold steady for thresholds where $\theta < 0.2$. To ensure reliability, we reevaluate our analysis with diverse initial conditions and notice that, for a simulation length of $T \geq 500$, the causality inference and equation derivation remain stable. Furthermore, our algorithm undergoes extensive testing on varied chaotic synthetic systems, demonstrating the ability to reconstruct equation terms via their linear and nonlinear causalities.

Estimating Equation Parameters of Synthetic Systems

The results of numerous optimizations on various three-dimensional chaotic systems are presented in the following Table 3.1. In our experiments, we find that the outcomes remain stable despite fluctuations in the hyperparameters α and l_e as long as they are within reasonable limits. Our results demonstrate that parameter reconstruction is accurate to the first decimal, allowing us to forecast the system several Lyapunov times in advance. Additionally, we discover that our optimization algorithm is resistant to initial conditions and, as a result, to the position on the attractor:

Table 3.1: Parameter Estimation for 3-Dimensional Chaotic Systems. This Table displays the Synchronization Error E_θ and Forecast Horizon τ for various three-dimensional systems. Mean and standard deviation values are provided for each system, which was tested 5 times. The systems utilize an integration step of $dt = 10^{-3}$, a coupling strength of $\alpha = 10^3$, and an evaluation length of $l_e = 10^4$.

System	Parameters	Synch. Error E_θ	Forecast Horizon τ
Thomas [98]	$f(x_i) = \sin x_i, b=0.21$	$1.3 \pm 3 \times 10^{-5}$	3.1 ± 0.2
Sprott [99]	$a=2.07, b=1.79$	$3.1 \pm 1.2 \times 10^{-4}$	9.8 ± 0.8
Lorenz [7]	$\sigma=10, \rho=28, \beta=\frac{8}{3}$	$4.2 \pm 0.5 \times 10^{-3}$	6.88 ± 0.05
Dadras-Momeni [100]	$a=3, b=2.7, c=1.7, d=2, e=9$	$7.9 \pm 1.7 \times 10^{-3}$	1.89 ± 0.05
Rössler [101]	$a=0.1, b=0.1, c=14$	0.011 ± 0.015	4.9 ± 0.2
Halvorsen [102]	$a=1.89$	0.0147 ± 0.0007	1.094 ± 0.002
Lorenz86 [103]	$a=0.25, b=4, f=1.1, g=8$	0.023 ± 0.006	0.75 ± 0.05
Three-Scroll [104]	$a=40, c=\frac{5}{6}, d=0.5, e=0.65, f=20$	0.037 ± 0.002	0.3 ± 0.2

Thus far, the optimization algorithm has been exclusively applied to three-dimensional systems encompassing a small pool of parameters. In the ensuing discussion, an assessment of the applicability of this methodology to high dimensional problems will be made. To this end, we will employ the Lorenz96 system [105], which can be scaled to any desirable dimensionality. All variables are assigned a coefficient with equal value of one, thereby constructing an N -dimensional system:

$$\dot{x}_d = (x_{d+1} - x_{d-2}) x_{d-1} - x_d + F \quad d \in \{1, \dots, N\} .$$

We assume cyclic boundary conditions as follows: $x_{-1} = x_{N-1}$, $x_0 = x_N$, and $x_{N+1} = x_1$. This system almost perfectly aligns with our agnostic guess of 1 for each parameter. In order to extend this model to suit our goals, we introduce a parameter for each variable in each dimension:

$$\begin{aligned} \dot{x}_d = & \quad (p_{3(d-1)+1} x_{d+1} - p_{3(d-1)+2} x_{d-2}) x_{d-1} \\ & - p_{3(d-1)+3} x_d + p_{3N+1} \quad d \in \{1, \dots, N\} . \end{aligned}$$

The parameters are stored in a $(3N + 1)$ -dimensional vector, denoted as \underline{p} . In our parametrized Lorenz96 system, we maintain the same value of forcing constant F across all dimensions. We store this constant as the final entry in the vector \underline{p} , which is renamed to p_{3N+1} in the equation above. The variables are assigned parameters p_1 to p_{3N} from a uniform distribution within the interval of $[0, 1]$ while the forcing constant, p_{3N+1} , is given the value of 8 consistent with the original model [105].

Optimization tests are conducted on several Lorenz96 systems of gradually increasing dimensionality. We run five experiments for every dimension and use a different parameter vector \underline{p} for each. Table 3.2 reveals the outcomes of the optimizations, showcasing the effectiveness of the process across a wide range of dimensions. Our optimization algorithm can accurately fit parameters with a high count. Furthermore, increasing dimensionality does not hinder predictability. For each dimension, our algorithm can predict multiple Lyapunov times ahead, as evidenced by Table 3.1. It is important to note that our algorithm does not have a preference for any specific dimension region. This demonstrates the algorithm's ability to function effectively with high-dimensional systems containing multiple free parameters.

Table 3.2: Parameter Estimation for N -Dimensional Lorenz96 Systems. This Table presents the results of the Lorenz96 system as dimensions increase. Some trials yielded outlier predictable times in the positive direction — therefore, we only report the minimum predictable time for each dimensionality to prevent skewed results. We utilize a precise time step of $dt = 10^{-4}$, an evaluation length of $l_e = 10^5$, and a coupling strength of $\alpha = 10^5$. Due to the steep computational requirements, we only conduct a single experiment for $N = 64$.

Dimensionality N	# Parameters	Synch. Error E_θ	Min. Forecast Horizon τ
5	16	0.07 ± 0.07	2.3
6	19	$2.4 \pm 1.2 \times 10^{-3}$	6.3
7	22	$1.3 \pm 1.2 \times 10^{-3}$	8.1
8	25	$7 \pm 8 \times 10^{-3}$	8.0
9	28	$2.9 \pm 1.9 \times 10^{-3}$	3.6
10	31	0.05 ± 0.07	3.2
12	37	$4 \pm 3 \times 10^{-3}$	2.0
14	43	0.04 ± 0.05	2.0
16	49	$7 \pm 6 \times 10^{-3}$	1.5
32	97	0.05 ± 0.04	2.1
64	193	1.04×10^{-3}	2.0

Robustness against Noise

Up to this point, we conducted all computations utilizing synthetic systems that are inherently precise and free from errors. This Section aims to examine our optimization’s resilience to noise in the primary system’s data. Accordingly, we simulate the classical Lorenz system with its standard parameters and introduce Gaussian noise to each data point and coordinate. The zero-mean Gaussian noise draws from a normal distribution with a standard deviation of σ . In our experiments, the standard deviations range between 0.01 and 2. Prior to incorporating Gaussian noise into the accurate data, we determine the *Signal-to-Noise Ratio* (SNR) by dividing the power of the signal by the power of the noise. The power is assessed using squared amplitude A , allowing us to express SNR as follows:

$$\text{SNR} = \left(\frac{A_{\text{signal}}}{A_{\text{noise}}} \right)^2$$

$$\text{SNR} = 10 \log_{10} \left(\frac{A_{\text{signal}}}{A_{\text{noise}}} \right)^2 \text{ dB}.$$

The amplitude A of a time series $\underline{x}(t)$ can be calculated with:

$$A = \sqrt{\frac{1}{t} \int_0^t \|\underline{x}(\tilde{t})\|^2 d\tilde{t}}.$$

The following Figure 3.7 illustrates the results obtained by varying the coupling strength. Notably, we observe an intriguing trend wherein the parameter errors follow a power law within a certain range. The observed cut-off for this power law behavior at 10^{-2} arises from our chosen convergence criterion for these experiments. A decrease in coupling strength leads to a more negative exponent in the observed power law fit. This indicates that a performance boost occurs at lower coupling strengths, as errors do not increase as rapidly with rising noise levels.

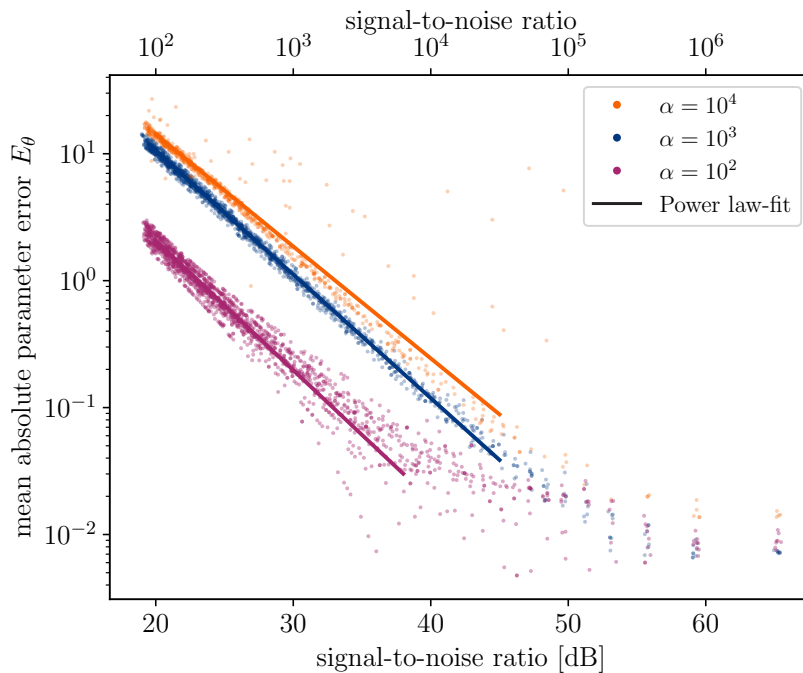


Figure 3.7: Robustness Against Noise. This Figure presents the performance for varying coupling strengths. Lower coupling strengths are found to perform better on noisy data.

In real-world data analysis, this suggests that setting the coupling strength too high produces inferior outcomes than weaker coupling strengths. One possible explanation is that a strong coupling compels the secondary system to perceive the primary system’s noise as real dynamics, whereas a weaker coupling allows the dynamics term $F(\underline{y})$ in Equation 3.5b to hold more influence. Due to the similarity between our algorithm and the one proposed by Mariño & Míguez [93], we use their algorithm as a benchmark. As stated in the previous subsection, we implement a moderate coupling strength of $\alpha = 10^2$. Fig. 3.8 demonstrates that, although the benchmark algorithm occasionally performs better, our algorithm proves to be more stable and consistently produces superior results.

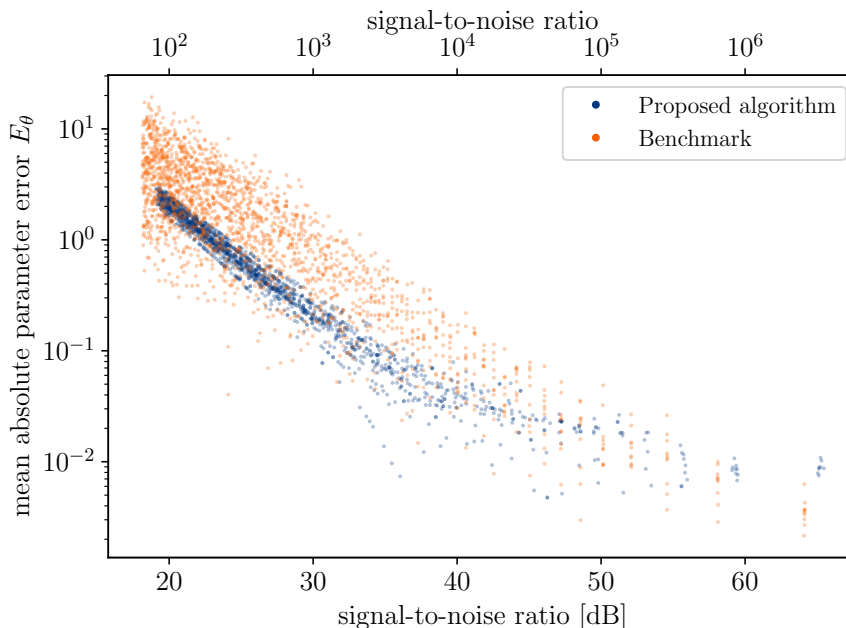


Figure 3.8: Comparison against Benchmark. This Figure depicts a comparison of the performance of our algorithm (blue) with the benchmark algorithm (orange) as presented by Mariño & Míguez [93].

Incorrect Prior and Equation Terms

Another potential source of uncertainty could arise due to an inadequate prior understanding of the governing equations. To investigate this, we incorporate an extra linear and nonlinear term into the original Lorenz system, resulting in the following revised system:

$$\dot{x}_1 = -\sigma x_1 + \sigma x_2 + \kappa x_2 x_3 \quad (3.18a)$$

$$\dot{x}_2 = \rho x_1 - x_2 - x_1 x_3 + \xi x_2 \quad (3.18b)$$

$$\dot{x}_3 = -\beta x_3 + x_1 x_2. \quad (3.18c)$$

This adds the following two gradients to the set of Equations 3.12:

$$\begin{aligned} \frac{\partial \ell}{\partial \kappa} &= -2(\dot{x}_1 - \dot{y}_1) y_2 y_3 \\ \frac{\partial \ell}{\partial \xi} &= +2(\dot{x}_2 - \dot{y}_2) y_3. \end{aligned}$$

As in previous experiments, all parameters are initialized with a value of one, as we have no prior knowledge of the parameters. By using a coupling strength of $\alpha = 10^2$ and an evaluation length of $l_e = 10^3$, the parameters are found as presented in Table 3.3. The algorithm successfully

sets the parameters almost to zero, thus minimizing their impact on the governing equations. The remaining true parameters of the system are also reconstructed correctly. Adding incorrect terms to the prior knowledge of the primary system causes a divergence of the coefficients for the benchmark algorithm.

Table 3.3: Parameter Estimation for Incorrect Equation Terms. This Table displays the outcomes of executing our suggested optimization on information derived from the Lorenz system onto the system defined by Equations 3.18. The experiment is conducted 50 times with varying initial conditions, and the error reflects one standard deviation of the measurements.

Parameter	True Value	Reconstructed Value
σ	10	10.027±0.015
ρ	28	27.795±0.014
β	2.6̄	2.676±0.003
κ	0	-0.0014±0.0004
ξ	0	0.000±0.002

We want to note two limitations of our algorithm: the coefficients κ and ξ are correctly set to a value close to zero. However, it is currently not possible to determine whether terms with small coefficients are irrelevant for the dynamics of the system. Therefore, we advise against discarding terms with coefficients close to zero. Additionally, our algorithm is susceptible to under-inclusion of terms. This implies that if a term is not present in the prior knowledge of the system, our algorithm will fail and diverge instead. However, this can be easily resolved since the reconstruction process tends to over-include terms, which makes it a non-issue.

Chapter 4

Predicting Chaos with Binary and Minimal Reservoir Computing

H. Ma, D. Prosperino,
A. Haluszczynski & C. R ath
“Efficient forecasting of chaotic systems
with block-diagonal and binary
reservoir computing”
*Chaos: An Interdisciplinary
Journal of Nonlinear Science*
vol. 33, no. 6, 2023

H. Ma,
D. Prosperino & C. R ath
“A novel approach
to minimal
reservoir computing”
*Scientific Reports
Nature Portfolio*
vol. 13, no. 1, p. 12970, 2023

Next to the construction of governing equations, the use of machine learning to predict complex dynamic systems has gained popularity across scientific fields. Reservoir computing, a successful technique for reproducing such systems, utilizes a sparse, random network to create the system’s memory. In this dissertation, we advocate the use of a block-diagonal reservoir, which is essentially composed of several smaller reservoirs, each of which has unique dynamics. This approach challenges the traditional notion of the reservoir as a single network. Additionally, we remove any randomness from the reservoir through matrices filled with ones for each block. We assess the prediction performance of block-diagonal reservoirs and their sensitivity to hyperparameters. Our results indicate that they perform similarly to sparse random networks, prompting a discussion regarding their implications on scalability, comprehensibility, and the physical implementation of reservoir computers. Furthermore, current advancements in reservoir computation concentrate on linear and nonlinear regressions of input data combinations, including time lags and polynomial derivatives, completely eliminating the need for randomness. However, handling high-dimensional and nonlinear data results in a considerable rise in the number of possible combinations. Thus, our research demonstrates that modifying the traditional architecture of reservoir computers to reduce computational requirements can significantly and consistently improve predictive accuracy over long and short time scales. This improvement is evident even when using relatively small training datasets and when compared to similar models. This efficient design with minimal data requirements offers a practical solution for implementation in real-world situations where data gathering is challenging or costly.

4.1 Background and Motivation

While machine learning techniques have shown promise in accurately predicting the behavior of dynamic systems [106], their usefulness in certain scientific applications is limited by challenges related to the vast amounts of required data, numerous hyperparameters, and limited interpretability [107]. In many areas, however, a fundamental understanding of the models is necessary to avoid misinterpretations in the absence of deeper methodological knowledge [108].

In the field of complex systems research, *Reservoir Computing/Computers* (RC) [36] has/have emerged to quantify and predict the spatiotemporal dynamics of chaotic nonlinear systems. RCs are a type of *Recurrent Neural Network* (RNN) and are commonly known as *Echo-State Networks* (ESNs) [109]. The heart of the model consists of a fixed reservoir, a complex network with connections according to a predefined network topology, which can have a significant impact on the prediction performance [110]. In current state-of-the-art models, the topology of the reservoir is often chosen randomly [111] in the hope that the resulting dynamics will be sufficiently complex to allow good performance on a given task. Nonetheless, this method can be hit-or-miss [35], and it is unfeasible to foretell beforehand how the topology of the reservoir will affect the system’s performance. While reservoirs modeled as random *Erdős–Rényi* networks were introduced by Maass, Natschläger, & Markram [33] and Jaeger [32], research by Watts & Strogatz [112], Albert & Barabási [113], and others has revealed that random networks are not common in physics, biology, or sociology [114]. Instead, real-world applications often exhibit more complex networks such as scale-free, small-world, or intermediate forms [115].

In recent years, several new approaches to improve the explainability of RC have emerged. For instance, Haluszczynski & Răth [35] compared various network construction algorithms, Griffith, Pomerance, & Gauthier [116] introduced low-connectivity networks, and Carroll & Pecora [117] analyzed how network symmetries affect prediction performance. However, questions about the functionality of RCs must still be addressed to develop new algorithms, optimize the system for specific applications, and build efficient hardware realizations of RCs.

Our work challenges the interpretation of the reservoir as a single network by intentionally using block-diagonal matrices as reservoirs. Therefore, the reservoir is divided into multiple smaller ones, each of which has unique dynamics. Furthermore, we utilize matrices of ones as the blocks, which eliminates any randomness in the network altogether. Our inspiration comes from an experiment where we constructed the reservoir as a two-dimensional *Ising* model [118] and wanted to observe how a phase transition would affect the prediction performance. After observing no significant decline in performance, we concluded that the network can be created as a block-diagonal matrix consisting of ones.

Further research has revealed new algorithms that do not rely on randomization. These algorithms employ regressions on extensive libraries of linear and nonlinear combinations constructed from data observations and their respective time lags [119]. Innovations include *Next Generation Reservoir Computers* (NG-RCs) [120] and *Sparse Identification of Nonlinear Dynamics* (SINDy) [91]. These algorithms are based on *Nonlinear Vector Autoregression* (NVAR) [121] and the mathematical fact that a powerful universal approximator can be constructed by using an RC with a linear activation function [34], [122]. The model presented in this dissertation is built on the same mathematical principles as the traditional reservoir architecture. The input weights are restructured to separately feed all coordinate combinations into the reservoir. The randomness of the reservoir is removed by replacing it with a block-diagonal matrix of blocks of one. Instead of introducing nonlinearity to the activation function, we add higher order reservoir states in the readout. Applying this novel architecture to synthetic chaotic systems, we show that these changes result in short- and long-term predictions that outperform traditional RC, NG-RC, and SINDy.

4.2 Prediction of Dynamical Systems

The prediction of dynamical systems is crucial for physics and computational science. While traditional modeling approaches often encounter difficulties in addressing the growing complexity of modern systems, machine learning often lacks interpretability. In this Section, we explore the prediction of dynamical systems through RC and related approaches. We detail the advancements of NG-RC, emphasizing its capabilities and differences. Lastly, we discuss SINDy, a method that efficiently identifies governing equations from data, thus bridging the gap between data-driven modeling and classical physics.

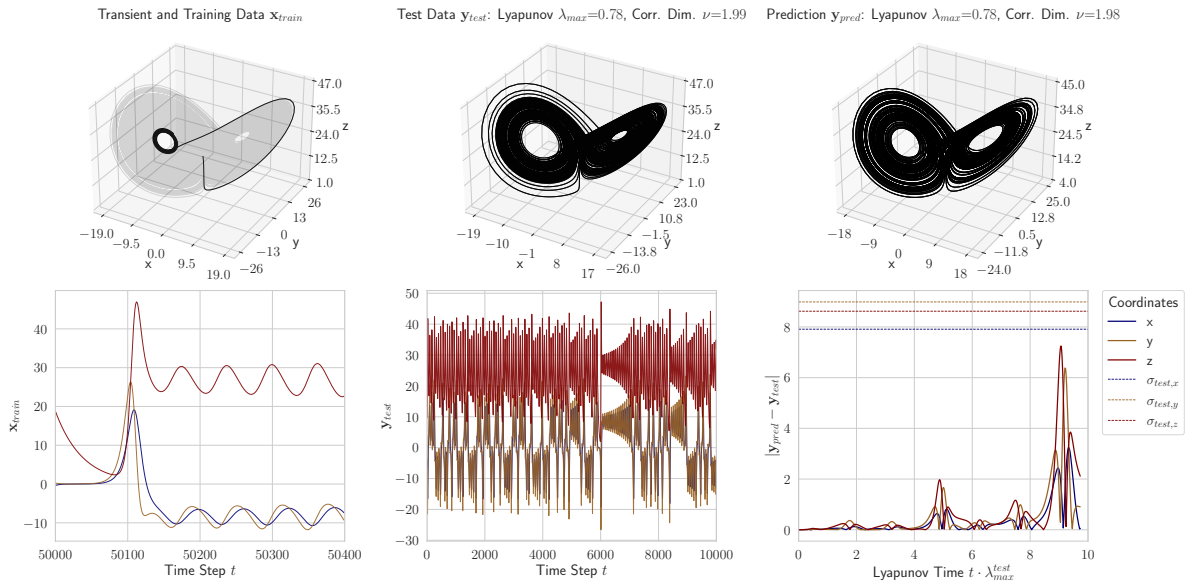


Figure 4.1: Prediction of Chaotic Systems. The leftmost column displays the Lorenz attractor’s attractor (top) and trajectories (bottom) of training data points (including the discarded transient). In the middle column, the attractor and trajectories of the test data are presented. The rightmost column demonstrates the attractor (top) and absolute prediction error (bottom) of the prediction. Dashed lines indicate the standard deviation of the test data.

Reservoir Computing

Reservoir Computing (RC) is a computational framework that has gained significant attention in the field of machine learning and nonlinear time series analysis. It originated from two independently proposed *Recurrent Neural Network* (RNN) architectures, which were introduced as *Echo State Networks* (ESNs) within the field of machine learning by Jaeger [32], and *Liquid State Machines* (LSMs) from the field of computational neuroscience.

The idea behind RC is to leverage the dynamics of a large, fixed, and randomly connected recurrent neural network as a reservoir to process temporal data. This approach simplifies the training process by only requiring the linear optimization of a readout layer, leading to computational efficiency and improved performance on various tasks. Hence, it offers a powerful approach for processing complex, temporal data, particularly in the context of prediction and pattern recognition tasks [111]. One type of time series that has been shown to be well predictable with RC are the trajectories of chaotic systems. In these cases, RC excels not only at creating accurate short-term predictions, but also at replicating the ergodic properties of the system’s attractor [123].

Fundamental Concept

In the following discussion, we limit the use of RC to tasks that create a connection between time-varying input data $\mathbf{u}(t)$ and the desired output data $\mathbf{y}(t)$. Here, the time series data that has been injected interacts with a complex and nonlinear dynamical system of high dimensionality, referred to as reservoir \mathbf{A} . As a result, the state of the reservoir $\mathbf{r}(t)$ evolves as a reflection of both current and previous inputs, with the influence of older inputs diminishing over time. The high-dimensional reservoir state $\mathbf{r}(t)$ is dynamically linked to the time-dependent input $\mathbf{u}(t)$ via the following equation:

$$\mathbf{r}(t+1) = f(\mathbf{W}_{in}\mathbf{u}(t) + \mathbf{A}\mathbf{r}(t)), \quad (4.1)$$

where $\mathbf{r}(t)$ is the reservoir's r_{dim} -dimensional state vector at time t , $\mathbf{u}(t)$ is the input vector of dimension u_{dim} at time t . \mathbf{W}_{in} is an $r_{dim} \times u_{dim}$ matrix, representing the input weights, \mathbf{A} is an $r_{dim} \times r_{dim}$ adjacency matrix, representing the reservoir weights, and f is a nonlinear activation function, typically the *hyperbolic tangent*. A visual depiction of the fundamental RC structure is available in Figure 1.3.

Input Weights

Before feeding the data to the reservoir, it is embedded into a high-dimensional, random space through the use of an input matrix. The elements of the input matrix, noted as $W_{in,ij}$, indicate the connection strength between the input variable u_j and the reservoir node r_i . Commonly, \mathbf{W}_{in} is populated with random numbers originating from a uniform distribution spanning $[-\sigma, \sigma]$. In this context, σ represents the *input strength*. A higher input strength signifies a stronger linkage from the input to the reservoir. As proposed by Lu, Hunt, & Ott [124], we employ a configuration, wherein each reservoir node \mathbf{r}_i is exclusively connected to a random input variable, ensuring that each row of \mathbf{W}_{in} contains just one non-zero element.

Reservoir

Our approach leans toward the conventional method of constructing a large, random, and sparse network, an approach known to invoke complex reservoir dynamics, as underscored in the pioneering research of Jaeger [32]. Following Lu, Hunt, & Ott [124], the network is constructed using a weighted *Erdős-Rényi* random network [125]. There, we begin by initializing an all-zero adjacency matrix. Then, with a probability of p , we assign a value of 1 for every off-diagonal element. On average, each node is connected to $d = p \times (r_{dim} - 1)$ distinct nodes. This is a crucial parameter referred to as the *average node degree*. Subsequently, we assign a random value within the uniform distribution between -1 and 1 to each previously assigned 1, weighting the matrix.

Ensuring optimal reservoir operation requires adherence to the *Echo State Property* (ESP), which dictates that reservoir states should gradually become independent of initial conditions over time. To maintain the ESP, it is necessary to scale the *spectral radius* ρ of the reservoir matrix \mathbf{A} . The spectral radius of a square matrix corresponds to its largest absolute eigenvalue, representing the maximum scaling factor that occurs after matrix operations:

$$\rho(\mathbf{A}) = \max \{ |\lambda_i(\mathbf{A})| \}, \quad (4.2)$$

where $\lambda_i(\mathbf{A})$ refers to the eigenvalues of \mathbf{A} . Therefore, the chosen spectral radius target ρ^* within the reservoir network is critical for determining the impact of previous reservoir states on future ones. As such, \mathbf{A} undergoes scaling to adjust to the desired spectral radius, where $0 < \rho^* < 1$, in order to finalize the reservoir matrix.

$$\mathbf{A} \mapsto \frac{\rho^*}{\rho(\mathbf{A})} \mathbf{A}. \quad (4.3)$$

Readout

The iterative equation, represented by Equation 4.1, yields the reservoir states, denoted as \mathbf{r} . To enhance prediction accuracy by breaking symmetries, Herteux & R  th [126] introduced an innovative technique that involves applying a *readout* function to these reservoir states before the training. Following this methodology, we further enrich our states by including squared reservoir states, resulting in an augmented reservoir state matrix termed $\tilde{\mathbf{r}}$:

$$\tilde{\mathbf{r}} = [r_1, \dots, r_{r_{dim}}, r_1^2, \dots, r_{r_{dim}}^2], \quad (4.4)$$

where each element in the reservoir states undergoes a squaring operation. Hence, the *generalized reservoir states* at time t are denoted as $\tilde{\mathbf{r}}(t)$.

Training and Prediction

The training process centers around fine-tuning \mathbf{W}_{out} to minimize the discrepancy between the prediction and the actual data. A commonly used methodology for this purpose is *ridge regression* in linear regression analysis, which helps to address multicollinearity and avoids overfitting in situations with numerous predictor variables [127]. The main difference between ridge regression and *Ordinary Least Squares* (OLS) regression, is the inclusion of a regularization term in the loss function. In its matrix expression, ridge regression is given by:

$$\mathbf{W}_{out} = \left(\tilde{\mathbf{R}}^T \tilde{\mathbf{R}} + \beta I \right)^{-1} \tilde{\mathbf{R}}^T \mathbf{Y}, \quad (4.5)$$

where \mathbf{W}_{out} is the output weights matrix with dimensions $u_{dim} \times 2r_{dim}$. It is the only component of the RC that undergoes training. \mathbf{Y} is the training data matrix, with dimensions $u_{dim} \times T$, that carries the desired output values. $\tilde{\mathbf{R}}$ is the matrix of size $2r_{dim} \times T$ containing the stacked generalized reservoir states. The *regularization parameter* β controls the strength of the penalty for large parameter values, which ensures stability. For $\beta = 0$ the ridge regression reduces to an OLS regression. I is the identity matrix.

Thus, the output of the RC is a direct linear projection of the reservoir states and the output weights.

$$\mathbf{y}(t) = \mathbf{W}_{out} \tilde{\mathbf{r}}(t). \quad (4.6)$$

The benefit of the recursive Equation 4.1 lies in its ability to make predictions of arbitrary lengths. Thus, the iterative prediction equation is:

$$\mathbf{r}(t+1) = f(\mathbf{W}_{in} \mathbf{W}_{out} \tilde{\mathbf{r}}(t) + \mathbf{A} \mathbf{r}(t)). \quad (4.7)$$

Next-Generation Reservoir Computing

In 2021, Gauthier, Bollt, Griffith, et al. [120] unveiled the advanced architecture of the *Next-Generation Reservoir Computing* (NG-RC). This novel approach stood out due to its deterministic nature, fewer variables, and notably faster performance relative to conventional methodologies. While traditional RC relies on reservoir matrices with random initialization for its network structure and utilizes a linear readout, NG-RC employs a series of distinct polynomials for nonlinear dimensionality enhancement at its foundation. Once the input data's state space is defined, it undergoes systematic training via ridge regression to align with the intended output.

The d -dimensional data points \mathbf{x} of the input data $\mathbf{X} = [\mathbf{x}_0, \dots, \mathbf{x}_n]$ undergo transformation via a polynomial multiplication dictionary \mathbf{P} into a space of increased dimensions. Within this dictionary, unique polynomials of specific orders O present in $\mathbf{P}^{[O]}$ are identified by an index. To provide clarity, we consider an input data point $\mathbf{x}_t = (x_{t,1}, x_{t,2})^T$. This data point is then altered using the unique polynomials of the first and second order:

$$\mathbf{P}^{[1,2]}(\mathbf{x}_t) = \begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t,1}^2 \\ x_{t,2}^2 \\ x_{t,1} \cdot x_{t,2} \end{pmatrix} \quad (4.8)$$

Moreover, Gauthier, Bollt, Griffith, et al. [120] introduced the concept of a *time-shift expansion* \mathbf{L}_k^s to the input data, setting the NG-RC apart from traditional NVAR techniques [121]. Here, the embedding dimension k represents the number of past data points merged with the current data point, while the lag s depicts the temporal distance between these points. When this expansion is applied to the input, it shapes the NG-RC's linear reservoir layer, which yields:

$$\mathbf{r}(t+1) = \mathbf{P}^{[1,2]}(\mathbf{L}_k^s(\mathbf{x}_t)) = \mathbf{P}^{[1,2]} \left(\begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t-1,1} \\ x_{t-1,2} \end{pmatrix} \right) = \begin{pmatrix} x_{t,1} \\ x_{t,2} \\ x_{t-1,1} \\ \vdots \\ x_{t,1} \cdot x_{t-1,2} \\ x_{t,2} \cdot x_{t-1,2} \end{pmatrix}, \quad (4.9)$$

where $\mathbf{r}(t+1)$ represents the state vector. Analogous to the conventional RC, this vector is mapped using an output matrix \mathbf{W}_{out} to achieve the desired output target \mathbf{y}_t . The training within NG-RC is also conducted using ridge regression.

During the training phase of NG-RC, the input training data, denoted as \mathbf{X}_T , of a specific duration T , undergoes a transformation to yield the state matrix $\mathbf{R} = \mathbf{P}^{[O]}(\mathbf{L}_k^s(\mathbf{X}_T))$. It is essential to consider that based on the values of k and s , a preparatory phase or *warm-up* duration of $\delta t = k \cdot s$ becomes necessary since for periods $t < \delta t$, the entries in the state matrix remain undefined.

One of the most notable attributes of the NG-RC is its ability to provide results that are interpretable. According to the Equation:

$$\begin{pmatrix} \dot{x}_i \\ \dot{x}_{i+1} \\ \vdots \\ \dot{x}_n \end{pmatrix} = \frac{1}{dt} \mathbf{W}_{out} \mathbf{P}^{[O]}, \quad (4.10)$$

the governing equations of the system can be discerned directly, where dt represents the incremental time step of the data.

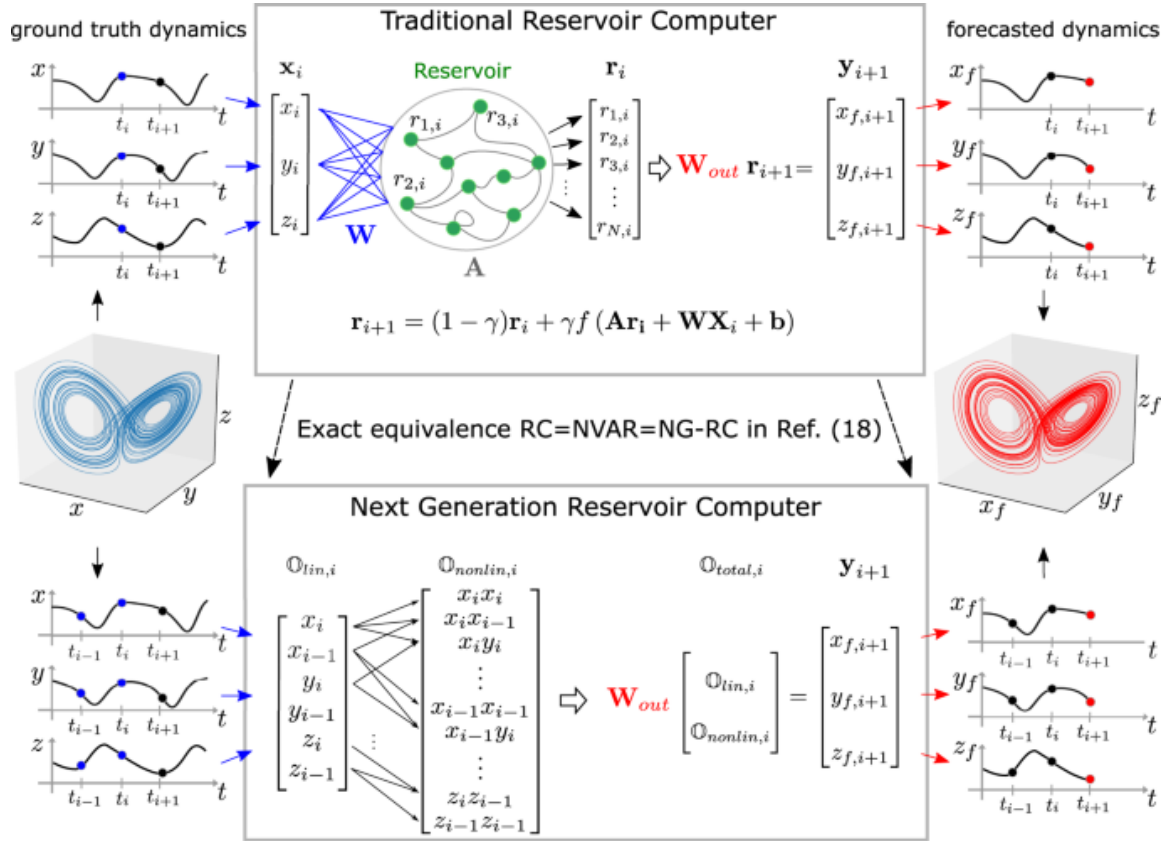


Figure 4.2: Next-Generation Reservoir Computing. The top part of the Figure shows a classical RC processing time series data of the Lorenz attractor (blue, center left). The predicted form of this attractor (red, middle right) is derived from a linear combination of the states within the reservoir. On the other hand, the lower part shows NG-RC, which generates predictions by using a linear combination of time-delayed states from the time series data. It is worth noting that two such time-delay states are demonstrated here. Further enhancing its capabilities, the NG-RC incorporates nonlinear transformations of the data, with the quadratic function demonstrated as an example in this context. Adapted from Gauthier, Bollt, Griffith, et al. [120].

Sparse Identification of Nonlinear Dynamics

Sparse Identification of Nonlinear Dynamics (SINDy) was introduced by Brunton, Proctor, & Kutz [91] in 2016 as a powerful technique for system identification and model discovery from data. The motivation behind SINDy was to develop a method that can identify the governing equations of a system directly from data, without requiring prior knowledge of the system's dynamics. By leveraging sparsity-promoting techniques and optimization algorithms, SINDy has proven to be effective in capturing the essential dynamics of complex systems.

The mathematical theory of SINDy revolves around the concept of sparse regression, where the goal is to identify a parsimonious set of governing equations that describe the dynamics of a system. Given a set of state variable measurements of a dynamical system $\mathbf{x}(t)$, the goal of SINDy is to find a sparse representation of the governing equations in the form:

$$\dot{\mathbf{x}}(t) = \Theta(\mathbf{x}(t))\xi, \quad (4.11)$$

where $\dot{\mathbf{x}}(t)$ is the time derivative of the state variables. $\Theta(\mathbf{x}(t))$ is a *library* of candidate functions constructed from the state measurements. It can include polynomial terms, trigonometric functions, or any other relevant nonlinear terms. ξ is a coefficient vector, which will be sparse, indicating that only a few terms in the library Θ are required to represent the dynamics.

Algorithm 5 Sparse Identification of Nonlinear Dynamics

- 1: **Data Collection.** Gather measurement data $\mathbf{x}(t)$ and compute its time derivative $\dot{\mathbf{x}}(t)$.
- 2: **Library Construction.** Construct the library of candidate functions $\Theta(\mathbf{x}(t))$. This can be constructed by considering a set of potential functions of the state variables, such as:

$$\Theta(\mathbf{x}) = \left[1 \quad \mathbf{x} \quad \mathbf{x}^2 \quad \sin(\mathbf{x}) \quad \dots \right]^T. \quad (4.12)$$

- 3: **Sparse Regression.** Solve the following optimization problem:

$$\min_{\xi} \|\dot{\mathbf{x}} - \Theta(\mathbf{x}(t))\xi\|_2^2 + \lambda\|\xi\|_1, \quad (4.13)$$

where $\|\cdot\|_2^2$ is the L2 norm, capturing the least squares fit. $\|\cdot\|_1$ is the L1 norm, promoting sparsity in the coefficients ξ . λ is a regularization parameter controlling the trade-off between fit and sparsity.

- 4: **Model Extraction.** From the sparse coefficient vector ξ , we extract the governing differential equations. Only the non-zero coefficients in ξ will correspond to the terms in the equations.
-

The effectiveness of the SINDy algorithm largely depends on the library Θ that is chosen [128]. A poorly constructed library might not capture the true dynamics, leading to imprecise outcomes. Equally crucial is the determination of the regularization parameter λ . In situations where λ is too small, the algorithm may produce overly complex models, a classic sign of overfitting. Conversely, an excessively large λ might yield an oversimplified model [129]. Another aspect to be wary of, is the presence of noise in the input data. SINDy, being a data-driven method, is susceptible to noisy data. To counteract this, one can employ strategies such as data filtering or the application of more resilient norms [130].

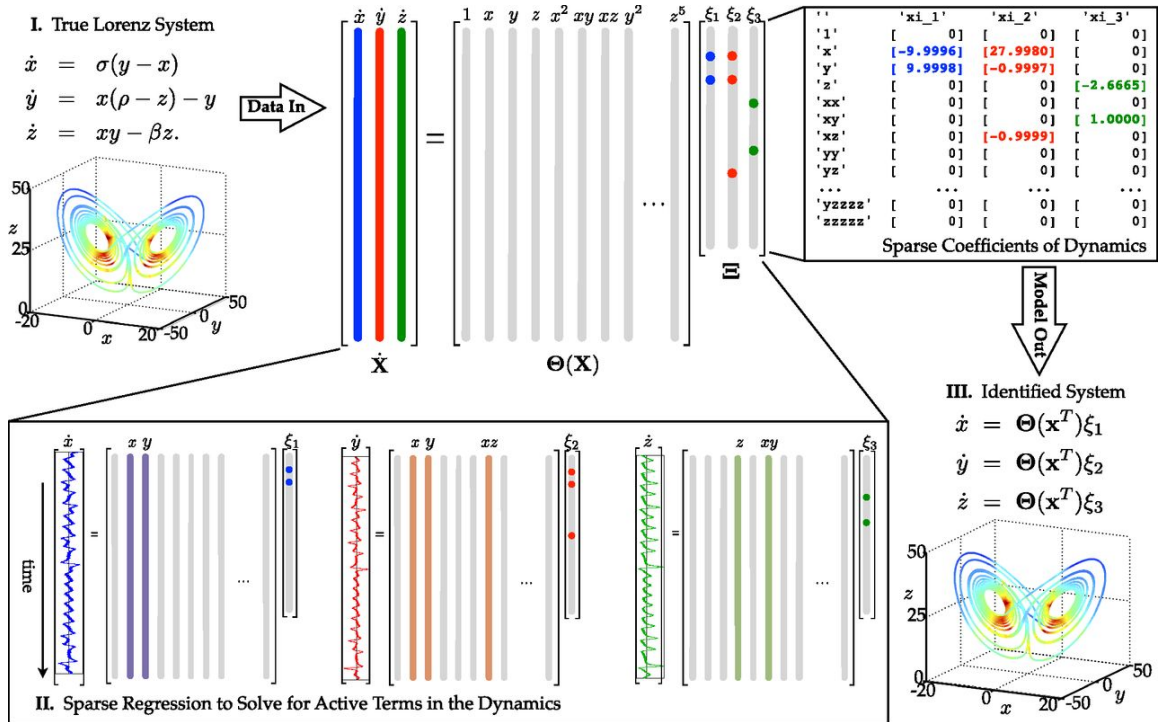


Figure 4.3: Sparse Identification of Nonlinear Dynamics. This Figure illustrates the SINDy algorithm implemented on the Lorenz system. The data is first collected from the system, comprising of both the states of \mathbf{X} and its derivatives of $\dot{\mathbf{X}}$. Following that, a library of nonlinear functions of the states, denoted $\Theta(\mathbf{X})$, is formulated. This library facilitates the identification of the minimal set of necessary terms to satisfy the equation $\dot{\mathbf{X}} = \Theta(\mathbf{X})\xi$. Within the vectors of ξ , which are determined through sparse regression, the few non-zero entries identify the important terms that make up the dynamic right-hand side of the system. Adapted from Brunton, Proctor, & Kutz [91].

4.3 Evaluating Predictions of Chaotic Systems

When predicting nonlinear dynamical systems with chaotic attractors, our objective exceeds the mere replication of short-term paths. It is equally essential to replicate the long-term statistical features of the system, often referred to as its *climate*. This emphasis stems from the fact that chaotic systems are intrinsically sensitive to initial conditions, which means that minor deviations can escalate exponentially over time. Even with flawless short-term forecasting, numerical inaccuracies can still cause the predicted path to diverge from the actual trajectory. However, in many cases, this divergence is insignificant as long as the predicted trajectory remains aligned with the same attractor. To accurately assess this complex dynamic, it is essential to use quantitative metrics that capture the multifaceted behavior of the system. Thus, adhering to the approach of Haluszczyński & R ath [35], we employ the following metrics: *Lyapunov exponents*, *correlation dimension*, and *forecast horizon*.

Lyapunov Exponents

Lyapunov exponents are named in honor of Russian mathematician Aleksandr Mikhailovich Lyapunov, who in the late 19th century conducted research on motion stability [3]. Instead of determining precise solutions, Lyapunov focused on examining their stability and created methods to evaluate the stability of equilibrium points. The concept of the Lyapunov exponent, which measures the average exponential divergence or convergence of trajectories within a phase space, first emerged in the 20th century. Its importance grew with the discovery of deterministic chaos, where it serves as a chaos quantifier. Today, the Lyapunov exponent has proven invaluable in nonlinear dynamics and is utilized across a wide range of scientific fields.

For a dynamical system trajectory $x(t)$ described by $\dot{x} = f(x)$, the i^{th} Lyapunov exponent, λ_i , is articulated as:

$$\lambda_i = \lim_{t \rightarrow \infty} \frac{1}{t} \ln \left(\frac{\|\delta x_i(t)\|}{\|\delta x_i(0)\|} \right), \quad (4.14)$$

with $\delta x_i(t)$ denoting minor perturbations in the i^{th} direction at a given time t .

An n -dimensional system yields n Lyapunov exponents. A sorted spectrum of these exponents paints a thorough picture of system dynamics, with interpretations as:

- $\lambda_i > 0$: Suggests trajectories diverging in the i^{th} direction, hinting at chaos.
- $\lambda_i = 0$: A neutral stance, often present in quasi-periodic systems.
- $\lambda_i < 0$: Signals trajectory convergence in the i^{th} direction, signifying stability.

The foremost focus in this work is the *largest Lyapunov exponent*, λ_{max} . A positive value confirms chaotic system dynamics. Its magnitude offers insight into trajectory divergence rates:

$$d(t) = C \cdot e^{\lambda_{max} \cdot t}, \quad (4.15)$$

Here, $d(t)$ represents the phase space distance between two initially proximate states, with C being the constant indicating the initial separation. Hence, by solely determining the largest exponent, one can largely decipher the underlying system behavior. The *Rosenstein* algorithm is the standard method for estimating λ_{max} from time series data [131]:

Algorithm 6 Rosenstein Algorithm

- 1: **Time Series Embedding.** Embed the time series $x(t)$ into higher-dimensional spaces using the embedding dimension κ and time delay τ :

$$X(t) = [x(t), x(t + \tau), x(t + 2\tau), \dots, x(t + (\kappa - 1)\tau)]. \quad (4.16)$$

- 2: **Nearest Neighbors.** For each point X_i in the phase space, find its nearest neighbor X_j (excluding temporally close points to avoid autocorrelation effects). The Euclidean distance metric is typically used to determine closeness.
- 3: **Time Evolution.** Track the distance between each point X_i and its nearest neighbor X_j as they evolve in time. More precisely, compute the average divergence $d(t)$ of pairs of trajectories over time t :

$$d(t) = \frac{1}{N-t} \sum_{i=1}^{N-t} \ln \|X_{i+t} - X_{j+t}\|, \quad (4.17)$$

where N is the total number of data points.

- 4: **Linear Fit.** Plot the average logarithmic divergence $\ln(d(t))$ against time t . The slope of the linear region of the resulting divergence plot provides an estimate for the largest Lyapunov exponent:

$$\lambda_{max} = \frac{1}{t} \ln(d(t)). \quad (4.18)$$

This process is illustrated in Figure 1.1.

Correlation Dimension

The concept of correlation dimension stems from the larger investigation of fractal geometry and the necessity of measuring the intricacy of strange attractors in chaotic dynamical systems. In the late 1970s to early 1980s, Grassberger & Procaccia [132] introduced this measure, which has since become fundamental in chaos theory. The correlation dimension allows for the categorization of an attractor's structure by determining its *fractional* dimensionality.

The correlation dimension ν is one of the scaling dimensions used to characterize chaotic dynamical systems and is especially useful when examining the geometry of strange attractors. While an attractor might be embedded in a higher-dimensional phase space, its actual structure might be constrained to a fractional or non-integer dimension. This fractional dimensionality indicates that the attractor fills its embedding space in a complex, fractal manner. Formally, the correlation dimension is defined using the correlation integral $C(r)$:

$$C(r) = \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i=1}^N \sum_{j=1}^N \Theta(r - |x_i - x_j|) \quad (4.19)$$

where Θ is the Heaviside step function, r is a small distance, and x_i and x_j are state vectors from the time series. The correlation dimension ν is then defined as:

$$\nu = \lim_{r \rightarrow 0} \frac{\log(C(r))}{\log(r)} \quad (4.20)$$

The *Grassberger-Procaccia* algorithm is a widely recognized method used for estimating the correlation dimension:

Algorithm 7 Grassberger-Procaccia Algorithm

- 1: **Time Series Embedding.** Embed the time series x into higher-dimensional spaces using the embedding dimension κ and time delay τ :

$$X(t) = [x(t), x(t + \tau), x(t + 2\tau), \dots, x(t + (\kappa - 1)\tau)]. \quad (4.21)$$

- 2: **Distance Calculation.** For each point X_i in this reconstructed phase space, compute the distance to every other point X_j in the space. Usually, the Euclidean distance is used:

$$d(X_i, X_j) = \sqrt{\sum_{k=1}^m (X_{i,k} - X_{j,k})^2}. \quad (4.22)$$

- 3: **Pairs Counting.** For a range of length scales r , count the number of point pairs (X_i, X_j) such that their distance $d(X_i, X_j)$ is less than r . The result is a function $C(r)$, where:

$$C(r) = \frac{2}{N(N-1)} \sum_{i=1}^N \sum_{j=i+1}^N \Theta(r - d(X_i, X_j)). \quad (4.23)$$

- 4: **Linear Fit.** Plot $\ln(C(r))$ as a function of $\ln(r)$. In the resulting plot, if the system behaves chaotically or has a fractal structure, there will be a linear region. The slope of this linear region gives the estimate of the correlation dimension ν :

$$\nu = \lim_{r \rightarrow 0} \frac{\ln(C(r))}{\ln(r)}. \quad (4.24)$$

Forecast Horizon

To evaluate the accuracy of our short-term trajectory predictions, we use the *forecast horizon*, denoted by τ , as introduced by Haluszczyński & R ath [35]. This method determines the consecutive time intervals during which the predicted trajectory $\mathbf{y}_{pred}(t)$ varies from the reference test data $\mathbf{y}_{test}(t)$ by a margin below the standard deviation of the test dataset, σ_{test} . Thus, the forecast horizon indicates the time duration during which the gap between the predicted trajectory and the actual trajectory remains within acceptable limits. For each coordinate i this condition can be expressed mathematically as:

$$|y_{pred,i}(t) - y_{test,i}(t)| < \sigma_{test,i}, \quad (4.25)$$

To relate the forecast horizon to the underlying chaotic dynamics of the system, we rescale it using the Lyapunov exponent. This expresses the forecast horizon in multiples of the Lyapunov time [133]. In detail, we scale the forecast horizon by multiplying the time discretization, dt , and the largest Lyapunov exponent of the test data $\lambda_{max,test}$:

$$\tau \mapsto \tau \cdot dt \cdot \lambda_{max,test}. \quad (4.26)$$

Our objective is to ensure that any small deviations around the actual trajectory do not prompt an immediate identification of the prediction as non-conforming. This metric evaluates the extent to which our prediction aligns with the correct trajectory before the system's inherent chaos triggers a rapid divergence. It should be noted that the trajectory distances typically exceed the threshold values as soon as there is any error in the short-term prediction due to the differing ranges of state variables across different systems.

4.4 Block-Diagonal Reservoirs

One central objective of this research is to validate the ability to divide a reservoir into numerous sub-reservoirs without compromising its predictive power. To achieve this, we utilize a block diagonal matrix as the topology for our $d \times d$ dimensional reservoir. The matrix is composed of \mathbf{J}_i blocks, each with dimensions of $b \times b$, where $i \in \{1, 2, \dots, \lfloor \frac{d}{b} \rfloor\}$. Hence, each block corresponds to an independent sub-reservoir:

$$\mathbf{A} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{\lfloor \frac{d}{b} \rfloor} \end{pmatrix}. \quad (4.27)$$

As previously mentioned, the reservoir topology is rescaled to a target spectral radius ρ^* . Thus, it is essential to compute the spectral radius and, consequently, the eigenvalues of the matrix \mathbf{A} . The time complexity for computing the eigenvalues of a d -dimensional matrix through bi-diagonalization is $\mathcal{O}(d^3)$ [134]. However, the eigenvalues of a block diagonal matrix are the list of eigenvalues of the blocks:

$$\rho(\mathbf{A}) = \max \left\{ \rho(\mathbf{J}_1), \dots, \rho\left(\mathbf{J}_{\lfloor \frac{d}{b} \rfloor}\right) \right\}, \quad (4.28)$$

which speeds up the computation by a factor of:

$$\frac{d^3}{b^3 \cdot \lfloor \frac{d}{b} \rfloor} \approx \left(\frac{d}{b}\right)^2. \quad (4.29)$$

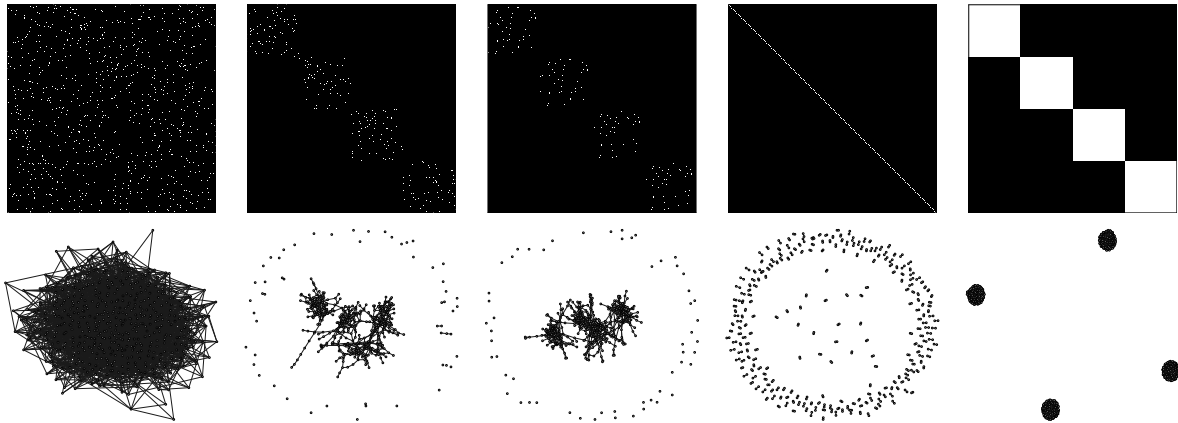


Figure 4.4: Reservoir Topologies. This Figure depicts different reservoir topologies containing $d = 500$ nodes. The top row displays the network connections via white entries. The bottom row presents the corresponding spring-layouts [135], where each node is a white circle outlined in black, and connections are depicted by black lines. The first topology is the ordinary topology utilizing an Erdős-Rényi graph. The second topology is a block-diagonal topology comprising different Erdős-Rényi graphs with a size of $b = 125$ as blocks. The third topology is also a block-diagonal topology but utilizes equal-sized Erdős-Rényi graphs with a size of $b = 125$ as blocks. The fourth topology uses matrices of ones with a size of $b = 2$ as blocks in a block-diagonal topology. Finally, the fifth topology uses matrices of ones with a size of $b = 125$ as blocks in a block-diagonal topology.

Blocks of Erdős–Rényi Networks

First, we choose the individual blocks \mathbf{J}_i as Erdős–Rényi networks [112]. In our analysis, we differentiate between two cases:

- *Individual Blocks*: each block, \mathbf{J}_i , is constructed separately using a distinct random seed.
- *Equal Blocks*: all blocks are identical, hence only one construction of the Erdős–Rényi network is needed. Furthermore, this leads to an additional acceleration in the eigenvalue computation by a factor of $\lfloor \frac{d}{b} \rfloor$:

$$\mathbf{J}_1 = \mathbf{J}_2 = \dots = \mathbf{J}_{\lfloor \frac{d}{b} \rfloor}.$$

Blocks of Matrices of Ones

To eliminate the reservoir’s randomness, we construct each block \mathbf{J}_i as a matrix of ones. This approach has multiple implications. Firstly, there is no need to calculate the reservoir’s spectral radius $\rho(\mathbf{J})$ because it is equivalent to the block size b :

$$\mathbf{A} = \frac{\rho^*}{b} \mathbf{J}. \quad (4.30)$$

Furthermore, this reservoir architecture implies that in each iteration, each block \mathbf{J}_i functions as an averaging operator on the reservoir states, resulting in a reduction of dimensionality. This mechanism is akin to the *average pooling layers* of other machine learning methodologies in which the features that are accentuated are more robust and less susceptible to noise [136]. The mean between the i^{th} and j^{th} row of the reservoir state $\mathbf{r}(t)$ is denoted as:

$$\bar{r}_{i:j}(t) \equiv \frac{1}{j-i+1} \sum_i^j r_i(t). \quad (4.31)$$

Each block produces a vector of size $b \times 1$ with identical values. For instance, the initial row of the multiplication $\mathbf{J} \cdot \mathbf{r}(t)$ is:

$$\underbrace{[1, \dots, 1]}_b \underbrace{[0, \dots, 0]}_{d-b} \cdot \mathbf{r}(t) = \sum_{i=1}^b r_i(t) = b \cdot \bar{r}_{1:b}(t). \quad (4.32)$$

This is repeated for the first b rows. Thus, the product of the reservoir \mathbf{A} and $\mathbf{r}(t)$ from Equation 4.42 produces:

$$\mathbf{A} \cdot \mathbf{r}(t) = \frac{\rho^*}{b} \mathbf{J} \cdot \mathbf{r}(t) = \rho^* \begin{pmatrix} \bar{r}_{1:b}(t) \\ \bar{r}_{1:b}(t) \\ \vdots \\ \bar{r}_{(i-1) \cdot b + 1 : i \cdot b : k \cdot b}(t) \\ \bar{r}_{(i-1) \cdot b + 1 : i \cdot b : k \cdot b}(t) \\ \vdots \\ \bar{r}_{d-b+1:d}(t) \\ \bar{r}_{d-b+1:d}(t) \\ \vdots \end{pmatrix}. \quad (4.33)$$

Therefore, the multiplication contribution of the reservoir is identical for every block, resulting in uniform reservoir memory for each training step. Consequently, this leads to a reduction in computational expenses.

4.5 Minimal Reservoir Computing

While block-diagonal and binary reservoirs do enhance the computational efficiency of RC, the randomness of the input weights remains a factor of uncertainty. While the use of randomness has demonstrated success in certain tasks, its effectiveness is inconclusive [35]. Predicting how the performance of the system will be affected by the reservoir’s topology is unattainable before any experimentation takes place. Moreover, the random components of RC pose challenges in interpretation despite its simple architecture. Therefore, we recommend slight modifications to the standard RC architecture that not only augment transparency but also eliminate randomness, following the methods proposed by Gauthier, Bollt, Griffith, et al. [120] and Brunton, Proctor, & Kutz [91]. To aid comprehension, we first provide a brief overview of the adjustments and the algorithm for minimal RC at a high level:

Algorithm 8 Minimal Reservoir Computing

- 1: **Combinatory Input Weights.** The input weights \mathbf{W}_{in} are designed so that each combination of the coordinates of the data is fed into the reservoir separately.
- 2: **Block-Diagonal and Binary Reservoir.** The reservoir \mathbf{A} is a block-diagonal matrix consisting of matrices of ones with block size b .
- 3: **Linear Reservoir States.** We do not use a nonlinear activation function in order to construct the reservoir states $\mathbf{r}(t)$. Hence the iterative update equation reduces to:

$$\mathbf{r}(t+1) = \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t), \quad (4.34)$$

where $\mathbf{u}(t)$ denotes the training data at time t . It is important to note that the reservoir states are only linear. For the specific case where the target spectral radius is $\rho^* = 0$, this means that linear combinations of the data are fed directly into the reservoir.

- 4: **Nonlinear Readout.** Instead of only inserting the squared reservoir states [126], our generalized states $\tilde{\mathbf{r}}$ contain all orders up to a nonlinearity degree η :

$$\tilde{\mathbf{r}} = \{\mathbf{r}, \mathbf{r}^2, \dots, \mathbf{r}^{\eta-1}, \mathbf{r}^\eta\}. \quad (4.35)$$

Therefore, the readout fully captures the nonlinearity of the data.

- 5: **Training.** As in traditional RC, we stack the training data \mathbf{u} and the corresponding reservoir states $\tilde{\mathbf{r}}$ to matrices \mathbf{U} and $\tilde{\mathbf{R}}$ respectively. We subsequently solve the Equation $\mathbf{W}_{out} \cdot \tilde{\mathbf{R}} = \mathbf{U}$ through ridge regression as described in Equation 4.5.
- 6: **Prediction.** The prediction process for the reservoir states remains the same, utilizing the updated equation for updating:

$$\mathbf{r}(t) = \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{W}_{out} \cdot \tilde{\mathbf{r}}(t). \quad (4.36)$$

Note that the reservoir \mathbf{A} only operates on the *simple* reservoir state \mathbf{r} . The second term acting on \mathbf{W}_{out} is $\tilde{\mathbf{r}}$, which includes all the nonlinear powers. The predicted time series $\mathbf{y}(t)$ can be obtained through multiplication.

$$\mathbf{y}(t) = \mathbf{W}_{out} \cdot \tilde{\mathbf{r}}(t). \quad (4.37)$$

Combinatory Input Weights

To feed the input data $\mathbf{u}(t)$ into the reservoir, an input weight matrix \mathbf{W}_{in} is defined, which determines how strongly each coordinate influences each node of the reservoir network. In a traditional RC, the elements of \mathbf{W}_{in} are uniformly distributed random numbers between $[-1, 1]$. In our novel framework, we follow a structured approach for selecting the elements instead of random selection. Firstly, to eliminate randomness, we select a set of equally spaced values between 0 and 1 for each block size of b :

$$\mathbf{w} = (w_1, w_2, \dots, w_b)^T = \left(1, \frac{b-2}{b-1}, \dots, \frac{1}{b-1}, 0\right)^T. \quad (4.38)$$

To avoid non-invertible matrices and numeric instabilities, we take the square root values of all weights $\mathbf{w} = (\sqrt{w_1}, \dots, \sqrt{w_b})^T$. Then we specifically structure the input matrix so that the different combinations of input data coordinates, also called *features*, are fed separately into the reservoir. In the case of a 3-dimensional system with coordinates $\mathbf{u}(t) = (x, y, z)^T(t)$ and a nonlinearity order $\eta = 2$, the input matrix (multiplication) looks like:

$$\mathbf{W}_{in} \cdot \mathbf{u}(t) = \begin{pmatrix} \mathbf{w} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w} \\ \mathbf{w} & \mathbf{w} & \mathbf{0} \\ \mathbf{w} & \mathbf{0} & \mathbf{w} \\ \mathbf{0} & \mathbf{w} & \mathbf{w} \end{pmatrix} \cdot \mathbf{u}(t) = \begin{pmatrix} x \\ y \\ z \\ x+y \\ x+z \\ y+z \end{pmatrix} (t) \otimes \mathbf{w}, \quad (4.39)$$

where \otimes denotes the tensor product and hence each block represents one feature f . For n -dimensional data the feature space contains $n_f = 2^n - 2$ elements. Thus, the dimension of the reservoir is $d = n_f \cdot b$.

Block-Diagonal and Binary Reservoir

To maintain the independence of individual features, we choose the reservoir as a block diagonal matrix composed of blocks of ones of size b , as discussed in Section 4.4. As a result, each block \mathbf{J}_i is specifically associated with a particular feature:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{n_f} \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{J}_x & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_y & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{y+z} \end{pmatrix}. \quad (4.40)$$

As previously mentioned, we adjust the spectral radius $\rho(\mathbf{J})$ of the reservoir to match a target spectral radius ρ^* .

Linear Reservoir States

As is typical in traditional reservoir RC, we use a recurrent update equation to capture the dynamics of the system in the so-called reservoir states $\mathbf{r}(t)$. To accomplish this, a nonlinear activation function $g(\cdot)$, which captures the nonlinear properties of the data, is normally required.

However, as previously noted, we shift the nonlinearity entirely to the readout. Consequently, state evolution over time is determined through iterative processes:

$$\mathbf{r}(t+1) = g(\mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t)) \quad (4.41)$$

$$\longrightarrow \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t). \quad (4.42)$$

Due to our architecture selection, the states of the reservoir can be acquired separately for each feature:

$$\mathbf{r}(t) = \begin{pmatrix} \mathbf{r}_x \\ \mathbf{r}_y \\ \mathbf{r}_z \\ \mathbf{r}_{x+y} \\ \mathbf{r}_{x+z} \\ \mathbf{r}_{y+z} \end{pmatrix} (t) = \begin{pmatrix} \mathbf{r}_{x,1} \\ \mathbf{r}_{x,2} \\ \vdots \\ \mathbf{r}_{x,b} \\ \mathbf{r}_{y,1} \\ \vdots \\ \mathbf{r}_{y+z,b} \end{pmatrix} (t). \quad (4.43)$$

Therefore, we can gather all reservoir states that belong to a single feature $\mathbf{r}_f(t)$ — which are referred to as *feature states* — and examine them independently. This allows us to comprehend that the reservoir serves as an averaging mechanism for the feature states:

$$\mathbf{A}_f \cdot \mathbf{r}_f(t) = \left(\frac{\rho^*}{b} \sum_{i=1}^b \mathbf{r}_{f,i}(t) \right) \cdot I_b = \rho^* \cdot \bar{\mathbf{r}}_f(t), \quad (4.44)$$

where I_b is a vector of ones of size b . Thus, in each iteration, the features' states are normalized to the average of previous feature states $\bar{\mathbf{r}}_f(t)$ while adding a varying strength, determined by the input weight, to the new feature data $f(t)$:

$$\mathbf{r}_f(t+1) = \rho^* \cdot \bar{\mathbf{r}}_f(t) + \mathbf{w} \cdot f(t). \quad (4.45)$$

where f can be replaced by any other feature without loss of validity. The average, or *memory*, of each feature is monitored in the final row of the feature states since $w_b = 0$. This is indicative of the target spectral radius ρ^* and its impact on the retention of data memory during each iteration step.

Nonlinear Readout

While a quadratic readout, specifically the squared reservoir states \mathbf{r}^2 , is frequently included in a traditional RC to eliminate the symmetry of the activation function [126], a readout that can capture the nonlinearity of the data is necessary. As a result, we add even higher orders of nonlinearity to the so-called generalized states $\tilde{\mathbf{r}}$. For a given degree of nonlinearity η , they take the following form:

$$\tilde{\mathbf{r}} = \{\mathbf{r}, \mathbf{r}^2, \dots, \mathbf{r}^{\eta-1}, \mathbf{r}^\eta\}. \quad (4.46)$$

Therefore, given a degree of nonlinearity η and a block size of b , the number of elements in the readout, which also corresponds to the number of parameters to optimize, equals:

$$n_{out} = (2^\eta - 2) \cdot \eta \cdot b = \left(\sum_{k=1}^{\eta} \binom{\eta}{k} - 1 \right) \cdot \eta \cdot b, \quad (4.47)$$

which we rewrite to binomial coefficients for better comparison.

When dealing with high-dimensional data possessing a high level of nonlinearity, fewer variables need to be optimized as compared to comparable predictive models like NG-RC [120] or SINDy [91]. This is because NG-RC and SINDy require combinations of lagged time series, resulting in a larger feature space. For a system of dimension n and a nonlinearity degree η the number of features is (at least):

$$n_f = \sum_{k=1}^{\eta} \binom{n+k-1}{k} = \binom{n+\eta}{\eta} - 1, \quad (4.48)$$

which grows much faster for larger n and η than the expression for n_{out} in Equation 4.47.

Addressing Collinearity

To address the challenge of collinearity in linear regression models, which can lead to instability in the coefficient estimates and difficulty in interpreting the results, one effective strategy is to eliminate linearly dependent features. This approach ensures a more robust model without sacrificing predictive performance. By including this step before the training of minimal RC, we can further enhance prediction performance and robustness.

In the context of a three-dimensional system, consider the state variables x , y , and z . Linear combinations such as $x + y$, $x + z$, and $y + z$ can be expressed as linear dependencies on the original variables x , y , and z when the coefficient η is equal to 1. By omitting these linearly dependent combinations, we maintain the integrity of the model and ensure a more accurate and interpretable representation of the underlying relationships in the data.

4.6 Prediction Results and Parameter Robustness

In this dissertation, we introduce block-diagonal reservoirs, indicating the possibility of composing a reservoir from smaller ones, each with distinctive dynamics. Additionally, we eliminate the randomness of the reservoir by using matrices of ones for the individual blocks. This deviates from the common interpretation of the reservoir as a single network. On the example of the Lorenz and Halvorsen systems, we examine the performance of block-diagonal reservoirs and their sensitivity to hyperparameters. We discover that the performance is comparable to sparse random networks and explore the implications of scalability, explainability, and hardware realizations of reservoir computers.

Additionally, we introduce a new RC framework that surpasses similar approaches in both short- and long-term forecasting with an equivalent demand for minimal training data and computational resources. The architecture is modified by restructuring the input weights and the reservoir so that combinations of input data coordinates are fed into the reservoir separately. Thus, a block-diagonal matrix of ones serves as the reservoir, operating as an averaging mechanism for the reservoir states during each update step. Comparable to average pooling layers in other machine learning approaches, this can be explained as a technique for mainly acquiring features that demonstrate greater robustness. Instead of utilizing a nonlinear activation function for generating the reservoir states, we capture the nonlinearity of the data in the readout layer by adding higher orders of the reservoir states before performing ridge regression. Using the Lorenz system and other synthetic chaotic systems as examples, we show that these changes lead to excellent short- and long-term predictions that significantly outperform traditional RC, NG-RC, and SINDy.

While the evaluation of prediction performance on chaotic systems is typically assessed through visual means, we utilize three quantitative metrics described in Section 4.3: the largest Lyapunov exponent, correlation dimension, and forecast horizon. Our results are also rigorously validated through the use of multiple attractor starting points, varying training data size, and discretizations.

Block-Diagonal Reservoirs

In this Section, we present the results of our investigation into variations of different parameters that demonstrate the strength of our improved design. The parameters that underwent changes are as follows:

- *Network Dimension*: we vary the network dimension between $d \in \{400, 450, \dots, 600\}$ as these are reasonable values in RC research. We intentionally choose multiples of 50 to ensure a sufficient number of block sizes b , as determined by $\lfloor \frac{d}{b} \rfloor$.
- *Block-Size*: we vary the block-size b by setting it to all divisors of the network dimension d , excluding the divisor $b = 1$ as it essentially removes the reservoir. Additionally, we do not utilize $b = d$ as it represents a single network, similar to the traditional architecture.
- *Target Spectral Radius*: we vary the target spectral radius from $\rho^* \in \{0.1, 0.2, \dots, 2.0\}$ and find that, similarly to the traditional architecture, the target spectral radius of $\rho^* = 0.1$ yields the most robust prediction results. Henceforth, we set $\rho^* = 0.1$ as default.
- *Attractor Starting Points*: we choose 500 different starting points on the attractor.
- *Random Seeds*: we choose 100 different random seeds across all components of the RC architecture which have randomness — the input weights \mathbf{W}_{in} and the reservoir \mathbf{A} for block-diagonal Erdős–Rényi networks.

We solve the differential equations of the synthetic system utilizing the Runge-Kutta method [18] for 70,000 steps with a discretization of $dt = 0.02$. We exclude the initial transient of $T = 50,000$ steps to ensure a sufficient manifestation of the attractor. The remaining steps are used for training $T_{train} = 10,000$ and testing $T_{test} = 10,000$ of the RCs. To ensure reliable results, we use rounded last points of one data sample as starting points for the next, thereby varying the attractor’s starting points. The Lorenz and Halvorsen systems are initiated with $(-14, -20, 25)$ and $(-6.4, 0, 0)$, respectively. This configuration is similar to those utilized by Griffith, Pomerance, & Gauthier [116] and Haluszczynski & R ath [35].

In order to make the figures easier to visualize, we calculate the fraction of connection and use the root of it as the x-axis:

$$\sqrt{\frac{b}{d}}. \quad (4.49)$$

Thus, the higher the fraction the bigger the blocks and the number of connections in the network. This is necessary because the number of divisors for each network size d is different and the divisors are not equally spaced.

Blocks of Erdős–Rényi Networks

As mentioned above, we distinguish between two cases for the Erdős–Rényi blocks. First, where all the blocks are individual networks and second, where all blocks are equal.

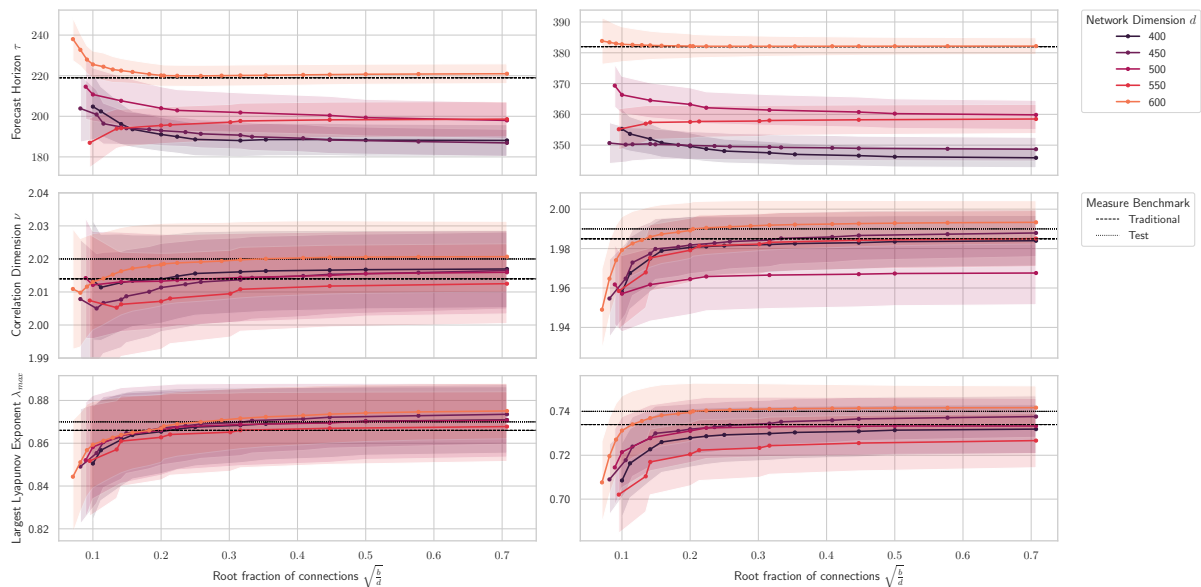


Figure 4.5: Prediction for Different Random Seeds of Individual Erdős–Rényi Blocks. This Figure shows the prediction performance of the Lorenz (left column) and Halvorsen (right column) systems for different random seeds of individual Erdős–Rényi blocks. The differently colored lines represent different network dimensions, and the corresponding shaded area represents the standard deviation over the variations. The lower, dashed line reflects the average prediction performance of the traditional RC architecture, while the higher, dotted line displays the average correlation dimension and largest Lyapunov exponent of the test data.

As shown in Figure 4.5, all prediction measures for individual blocks in the Lorenz and Halvorsen systems closely match the respective benchmark values of traditional RC and the test data. Furthermore, they even exceed these benchmarks for a network size of 600. Generally, we observe that small block-sizes have a worse long-term prediction quality with regards to the correlation dimension and the largest Lyapunov exponent. However, they appear to have a better short-term forecast horizon. The standard deviation over the variation of random seeds is comparable to the benchmarks. The results are illustrated in Figure 4.5.

Similar results can be observed for the equal blocks as shown in Figure 4.6 below — however, the standard deviation is slightly lower. This can be explained by the reduced level in randomness since only one block is randomly constructed.

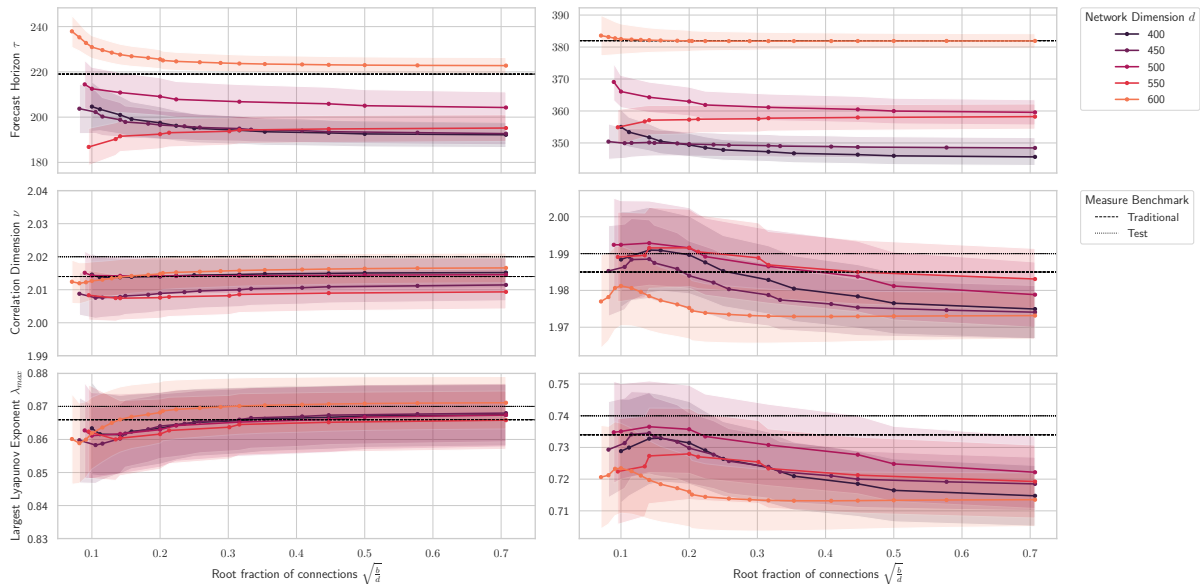


Figure 4.6: Prediction for Different Random Seeds of Equal Erdős–Rényi Blocks. This Figure shows the prediction performance of the Lorenz (left column) and Halvorsen (right column) systems for different random seeds of equal Erdős–Rényi blocks. The configuration is analogous to Figure 4.5.

Generally, we observe that the prediction performance stabilizes for $\sqrt{\frac{b}{d}} > 0.3$, for both individual and equal blocks. This has been detailed in Equation 4.29, which demonstrates that it quickens the calculation of the reservoir spectral radius by a factor of approximately 123 for individual blocks and 412 for equal blocks.

Furthermore, the modified architecture displays superior performance over traditional RC in long-term predictions, though slightly lower performance in short-term predictions. This is evident from the higher correlation dimensions and largest Lyapunov exponents, which surpass those of the traditional RC architecture and approach the *true* values of the test data.

Comparable short-term prediction performance can be achieved by increasing the network dimension to $d = 600$. This phenomenon indicates that the forecast horizon matches the traditional architecture’s value and outperforms it slightly for small block-sizes of $\sqrt{\frac{b}{d}} \approx 0.05$. Remarkably, the spectral radius’ calculation speed is increased by a factor of 160,000 and 3,200,000 for individual and equal blocks, respectively. This finding is valid for both the Lorenz and Halvorsen systems.

Blocks of Matrices of Ones

For the blocks of matrices of ones, the randomness of the reservoirs is removed. Therefore, our focus is on the remaining randomness present in the input weights \mathbf{W}_{in} and the variation of the starting point of the attractor.

We find that for varying the input weights, all prediction measures for both the Lorenz and Halvorsen systems are close to the respective benchmarks, and for some network sizes even exceed them. We observe in general that the performance is slightly worse than using blocks of Erdős–Rényi networks. As anticipated, the standard deviation over the input weight variation is comparable to the benchmarks and lower than for randomly constructed reservoirs. The results are shown in Figure 4.7.

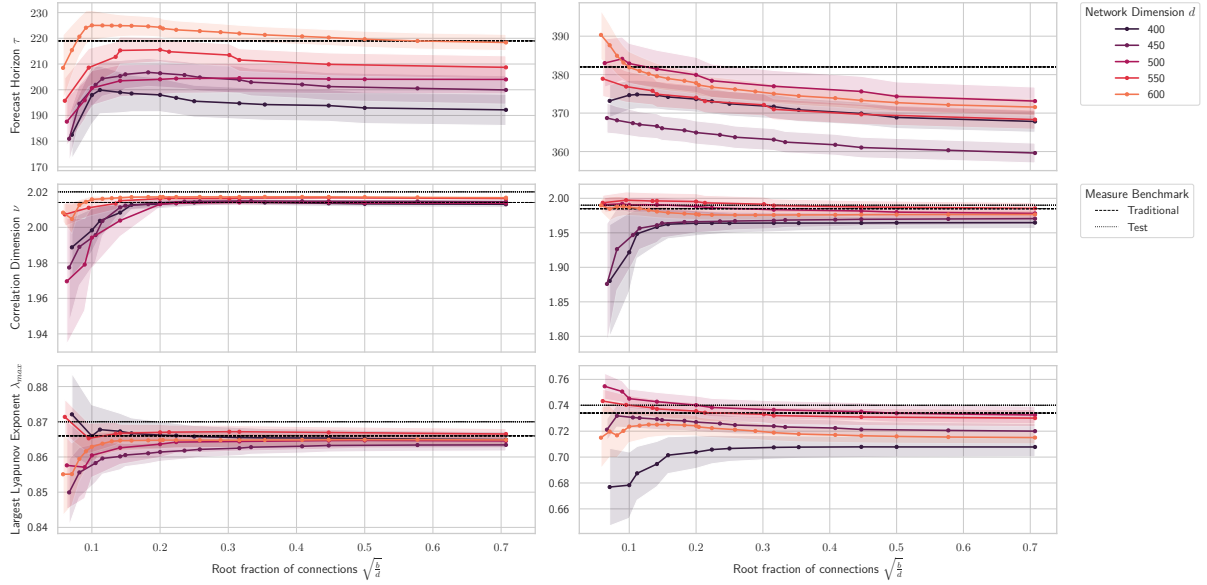


Figure 4.7: Prediction for Different Input Weights. This Figure shows the prediction performance of the Lorenz (left column) and Halvorsen (right column) systems for different input weights. The configuration is analogous to Figure 4.5.

A comparable behavior can be observed for the variation of the attractor starting points. Nonetheless, it is noteworthy that for certain block sizes, network dimensions of $d = 600$ surpass the benchmarks. Figure 4.8 depicts the results.

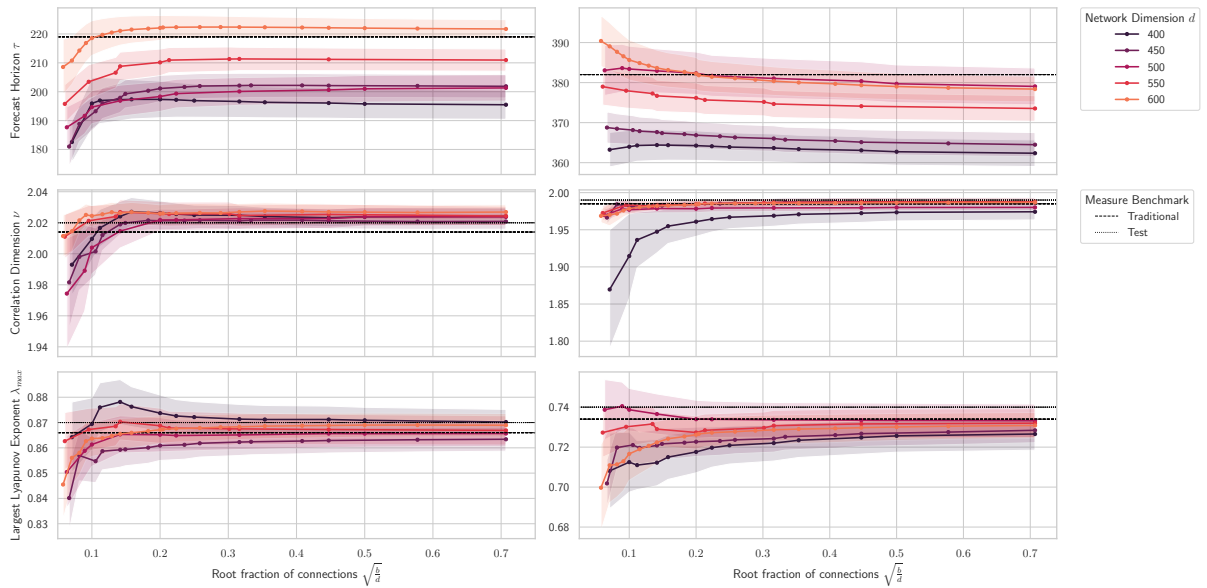


Figure 4.8: Prediction for Different Attractor Starting Points. This Figure shows the prediction performance of the Lorenz (left column) and Halvorsen (right column) systems for different starting points on the attractors. The configuration is analogous to Figure 4.5.

Overall, we observe that utilizing blocks of ones as the reservoir leads to stable prediction quality, even when input weights and attractor starting points are varied. For specific block-sizes, the modified architecture exceeds benchmark performance for both short- and long-term predictions.

Determining the most effective instance of these reservoirs can be accomplished through a few iterations while varying the block size for a sufficiently large network dimension. Since calculating the spectral radius of the reservoir is not necessary in this setup, fine-tuning the RC architecture through a parameter scan is both efficient and scalable.

Minimal Reservoir Computing

In this dissertation, we show how small changes to the traditional RC architecture can significantly improve its prediction capability of chaotic systems especially for low data requirements. Using the Lorenz system as an example of synthetic chaotic systems, we demonstrate that the mentioned modifications result in superior short- and long-term predictions, surpassing those produced by traditional RC, NG-RC, and SINDy models.

Therefore, similar to Gauthier, Boltt, Griffith, et al. [120], we use the minimal data setup for the Lorenz system with a discretization of $dt = 0.025$ and $T_{train} = 400$ training data points. The default RC architecture used in this work has block-size $b = 3$, spectral radius $\rho^* = 0.1$, and a nonlinearity degree $\eta = 2$. This equals 36 variables per coordinate.

In order to obtain robust results we repeat the analysis for 1,000 different starting points on the attractor and compare the prediction performance to the other models. In Figure 4.9 we see that the novel RC architecture significantly outperforms them with regards to short-term predictions with an average forecast horizon of approximately 7.0 Lyapunov times — this is around 2.5 times more than the averages of the other models.

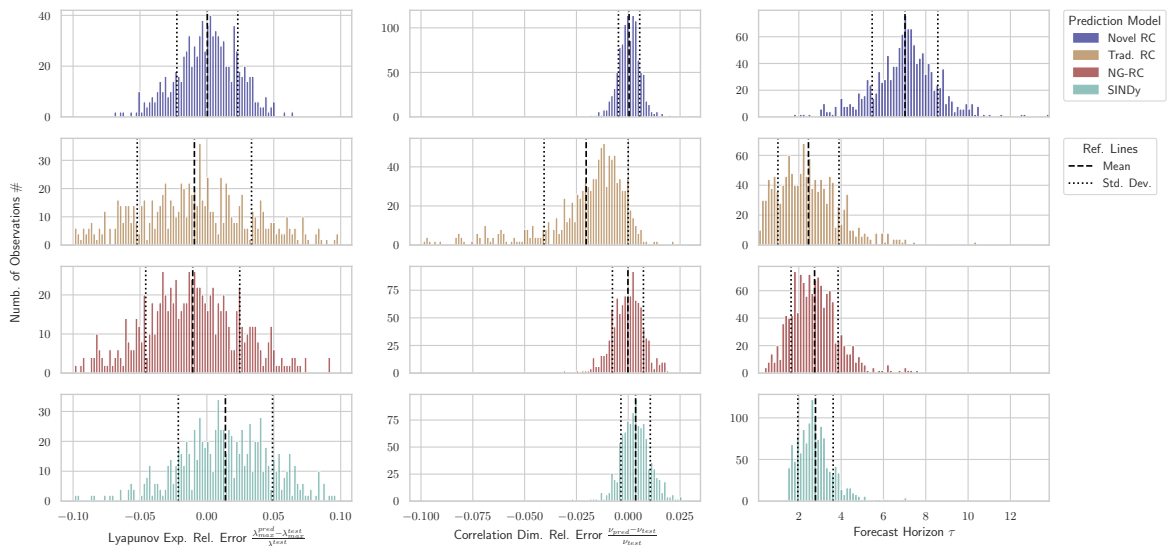


Figure 4.9: Prediction Performance of Minimal Reservoir Computing. This Figure shows the prediction measures (columns) of different models (rows) for 1000 different starting points on the Lorenz attractor. The relative error with respect to the test data is computed for the correlation dimension and the Lyapunov exponent. The mean and standard deviation of each distribution are indicated by a dashed and a dotted black line respectively.

The long-term prediction is marginally superior due to the average relative errors of the correlation dimension and the Lyapunov exponent, which are approximately $3.5 \cdot 10^{-4}$, respectively. This equates to around 9.0 and 39.7 times less than that of comparable models.

We verify the robustness of our novel RC to variations in discretization and length of training data. In Figure 4.10 we observe that it is quite robust and as expected, performs significantly better than comparable models especially with regards to short-term prediction. Here, we only see a decline in prediction performance for coarse discretizations $dt > 0.045$. The robustness of the long-term prediction is similar to traditional RC and SINDy. Interestingly, we see a decline in performance of NG-RC for larger training lengths $T_{train} > 700$ and finer discretizations $dt < 0.02$.

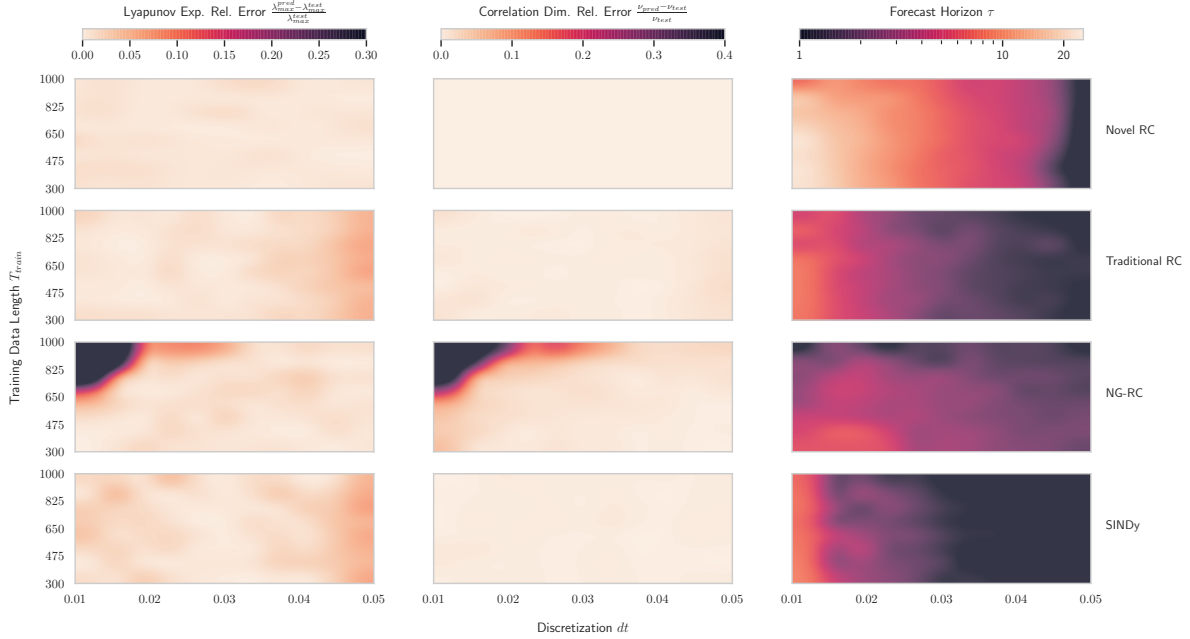


Figure 4.10: Prediction Performance for Different Data Discretizations and Lengths. This Figure displays the various prediction measures (columns) for differing models (rows) when using varying discretizations of the training data and lengths of the Lorenz system. The training data discretization (x-axis) and length (y-axis) varies between (0.01, 0.05) and (300, 1000) respectively. We calculate the relative error to the respective test data for the correlation dimension and Lyapunov exponent. The forecast horizon is scaled logarithmically in color. Each heatmap value represents the average of 100 variations of attractor starting points.

Furthermore, we find out the model to be reasonably robust to changes in hyperparameters and noise up to a SNR of ~ 38 dB. Additionally, we analyze the prediction performance of our model on different chaotic systems, which have different nonlinear behavior. We choose the models so that we can understand the inner workings of our RC better. For example, the Halvorsen system has only quadratic nonlinearities with no interacting coordinates and hence the input matrix only needs the first three blocks (which represent the distinct coordinates). Another example to point out is the *Rabinovich-Fabrikant* system [137], which has cubic nonlinearities. Here, we see that a nonlinearity degree of $\eta \geq 3$ is necessary for a reasonable prediction.

We addressed the issue of collinearity in further research, as outlined in Section 4.5. Through a recent parameter scan, we uncovered intriguing patterns in prediction performance. The results of the parameter analysis are illustrated in Figures 4.11, 4.12, and 4.13. The Figures display the largest Lyapunov exponent, correlation dimension, and forecast horizon for various parameterizations under minimal data requirements with $T_{\text{train}} = 400$ and $dt = 0.025$.

An important finding is the achievement of a decent prediction horizon of about 4 Lyapunov times for an even more minimal architecture with a block size of $b = 1$ and a target spectral radius of $\rho^* = 0$. This implies that linear combinations of the data can be directly fed into the reservoir, and that training a chaotic system with an OLS regression is sufficient. Additionally, we observe that the climate of the attractor can be accurately predicted using a minimal architecture with a block size of $b = 1$ and small degrees of nonlinearity. In this case, the extraction of recursive equations that reconstruct the Lorenz attractor becomes easily obtainable.

Furthermore, increasing the block size to 2 significantly improves the forecast horizon for nonlinearity of $\eta > 3$. Certain parameterizations of the target spectral radius and regularization parameter allow for a forecast horizon to extend beyond 10 Lyapunov times. Therefore, incorporating input weights with a value of 0 to average the feature states, as described in Section 4.5, yields remarkable results.

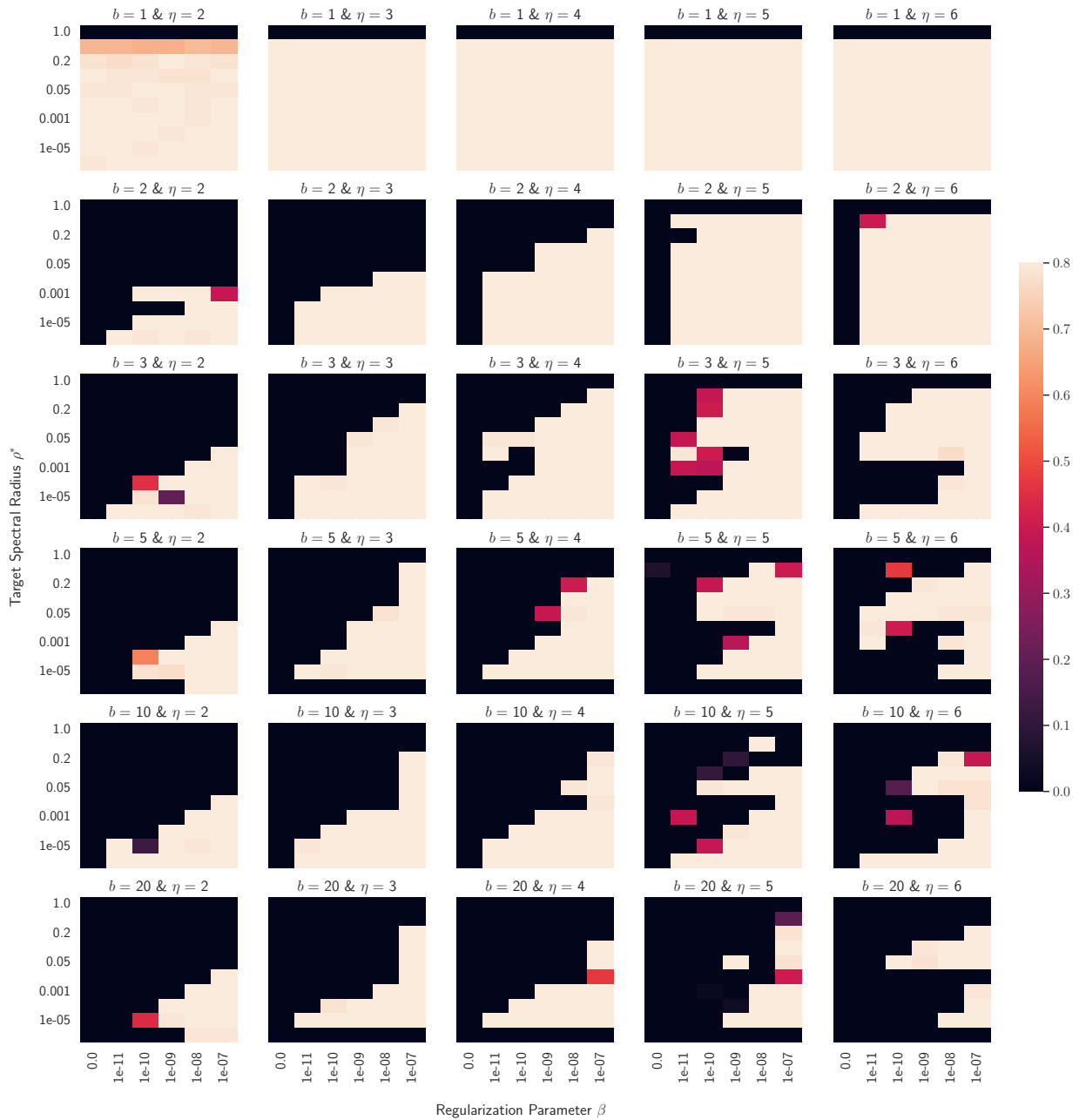


Figure 4.11: Largest Lyapunov Exponent. This Figure illustrates the largest Lyapunov for different parameterizations of minimal RC. Each heatmap varies the regularization parameter β along the x-axis and the target spectral radius ρ^* along the y-axis. β is varied between 0 and 1, while ρ^* is varied between 0 and $1e-7$. The arrangement of heatmaps forms a grid, where the block size b increases from top to bottom in each row, and the nonlinearity degree increases from left to right in each column. The block sizes vary between 1 and 20, while the nonlinearity degree varies between 1 and 6. Therefore, the size of the architecture increases from top left to bottom right. Each value displayed on the heatmap represents the mean of 100 attractor starting point deviations. A brighter color indicates a better performance.

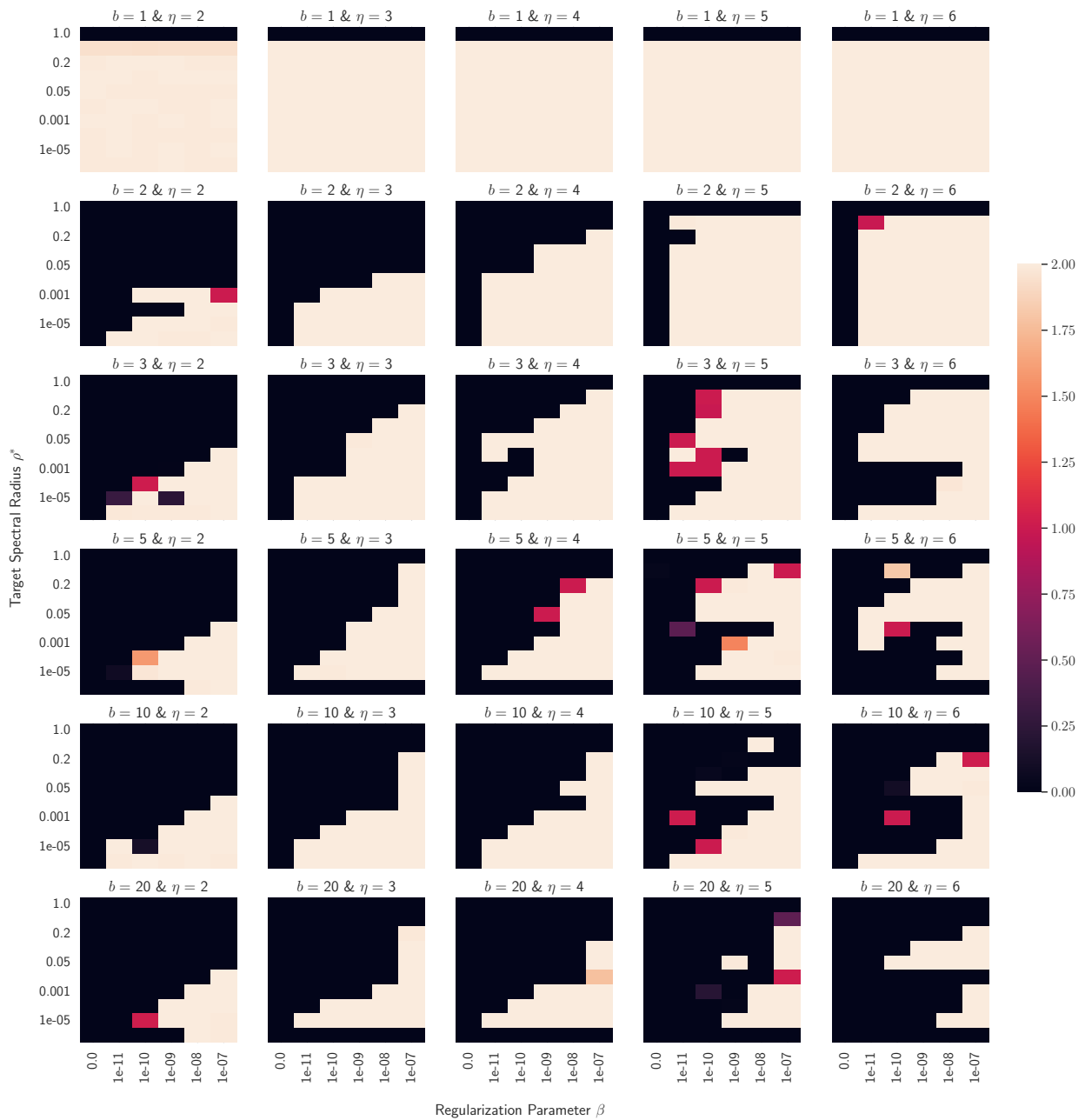


Figure 4.12: Correlation Dimension. This Figure illustrates the correlation dimension for different parameterizations of minimal RC. The configuration is analogous to Figure 4.11.

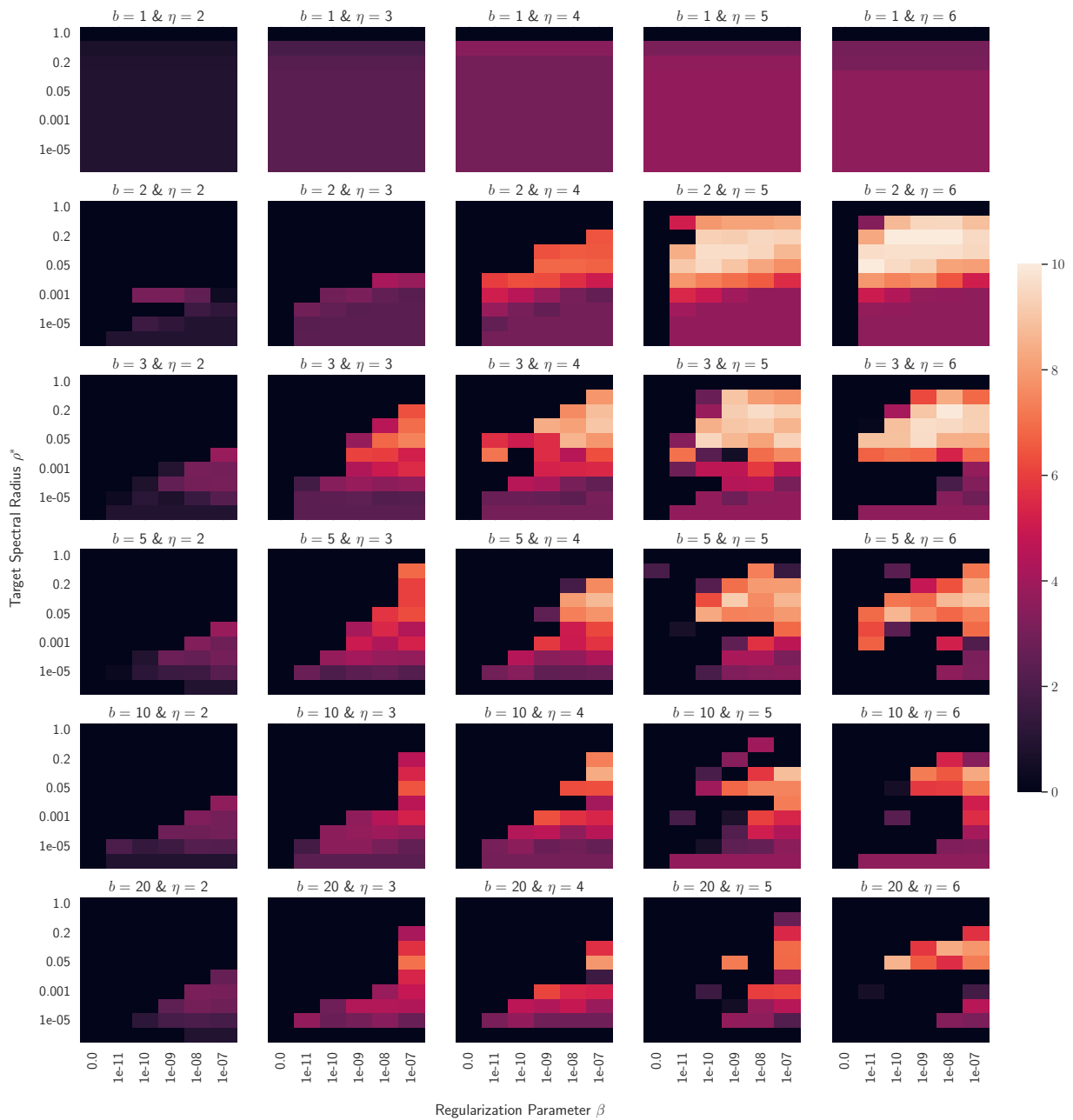


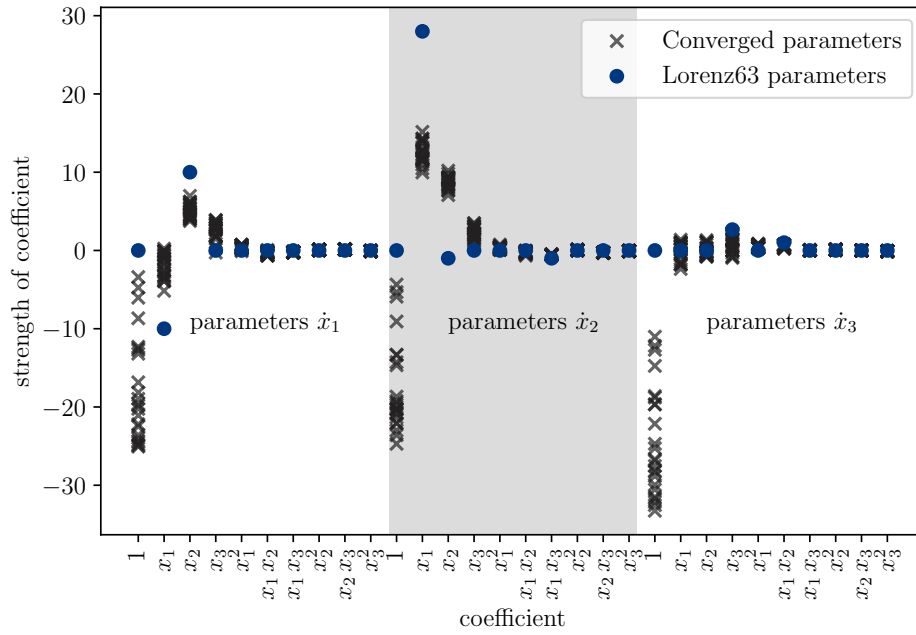
Figure 4.13: Forecast Horizon. This Figure illustrates the forecast horizon for different parameterizations of minimal RC. The configuration is analogous to Figure 4.11.

Summary and Outlook

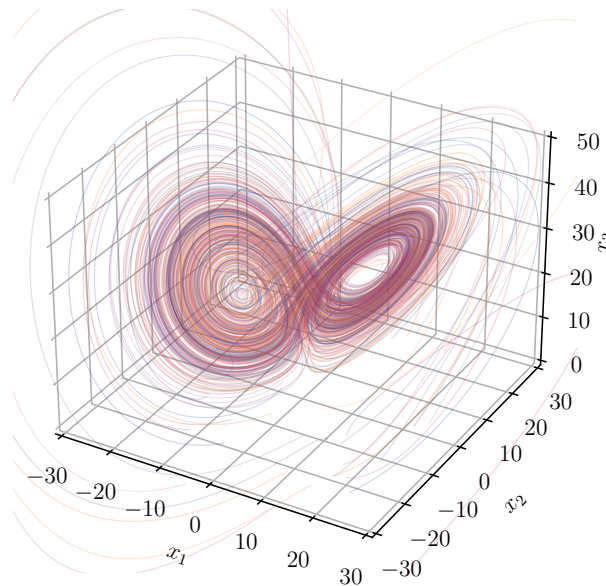
The inherent complexity of the world we inhabit presents a compelling challenge: to analyze and comprehend its intricate dynamics for better decision-making. This dissertation aimed to contribute to the body of research focused on demystifying these complex systems and enhancing our understanding and interaction with them. By addressing the nuances of nonlinearity in causal inference, deriving and calibrating governing equations using synchronization of chaos, and advancing machine learning techniques for the prediction of chaotic systems, we moved closer to making these complex systems more transparent and comprehensible. The methodologies and findings presented in this dissertation offer both theoretical advancements and practical applications. They are particularly relevant in areas where interpretability plays a crucial role in decision-making processes. In the following we present the main findings of our research and suggest potential avenues for future research.

How to be aware of nonlinearity. We began our journey by focusing on the often overlooked aspect of nonlinearity in causal inference. Although methods for inferring causality are constantly evolving, adequately addressing the identification of its drivers has remained a major obstacle. This problem is especially crucial when dealing with complex systems, where determining whether causality arises from linear or nonlinear elements proves extremely valuable. We validated our framework on synthetic chaotic systems and demonstrated the significance of nonlinear features in causality. We applied our findings to financial markets, where researchers and practitioners need to identify and measure interdependencies among financial instruments. However, conventional techniques like Pearson correlation have limited descriptive ability and only capture linear dependencies. We presented a comprehensive approach that incorporates both linear and nonlinear causalities. Our findings reveal significant nonlinear causality in stock indices in both Germany and the United States. While correlation may serve as a useful proxy for linear causality, it neglects to consider nonlinear factors, thus resulting in an underestimation of causality. Furthermore, this research highlights the potential use of causality in pair trading, portfolio optimization, and risk management. A natural extension is to incorporate further causal inference techniques into this framework and apply it to other real-world complex systems. As we have shown, it does not take much to enhance existing methodologies by simply being aware of nonlinearity.

How do machines learn chaos. After developing methods to better understand the linear and nonlinear causal links between state variables, we derived governing equations to describe the dynamics of the underlying system. Compared to other black-box approaches, our framework offers transparency by directly translating causal structures into differential equation terms. To refine the equation parameters, we utilize the synchronization of chaos. By coupling data to equations, we transform the inherently difficult chaotic optimization problem into convex one. This enables us to incorporate gradient descent, a widely used technique in machine learning, to identify the equation parameters that best reflect the data. While this effect was discovered empirically, the next step is to mathematically prove why the coupling results in convex optimization. Furthermore, our recent research revealed an interesting behavior: when we apply the algorithm to equations containing all possible combinations of variables, it identifies alternative equations to the Lorenz system that also yield the iconic butterfly attractor. Hence, one question that arises is how machine learning algorithms actually learn chaos. This question also arises in the next part of our journey, where we delve into predicting chaotic systems using machine learning.



a Parameters for second order library



b Trajectories

Figure 4.14: Learning Chaos. This Figure displays the results of fitting multiple samples of Lorenz data to an equation containing all possible combinations of state variables up to second order. The coefficients in Figure (a) differ from the *correct* Lorenz system equations — nonetheless, as presented in Figure (b), the integrated trajectories lead to a comparable butterfly shape. Adapted from Prosperino, Ma, & R ath [138].

How can machines help us. Acknowledging that not all complex systems can be accurately captured by governing equations, we explored the use of machine learning techniques, specifically reservoir computing. While providing data-driven advantages and a relatively simple architecture, challenges remain in ensuring its interpretability and efficient hardware implementations. Therefore, we proposed modifications to the traditional reservoir computing architecture to eliminate the need for randomness and render its components more interpretable. This novel approach of minimal reservoir computing consistently enhances predictive accuracy over long and short time scales, while significantly reducing computational requirements. This improvement is evident even when using very small training datasets. Moving forward, we seek to take a step toward incorporating reservoir computing to address real-world problems. Our latest research shows that an exciting application of reservoir computing is the control of nonlinear dynamics, which is useful in various stabilization tasks ranging from rocket launches to pacemakers and power grids. Initial results have already shown that our model performs exceptionally well in the example of forcing a chaotic parameterization of the Lorenz system into intermittent dynamics.

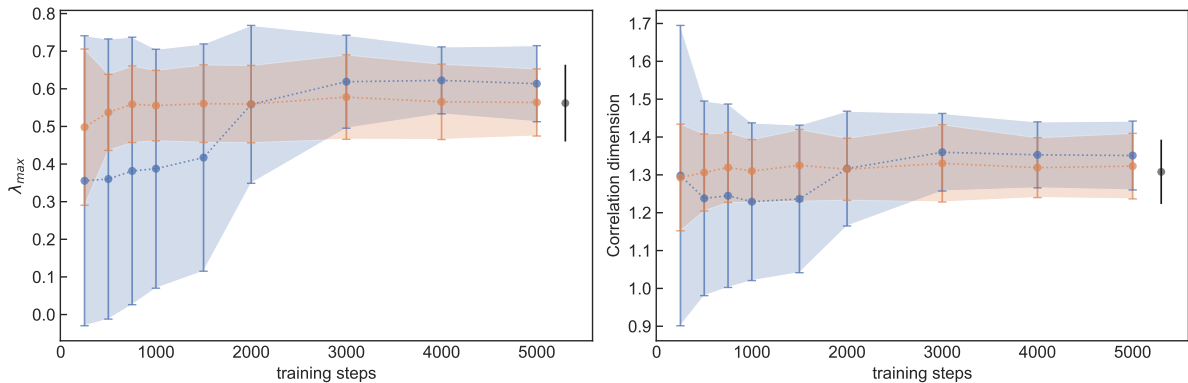


Figure 4.15: Controlling Chaos. This Figure illustrates the example of forcing a chaotic parameterization of the Lorenz system into intermittent dynamics using traditional (blue) and minimal (red) reservoir computing. It is demonstrated that minimal reservoir computing excels at this task and requires significantly less training data. Specifically, the Figure displays the average and standard deviation of the largest Lyapunov exponent (left) and correlation dimension (right) of the controlled system, based on various experiments with differing amounts of training data. The gray reference bar indicates the corresponding values for the simulated system in its original state. Adapted with updated results from Halaszczynski, Koeglmayr, & R ath [139].

While the journey of this dissertation ends here, as is the nature of research, many avenues for further exploration are now open. The reader may wonder what the author of this dissertation will do next. Well, after three years, he is looking forward to read for fun again. The reader is most warmly invited to send in recommendations. However, please do not suggest A Brief History of Time, the Feynman Lectures, or any other physics-related literature, as the author is already required to read them in preparation for his doctoral examination and are therefore not considered as fun.

Wish him luck

Synthetic Systems

In this Appendix, we present the governing equations and default parameterizations for all synthetic systems examined in this dissertation [140]. The differential equations are solved using the fourth-order Runge-Kutta method [18].

Lorenz

The Lorenz system was originally developed to model atmospheric convection [7]. Therefore, the state variables represent the convective flow, temperature variation, and vertical temperature variation:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - z) - y \\ \dot{z} &= xy - \beta z.\end{aligned}\tag{50}$$

The adjusted equations controlling for nonlinearity are:

$$\begin{aligned}\dot{x} &= \sigma(y - x) \\ \dot{y} &= x(\rho - \lambda_1 z) - y \\ \dot{z} &= \lambda_2 xy - \beta z.\end{aligned}\tag{51}$$

Parameter values	Initial point $\mathbf{u}(0)$	Time step dt	Lyapunov λ_{max}
$\sigma = 10, \rho = 28, \beta = 8/3$	$[-14, -20, 25]$	0.025	1.989

Halvorsen

While the Lorenz system contains mixed nonlinearity terms, the Halvorsen system has quadratic nonlinearities [140]. Furthermore, the equations have a cyclic symmetry:

$$\begin{aligned}\dot{x} &= -ax - 4y - 4z - y^2 \\ \dot{y} &= -ay - 4z - 4x - z^2 \\ \dot{z} &= -az - 4x - 4y - x^2.\end{aligned}\tag{52}$$

The adjusted equations controlling for nonlinearity are:

$$\begin{aligned}\dot{x} &= -ax - 4y - 4z - \lambda y^2 \\ \dot{y} &= -ay - 4z - 4x - \lambda z^2 \\ \dot{z} &= -az - 4x - 4y - \lambda x^2.\end{aligned}\tag{53}$$

Parameter values	Initial point $\mathbf{u}(0)$	Time step dt	Lyapunov λ_{max}
$a = 1.3$	$[-6.4, 0, 0]$	0.025	0.790

Fully Linear

To confirm that a system that is entirely linear only results in the identification of linear causality by the framework introduced in Chapter 2, we introduce the following *dummy* system:

$$\begin{aligned} \dot{x} &= \sin(y) \\ \dot{y} &= x + z \\ \dot{z} &= x - y. \end{aligned} \tag{54}$$

Initial point $\mathbf{u}(0)$	Time step dt
[1, 1, 1]	0.01

Thoai Fully Nonlinear

To ensure that the framework we introduced in Chapter 2 only detects nonlinear causality in a fully nonlinear system, we analyze the following system [141]:

$$\begin{aligned} \dot{x} &= \alpha y z \\ \dot{y} &= 1 - z^2 \\ \dot{z} &= \beta x^3 + y z. \end{aligned} \tag{55}$$

Parameter values	Initial point $\mathbf{u}(0)$	Time step dt
$\alpha = \beta = 1$	[1, 1, 1]	0.01

Coupled Difference

A simple example of a system that exhibits chaotic behavior is the coupled difference [52], which was also used by Sugihara, May, Ye, et al. [15] to illustrate *Convergent Cross Mapping*:

$$\begin{aligned} x(t+1) &= x(t) \cdot [r_x - r_x x(t) - \beta_{y \rightarrow x} y(t)] \\ y(t+1) &= y(t) \cdot [r_y - r_y x(t) - \beta_{x \rightarrow y} x(t)]. \end{aligned} \tag{56}$$

Parameter values	Initial point $\mathbf{u}(0)$
$r_x = 3.8, r_y = 3.5, \beta_{y \rightarrow x} = 0.02, \beta_{x \rightarrow y} = 0.1$	[0.2, 0.4]

Identifying causality drivers and deriving governing equations of nonlinear complex systems

Cite as: Chaos **32**, 103128 (2022); <https://doi.org/10.1063/5.0102250>

Submitted: 08 June 2022 • Accepted: 01 October 2022 • Published Online: 31 October 2022

Haochun Ma,  Alexander Haluszczynski, Davide Prosperino, et al.



View Online



Export Citation



CrossMark

ARTICLES YOU MAY BE INTERESTED IN

[Templex: A bridge between homologies and templates for chaotic attractors](#)

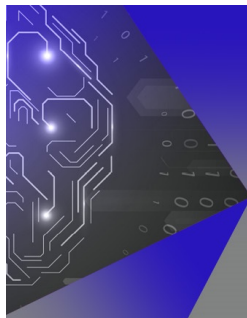
Chaos: An Interdisciplinary Journal of Nonlinear Science **32**, 083108 (2022); <https://doi.org/10.1063/5.0092933>

[Time-series forecasting using manifold learning, radial basis function interpolation, and geometric harmonics](#)

Chaos: An Interdisciplinary Journal of Nonlinear Science **32**, 083113 (2022); <https://doi.org/10.1063/5.0094887>

[Review of sample-based methods used in an analysis of multistable dynamical systems](#)

Chaos: An Interdisciplinary Journal of Nonlinear Science **32**, 082101 (2022); <https://doi.org/10.1063/5.0088379>



APL Machine Learning

Machine Learning for Applied Physics
Applied Physics for Machine Learning

Now Open for Submissions

Identifying causality drivers and deriving governing equations of nonlinear complex systems

Cite as: Chaos **32**, 1 031 28 (2022); doi: 10.1063/5.0102250

Submitted: 8 June 2022 · Accepted: 1 October 2022 ·

Published Online: 31 October 2022



View Online



Export Citation



CrossMark

Haochun Ma,^{1,2} Alexander Haluszczynski,²  Davide Prosperino,² and Christoph R ath^{3,a)} 

AFFILIATIONS

¹Ludwig-Maximilians-Universit at M unchen, Department of Physics, Schellingstra e 4, 80799 Munich, Germany

²Allianz Global Investors, risklab, Seidlstra e 24, 80335 Munich, Germany

³Deutsches Zentrum f ur Luft- und Raumfahrt (DLR), Institut f ur KI Sicherheit, Wilhelm-Runge-Stra e 10, 89081 Ulm, Germany

^{a)} Author to whom correspondence should be addressed: christoph.raeth@dlr.de

ABSTRACT

Identifying and describing the dynamics of complex systems is a central challenge in various areas of science, such as physics, finance, or climatology. While machine learning algorithms are increasingly overtaking traditional approaches, their inner workings and, thus, the drivers of causality remain elusive. In this paper, we analyze the causal structure of chaotic systems using Fourier transform surrogates and three different inference techniques: While we confirm that Granger causality is exclusively able to detect linear causality, transfer entropy and convergent cross-mapping indicate that causality is determined to a significant extent by nonlinear properties. For the Lorenz and Halvorsen systems, we find that their contribution is independent of the strength of the nonlinear coupling. Furthermore, we show that a simple rationale and calibration algorithm are sufficient to extract the governing equations directly from the causal structure of the data. Finally, we illustrate the applicability of the framework to real-world dynamical systems using financial data before and after the COVID-19 outbreak. It turns out that the pandemic triggered a fundamental rupture in the world economy, which is reflected in the causal structure and the resulting equations.

  2022 Author(s). All article content, except where otherwise noted, is licensed under a Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>). <https://doi.org/10.1063/5.0102250>

Understanding cause–effect relationships is a key challenge in many areas of science, as it forms the basis for developing analytical and predictive models. However, while methods for causal inference are constantly evolving, finding the drivers of causality is a critical aspect that is often not adequately addressed. Particularly when analyzing complex nonlinear systems, it is very useful to know whether causality stems from linear or nonlinear properties. In this work, we separate the causality in linear and nonlinear contributions and observe that a significant part can be attributed to nonlinear properties. Furthermore, we present a framework by which knowledge of the causal structure can be directly translated into equations that describe the underlying data. Potentially, this methodology can be used to find equations for real systems that allow for precise analysis and prediction.

I. INTRODUCTION

Causality, as one of the basic principles of scientific thought, has been intensively researched over many generations and different disciplines. Throughout history, interpretations of causality have evolved with the increasing effort and complexity of physical theories. While in Newton’s classical understanding action and reaction were defined as simultaneously coupled, Einstein introduced a temporal and spatial component by defining causality as events connected by the cone of light.¹ Subsequently, the disruption of quantum mechanics led to a probability-dominated understanding of physics, where causality is an unimaginable concept in a non-deterministic world. With the advent of chaos theory, causality was placed in the context of stability and equilibria of dynamical systems, which became known to the general public as the butterfly effect.²

Encouraged by the explosion of computation resources, the development of causal inference methods took a similar but accelerated path. Beginning with Granger causality in the 1960s,³ many techniques of increasing complexity were developed, ranging from information-theoretic measures⁴ to state-space reconstruction methods;⁵ Runge⁶ provides an excellent overview.

However, while causal inference is primarily concerned with measuring the presence of causality, research on its properties and drivers has remained secondary. A first step in this direction was taken by Paluš *et al.*⁷ who developed a diagnostic test for identifying nonlinear dynamic relationships in time series based on mutual information. Another approach, using Fourier transform surrogates, was taken by Haluszczyński *et al.*,⁸ who separated linear and nonlinear contributions of mutual information to capture nonlinear correlations in financial data. The contribution of nonlinearity to connectivity in climate data was quantified by Hlinka *et al.*⁹

While initial approaches for deriving governing equations from data in the 1990s were based on applying the flow method by inter alia Breden and Hübner¹⁰ and Eisenhammer *et al.*,¹¹ research on this topic has expanded considerably in the last few decades. In the context of nonlinear dynamical systems, Brunton *et al.*¹² introduced sparse identification on the chaotic Lorenz attractor. Other techniques include automated inference of dynamics¹³ and machine learning approaches.¹⁴

In this work, we combine the inference and analysis of causality with the derivation of governing equations in nonlinear complex systems. Therefore, we separate the linear and nonlinear contributions to causality using Fourier transform surrogates and develop a transparent rationale based only on the causalities to derive the differential equations.

II. BENCHMARK MODELS

In this work, we first validate our approach on four synthetic systems before demonstrating its applicability on a real-world example. If not stated otherwise, we solve the differential equations of the synthetic system using the Runge–Kutta method¹⁵ for

$T = 10\,000$ steps and a discretization of $dt = 0.01$. We discard the initial transient of $T = 50\,000$ steps for the analyses.

A. Lorenz system

In order to analyze the effect of nonlinearity on the causality structure, we introduce two additional parameters λ_1 and λ_2 to control the nonlinear terms of the Lorenz system, which models atmospheric convection,¹⁶

$$\begin{aligned} \frac{dx}{dt} &= \sigma(y - x), \\ \frac{dy}{dt} &= x(\rho - \lambda_1 z) - y, \\ \frac{dz}{dt} &= \lambda_2 xy - \beta z, \end{aligned} \tag{1}$$

where the standard parametrization is $\sigma = 10$, $\rho = 28$, $\beta = 8/3$, and $\lambda_1 = \lambda_2 = 1$. The implied linear and nonlinear connections between the variables are depicted in Fig. 1.

Figure 2 illustrates the attractor for a selection of different parameter configurations: While the system diverges for nonlinearity degrees less or equal to 0, the upper bounds can be chosen arbitrarily as we do not observe significant changes to the butterfly form even for extreme values ($\lambda_1, \lambda_2 \approx 1000$).

B. Halvorsen system

While the nonlinearity terms of the Lorenz system are mixed products of two different variables, the circulant Halvorsen system¹⁷ entails quadratic nonlinearities,

$$\begin{aligned} \frac{dx}{dt} &= ax - 4y - 4z - \lambda y^2, \\ \frac{dy}{dt} &= ay - 4z - 4x - \lambda z^2, \\ \frac{dz}{dt} &= az - 4x - 4y - \lambda x^2, \end{aligned} \tag{2}$$

where $a = 1.3$ and $\lambda = 1$ are the standard parameters.

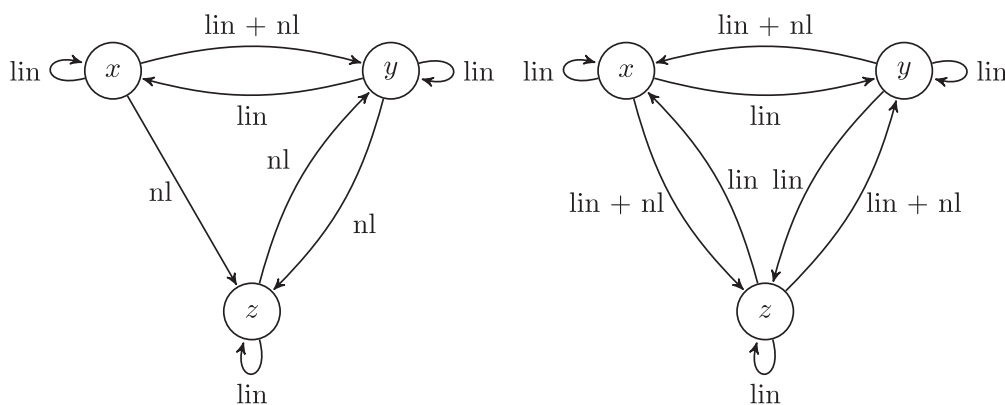


FIG. 1. Causality pictogram of the Lorenz (left) and Halvorsen (right) system. The linear (lin) and nonlinear (nl) causal links are depicted by the labeled arrows.

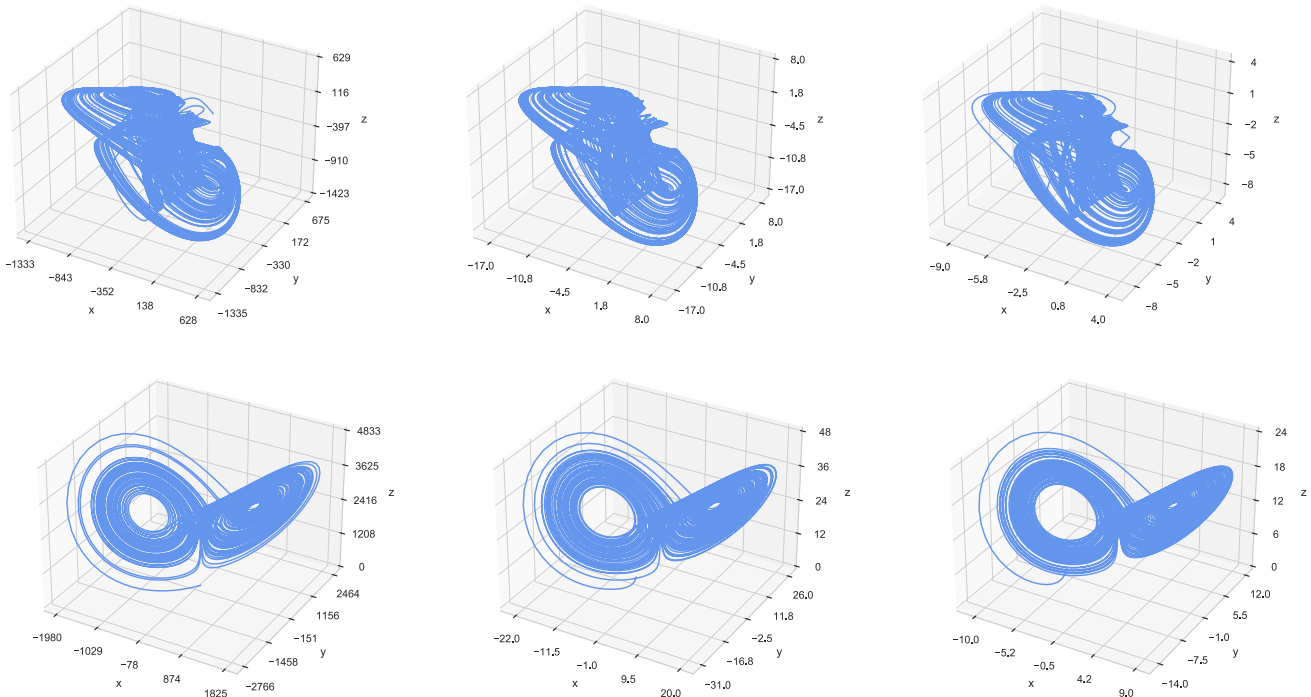


FIG. 2. Lorenz (top row) and Halvorsen (bottom row) attractors for different degrees of nonlinearity. While the standard parameters for the Lorenz system are set at $\sigma = 10$, $\beta = 8/3$, $\rho = 28$, the nonlinearity parameters from left to right are $\lambda_1 = \lambda_2 = 0.01$, $\lambda_1 = \lambda_2 = 1$, and $\lambda_1 = \lambda_2 = 2$. For the Halvorsen system, the nonlinearity parameters from left to right are $\lambda = 0.01, 1, 2$.

Analogously, we control the nonlinearity strength through the additional parameter λ . As observed for the Lorenz system, the basic form of the Halvorsen attractor also stays intact for variations in nonlinearity, as illustrated in Fig. 2.

C. Fully linear system

In order to verify that a fully linear system leads to only linear causality to be detected, we include the following system into our analysis:

$$\begin{aligned} \frac{dx}{dt} &= \sin(y), \\ \frac{dy}{dt} &= x + z, \\ \frac{dz}{dt} &= x - y. \end{aligned} \tag{3}$$

We would like to point out that purely linear systems do not exhibit chaotic behavior and that this system serves solely as a verification of our methods. The time series for the first $T = 30\,000$ steps after the initial transient are shown in Fig. 3.

D. Fully nonlinear system

In contrast, we also include a fully nonlinear system specified by the following equations:¹⁸

$$\begin{aligned} \frac{dx}{dt} &= \alpha yz, \\ \frac{dy}{dt} &= 1 - z^2, \\ \frac{dz}{dt} &= \beta x^3 + yz, \end{aligned} \tag{4}$$

where we set $\alpha = \beta = 1$ for chaotic behavior. The attractor of this system is depicted in Fig. 3.

E. Stock indices

In order to demonstrate the applicability of our framework to real-world systems, we consider the global financial market around the outbreak of the COVID-19 pandemic. Therefore, we choose the six major economies and their corresponding MSCI stock indices between November 2018 and May 2021: Europe (EU), United States (US), China (CN), Emerging Markets (EE), Japan (JP), and Pacific excluding Japan (PX). We convert the daily prices p_t to log-returns,

$$x_t \equiv \log p_t - \log p_{t-1}, \tag{5}$$

and divide the series into two phases: the time before the outbreak of the pandemic in February 2020 and the time after. This yields two sets of time series each with length $T = 325$, respectively. The

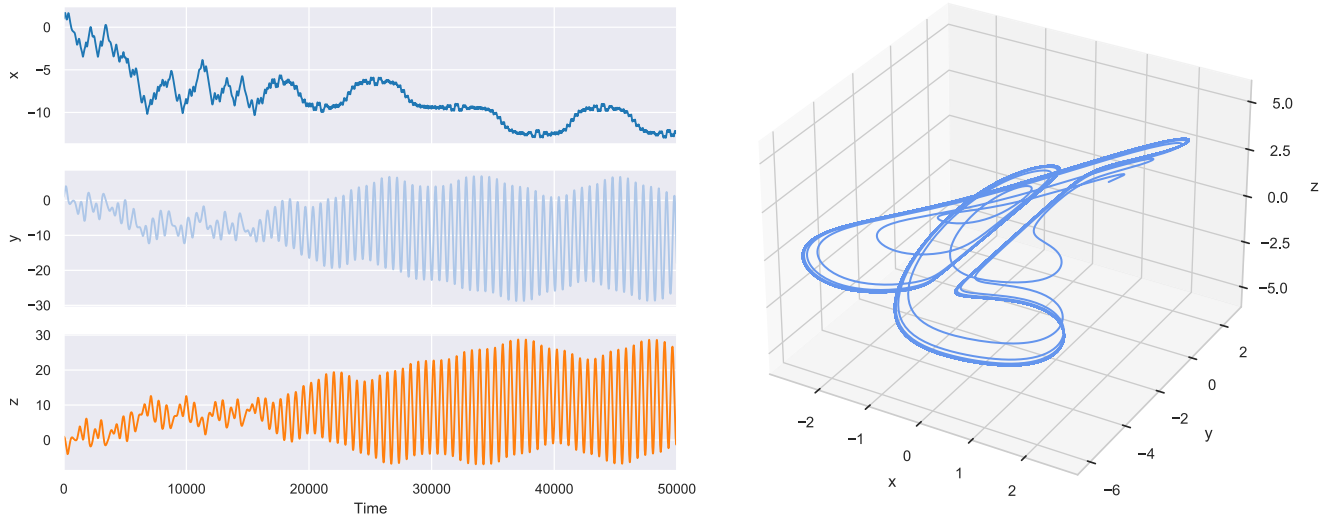


FIG. 3. Time series of the fully linear (left) and nonlinear (right) systems. The left figure depicts the first $T = 30\,000$ time series steps after the initial transient of the fully linear system, while the right figure shows the attractor of the fully nonlinear system.

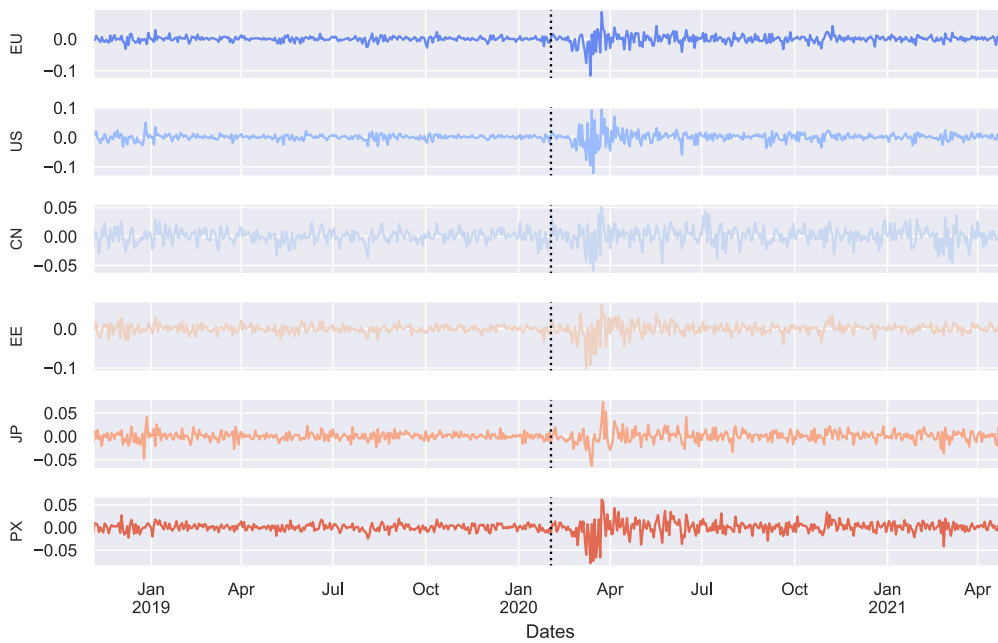


FIG. 4. Log returns of stock indices from six major economies. The illustrated time series are Europe (EU), United States (US), China (CN), Emerging Markets (EE), Japan (JP), and Pacific excluding Japan (PX). The black dashed vertical line depicts the outbreak of the COVID-19 pandemic in February 2020.

structural change in the dynamics of the stock indices triggered by the outbreak of COVID-19 can be observed in Fig. 4.

III. METHODS

In the following, we present the methods used in this work, which we assign to four different categories: Causality measures, Fourier transform surrogates, network measures, and the derivation of governing equations.

A. Causality measures

We select three techniques representing the main categories currently used in causal inference⁶—however, it is important to note that our framework is applicable to any method capable of detecting nonlinear causality.

1. Granger causality using linear autoregressive model

As one of the first causal inference approaches, Granger causality (GC) tests whether the prediction error of the next time step of a time series \mathbf{y} can be decreased by including the history of another time series \mathbf{x} —in this case, \mathbf{x} is said to Granger cause \mathbf{y} .¹⁹ Its original form compares the prediction error of a restricted autoregression model,

$$\hat{y}_t = \sum_{\tau=1}^{\tau_{max}} \alpha_{\tau} y_{t-\tau} + \varepsilon_t, \quad (6)$$

to its corresponding augmented model,

$$\hat{y}_t = \sum_{\tau=1}^{\tau_{max}} \alpha_{\tau} y_{t-\tau} + \sum_{\tau=1}^{\tau_{max}} \beta_{\tau} x_{t-\tau} + \eta_t, \quad (7)$$

where α_{τ} and β_{τ} are coefficients at lag τ and ε_t and η_t denote independent error terms. While GC is mostly used as a binary statistical hypothesis test,¹⁹ we quantify the strength of the causal coupling using the following normalization:

$$\psi_{GC}(\mathbf{x}, \mathbf{y}) = 1 - \min \left\{ 1, \left(\frac{\text{RSS}_{aug}}{\text{RSS}_{rest}} \right)^2 \right\}, \quad (8)$$

where RSS_{rest} and RSS_{aug} denote the residual sum of squares (RSS) of the restricted and augmented model, respectively. Hence, when the regression of the augmented model performs better than the restricted model ($\text{RSS}_{aug} < \text{RSS}_{rest}$), the fraction on the right-hand side is small—this implies stronger causality.

Since there exists no universal method to determine the optimal maximum lag τ_{max} , we repeat the procedure for several values of the maximum lag τ_{max} and average the result. Therefore, we take $N = 20$ equally distributed values between 1 and the time series length T : $\tau_{max} = 1, T/N, 2T/N, \dots, T$. As we do not find a significant difference, we conclude that the average is a good estimator within the scope of this work.

2. Transfer entropy

Following the proof of equivalence between GC and transfer entropy (TE) for Gaussian variables,²⁰ the measure introduced by

Schreiber⁴ has been widely regarded as the information-theoretical extension of GC. Analogously, TE quantifies the reduction of uncertainty on future values of \mathbf{y} by accounting for past values of \mathbf{x} given the history of \mathbf{y} . In essence, it is a special case of conditional mutual information (CMI),

$$\begin{aligned} \psi_{TE}(\mathbf{x}, \mathbf{y}) &\equiv I(\mathbf{y}; \mathbf{x}_{t-1:} | \mathbf{y}_{t-1:}) \\ &= H(\mathbf{y}, \mathbf{y}_{t-1:}) + H(\mathbf{x}_{t-1:}, \mathbf{y}_{t-1:}) \\ &\quad - H(\mathbf{x}_{t-1:}, \mathbf{y}, \mathbf{y}_{t-1:}) - H(\mathbf{y}_{t-1:}), \end{aligned}$$

where the colon indicates all previous steps of the time series and where H denotes the (joint) entropy of the time series calculated via

$$H(\mathbf{x}, \mathbf{y}) = - \sum_{t=1}^T \sum_{i=1}^T p(x_t, y_i) \log p(x_t, y_i). \quad (9)$$

For better comparability to other inference methods, we propose the following normalization:

$$\psi_{TE}(\mathbf{x}, \mathbf{y}) \mapsto \frac{\psi_{TE}(\mathbf{x}, \mathbf{y})}{\sqrt{H(\mathbf{y}, \mathbf{y}_{t-1:}) H(\mathbf{x}, \mathbf{x}_{t-1:})}}. \quad (10)$$

Our reasoning for this normalization is based on our interpretation of TE as an asymmetric causal measure, similar to covariance, which is rescaled to obtain the normalized cross correlation.

We would like to point out that the calculation of empirical probability densities p and, hence, information-theoretic measures raise unexpected difficulties exceeding the scope of this work. While it is common to use histograms with equally distributed bins to estimate densities, Mynter²¹ showed that this method potentially leads to biases since the estimation is highly dependent on the partition details—hence, finding a robust estimator is non-trivial. However, for the purpose of our research, we find that for time series of length T a number of $\sqrt{T/4}$ equally distributed bins performs reasonably well. This was also empirically confirmed by Baur and R ath²² who used this binning configuration for the construction of generalized local states in reservoir computing. Furthermore, it is worth mentioning that TE might capture false causalities depending on the dimension of conditioning.²³

3. Convergent cross mapping

Another category of causal inference is state-space methods such as convergent cross-mapping (CCM), which was developed by Sugihara *et al.*⁵ Its underlying idea is based on Takens' theorem, which states that the full state-space can be reconstructed from a single embedded coordinate of the system, also called shadow manifold. Due to transitivity, two coordinates within one system can then be mapped to each other through neighboring states in their respective shadow manifolds—this enables a cross prediction. Hence, if \mathbf{x} causes \mathbf{y} , the prediction of the future \hat{y}_t using the shadow manifold $\mathcal{M}_{\mathbf{x}}$ should be identical to the actual value y_t . In the context of CCM, the prediction is extended from a single value to a series. Therefore, both time series are divided into training and test sets, where the former are used to construct the shadow manifolds and the latter serve as benchmarks to evaluate the prediction performance.

While CCM is defined as the Pearson correlation ρ between the prediction $\hat{y}|\mathcal{M}_x$ and the test set of y , we propose another evaluation measure, the correlation distance $d = \sqrt{2(1-\rho)}$, in order to rescale the correlation to a positive interval. This entire procedure is repeated for an increasing training set fraction, which delivers a series \mathbf{d} consisting of D correlation distances. This series should theoretically converge to a maximum since the prediction is enhanced for finer resolutions of the shadow manifolds.

While originally CCM requires visual judgment of the convergence, we introduce an algorithmic approach using overlapping sliding windows of size \mathbf{d} . For each window, we calculate the standard deviation and set a threshold of 0.1. The convergence is fulfilled if the standard deviation decreases continuously and falls below the preset threshold. If \mathbf{d} converges, we calculate the mean of its last five values in order to smooth outliers. In case of non-convergence, we set the CCM causality to 0,

$$\psi_{CCM}(\mathbf{x}, \mathbf{y}) \equiv \begin{cases} \frac{1}{5} \sum_{i=1}^5 d_{D-5+i} & \text{if } \mathbf{d} \text{ converges} \\ 0 & \text{else} \end{cases} \in [0, 1]. \quad (11)$$

We would like to point out that there exist reservations toward CCM regarding some synthetically created systems—however, its wide range of successful applications is testament to its importance for causal inference.²⁴ We determined the optimal lag by finding the first minimum of the lagged mutual information—this yielded a lag $\tau = 1$. The optimal embedding dimension $\kappa = 3$ was found by using the false-nearest-neighbor algorithm.²⁵

4. Limits of causality measures

We would like to point out that we are aware of the limitations of the causal inference methods presented and of causal inference in general. However, in this paper, we use them only to illustrate a framework of how causality can be partitioned into linear and nonlinear contributions and how, assuming correct measurements, governing equations can be derived. It is beyond the scope of this paper to analyze whether they measure true causality and how robust the methods are. For a more detailed discussion of these points, we refer to their original papers Granger,³ Schreiber,⁴ and Sugihara *et al.*⁵

With respect to GC, we recognize that its main requirement, separability of variables, poses problems, especially when applied to deterministic dynamical systems.¹⁹ Therefore, GC only serves as a verification for our analysis, since it is based on autoregression and should, therefore, only capture causality arising from linear properties. We refer to Ref. 3 for more details.

Furthermore, we are aware that TE and CCM work with reconstructed spaces and that their application to variables within an attractor has theoretical weaknesses. However, the analysis in this paper is performed on simulated data and not on a theoretical basis. We refer the reader to Cummins *et al.*²⁶ for a detailed discussion of the effectiveness of state-space reconstruction methods in determining causality.

Lastly, we would like to note that we are aware that real-world system can be contaminated by different kinds of noise, which affects the performance of our methods. However, these issues lie beyond the scope of this work since they are addressed in the papers

which describe the causality inference methods. Since the methods work when the causality graphs are correct, their robustness to noise lies entirely in the robustness of the individual inference models against noise. We refer to Overbey and Todd²⁷ and Krishna and Tangirala²⁸ for analyses on TE and CCM, respectively. Empirically, we find our method to be robust to white noise for Signal-to-Noise ratios (SNRs) > 50 dB.

B. Fourier transform surrogates

In order to dissect the causality structure of time series systems into contributions from linear and nonlinear contribution drivers, we utilize Fourier transform (FT) surrogates. They destroy the nonlinear characteristics of a time series \mathbf{x} while keeping the linear ones unaffected.²⁹

1. Algorithm

First, we perform a Fourier transformation to separate the linear properties into the amplitudes and the nonlinear ones into the phases. Through randomizing the phases of its Fourier transformation by adding uniformly distributed numbers ϕ_k , solely the nonlinear features are destroyed. Hence, the back-transformation only contains the linear properties of the time series,

$$\tilde{\mathbf{x}}^{(k)} = \mathcal{F}^{-1} \{ \mathcal{F} \{ \mathbf{x} \} e^{i\phi_k} \}. \quad (12)$$

We increase the robustness of our results by averaging measures that are calculated on surrogate time series, over multiple realizations of random phases. Unless otherwise specified, we repeat our measurements for $K = 10$ realizations. A discussion on surrogate generation is provided by R ath *et al.*³⁰

2. Surrogate-based measures

Within the scope of this work we understand a bivariate measure $\psi(\mathbf{x}, \mathbf{y})$ as a function that maps two time series to a real number. Hence, we define its corresponding surrogate measure as the average over K surrogate realizations of both time series,

$$\psi^{surro}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\tilde{\mathbf{x}}^{(k)}, \tilde{\mathbf{y}}^{(k)}). \quad (13)$$

As indicated by the superscript k , we add the same random phases to both time series within one realization. This leaves the phase differences unaffected, which, for example, preserves the Pearson correlation.³¹

Furthermore, we define the cross-measure by only surrogating the first time series in the argument,

$$\psi^{cross}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\tilde{\mathbf{x}}^{(k)}, \mathbf{y}), \quad (14)$$

and analogously define the reverse as the anti-measure,

$$\psi^{anti}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\mathbf{x}, \tilde{\mathbf{y}}^{(k)}). \quad (15)$$

The intuition behind the cross- and anti-measure is to analyze the influence of the linear part of x on y under the measure ψ and vice versa.

3. Nonlinear measures

In the next step, we use these alterations to construct non-linearity measures extending the idea of nonlinear correlation.⁸ Therefore, we assume every measure to be composed of a linear part, represented by the surrogate, and a remaining nonlinear part.

Hence, the most intuitive form is given by the difference,

$$\psi^{nl} \equiv \psi - \psi^{surro}. \tag{16}$$

As we rule out negative nonlinearities attributing them to spurious effects, we propose the measure

$$\psi^{nl} \equiv \max \{0, \psi - \psi^{surro}\}. \tag{17}$$

Further nonlinearity measures can be easily derived by, e.g., normalization or interchanging surro-, cross-, and anti-measures.

C. Evaluation of causality matrices

Given an N -dimensional time series $\mathcal{S} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, we can hence compute the causality matrix corresponding to an arbitrary

measure $\psi(\mathbf{x}, \mathbf{y})$,

$$\Psi(\mathcal{S}) \equiv \begin{pmatrix} \psi(\mathbf{x}_1, \mathbf{x}_1) & \dots & \psi(\mathbf{x}_1, \mathbf{x}_N) \\ \psi(\mathbf{x}_2, \mathbf{x}_1) & \dots & \psi(\mathbf{x}_2, \mathbf{x}_N) \\ \vdots & \ddots & \vdots \\ \psi(\mathbf{x}_N, \mathbf{x}_1) & \dots & \psi(\mathbf{x}_N, \mathbf{x}_N) \end{pmatrix},$$

which fully describes the explicit links between the individual variables. In the case of causality measures, they are similar to an adjacency matrix representing finite graphs—hence, the entries Ψ_{ij} quantify the causal flow from x_i to x_j .

Especially for high-dimensional systems, it is useful to directly evaluate the measure of the whole system. Therefore, we develop intuitive matrix measures, which map a matrix Ψ to a real number. As indicated, a possibility could be to construct a graph from the measure matrix and to compute its corresponding properties. However, as causality measures do not necessarily fulfill the conditions of mathematical distances, we propose the matrix mean

$$\mu_{mean}(\Psi) \equiv \frac{1}{N^2 - N} \sum_{i=1}^N \sum_{j=1}^N (1 - \delta_{ij}) |\Psi_{ij}|, \tag{18}$$

where we use the Kronecker delta δ_{ij} to dismiss the diagonal entries of the matrix since $\psi(\mathbf{x}, \mathbf{x})$ is equivalent for arbitrary time series \mathbf{x} .

Considering our focus on causality measures, causality should only be present in a system if no impasse exists which breaks the causal chain. Therefore, we use the geometric mean as it only returns

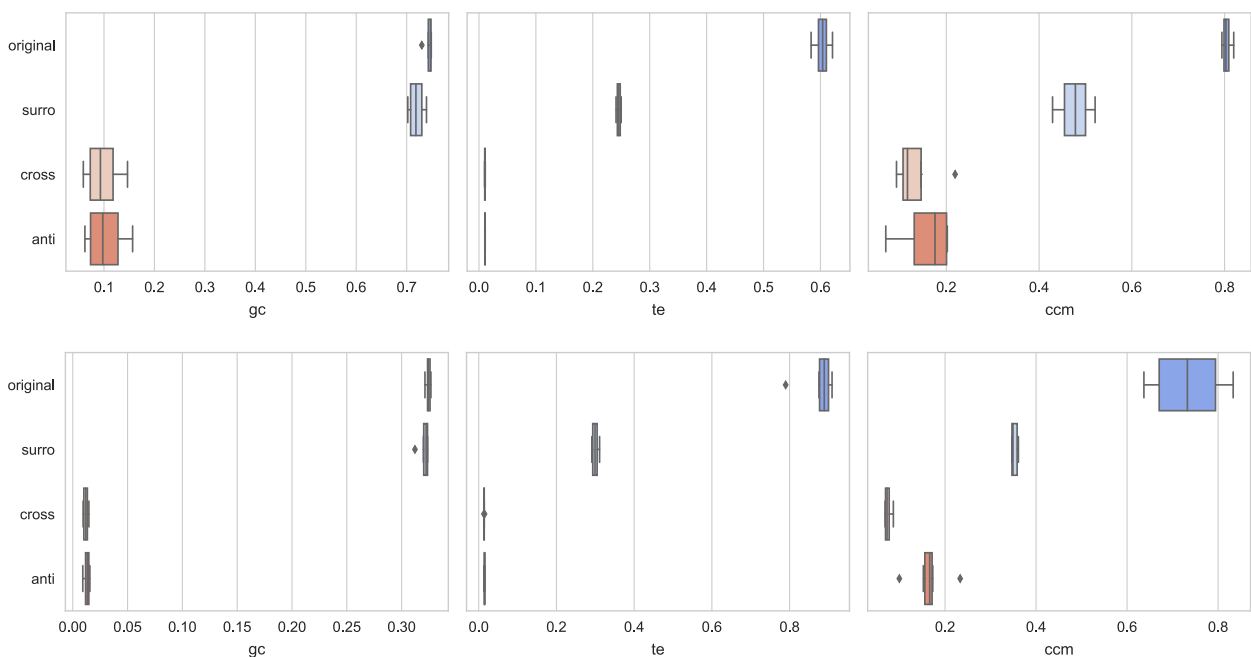


FIG. 5. Causality box plots of the standard Lorenz (top row) and Halvorsen (bottom row) systems. We compute the mean of the original-, surro-, cross-, and anti-matrices for GC, TE, and CCM, respectively. The sample consists of 50 simulations under the standard configuration. The surrogate-causalities are averaged over $K = 10$ surrogate realizations. The Lozenge symbols indicate outliers according to the interquartile range (IQR).³²

a nonzero value if all entries are nonzero,

$$\mu_{geo}(\Psi) \equiv \left(\prod_{i=1}^N \prod_{j=1}^N (1 - \delta_{ij} |\Psi_{ij}| + \delta_{ij}) \right)^{\frac{1}{N^2 - N}}. \quad (19)$$

Note that we include the matrix diagonals for cross- and anti-measures since their entries offer insights into the linear structure of the individual time series.

D. Derivation of governing equations

While extracting governing equations from data is key to build models in diverse fields of science,¹² existing methods often require sophisticated and specifically tailored algorithms. The major difficulty stems from the problem that there is an infinite number of possible governing equations that represent a finite time series. Even though the number of possibilities reduces for increasing length, the individual terms stay unidentifiable.

Hereby, we illustrate a simple rationale to derive equations directly from the causality matrices inferred from the underlying time series data. Therefore, we assume that the time series stem from a deterministic dynamical system, where a finite sample suffices to identify its underlying causal structure. Hence, by separating linear and nonlinear causalities, the terms of the governing equations

become separately deducible. Thus, we argue that the causal structure can be fully described by a linear matrix differential equation and a nonlinear part,

$$\frac{dx}{dt} = \left(\frac{dx}{dt} \right)_{lin} + \left(\frac{dx}{dt} \right)_{nl} = \Psi^{lin} x + \Psi^{nl} \odot x^n,$$

where \odot denotes our rationale for deriving the nonlinear terms and the superscript n indicates an n -dimensional Cartesian product. For simplicity, we assume all nonlinearity terms to be of order $n = 2$. However, this can be easily extended which is primarily relevant for high-dimensional systems.

First, we extract the linear terms from the surrogate- and cross-matrices. While the cross-matrix represents the linear causal flow of a variable to itself, we can extract the flow of the other variables from the surrogate-matrix,

$$\Psi^{lin} = \delta_{ij} \Psi_{ij}^{cross} + (1 - \delta_{ij}) \Psi_{ij}^{surro}. \quad (20)$$

Since we discard entries smaller than a preset threshold $\theta = 0.1$ attributing them to inaccuracies of the causal inference, the individual equations are given by

$$\left(\frac{dx_j}{dt} \right)_{lin} = \sum_i^N \Theta(\Psi_{ij}^{lin} - \theta) x_i, \quad (21)$$

where Θ is the Heaviside-function.

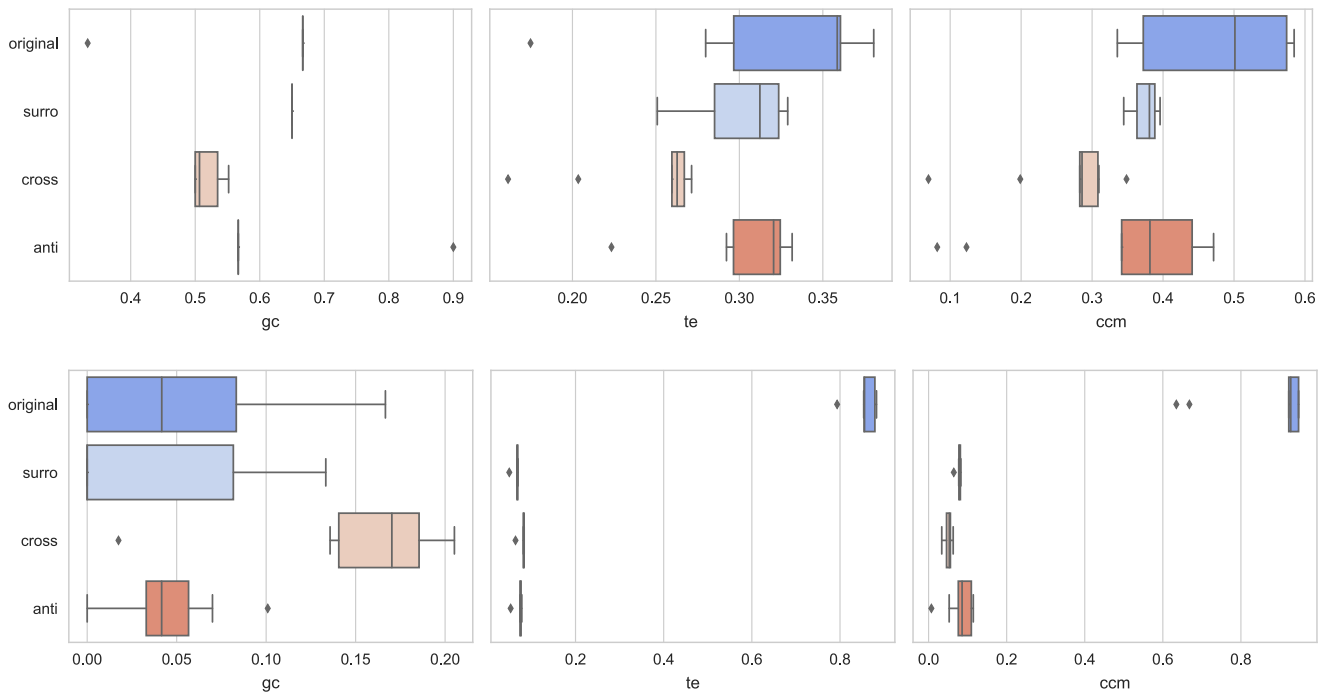


FIG. 6. Causality box plots of the fully linear (top row) and nonlinear (bottom row) systems. We compute the mean of the original-, surro-, cross-, and anti-matrices for GC, TE, and CCM, respectively. The sample consists of 50 simulations under the standard configuration. The surrogate-causalities are averaged over $K = 10$ surrogate realizations. The Lozenge symbols indicate outliers according to the interquartile range.

In the next step, we calculate the nonlinear causality-matrix Ψ^{nl} using the original- and surrogate-matrices. Since it incorporates inaccuracies stemming from two causal inferences, we raise the threshold to 2θ . The nonlinear part of the equations can then be constructed by adhering to two simple rules,

- If in one column x_j of Ψ^{nl} only one entry $x_i \neq x_j$ exceeds the threshold, then the nonlinear term entering the equation is

$$\left(\frac{dx_j}{dt}\right)_{nl} = \Theta(\Psi_{ij}^{nl} - 2\theta) x_i^2, \quad (22)$$

since we reason that the entire nonlinear causal flow of the system must be accumulated in x_i .

- If multiple entries $\{x_k, x_{k+1}, \dots, x_l\}$ in Ψ^{nl} exceed the threshold, then all permutation of pairs enter the equation

$$\left(\frac{dx_j}{dt}\right)_{nl} = \sum_{i=k}^n \sum_{j \leq i}^l \Theta(\Psi_{ij}^{nl} + \Psi_{ji}^{nl} - 4\theta) x_i x_j,$$

since we argue that the nonlinear causal flow must be split between all possible pairs.

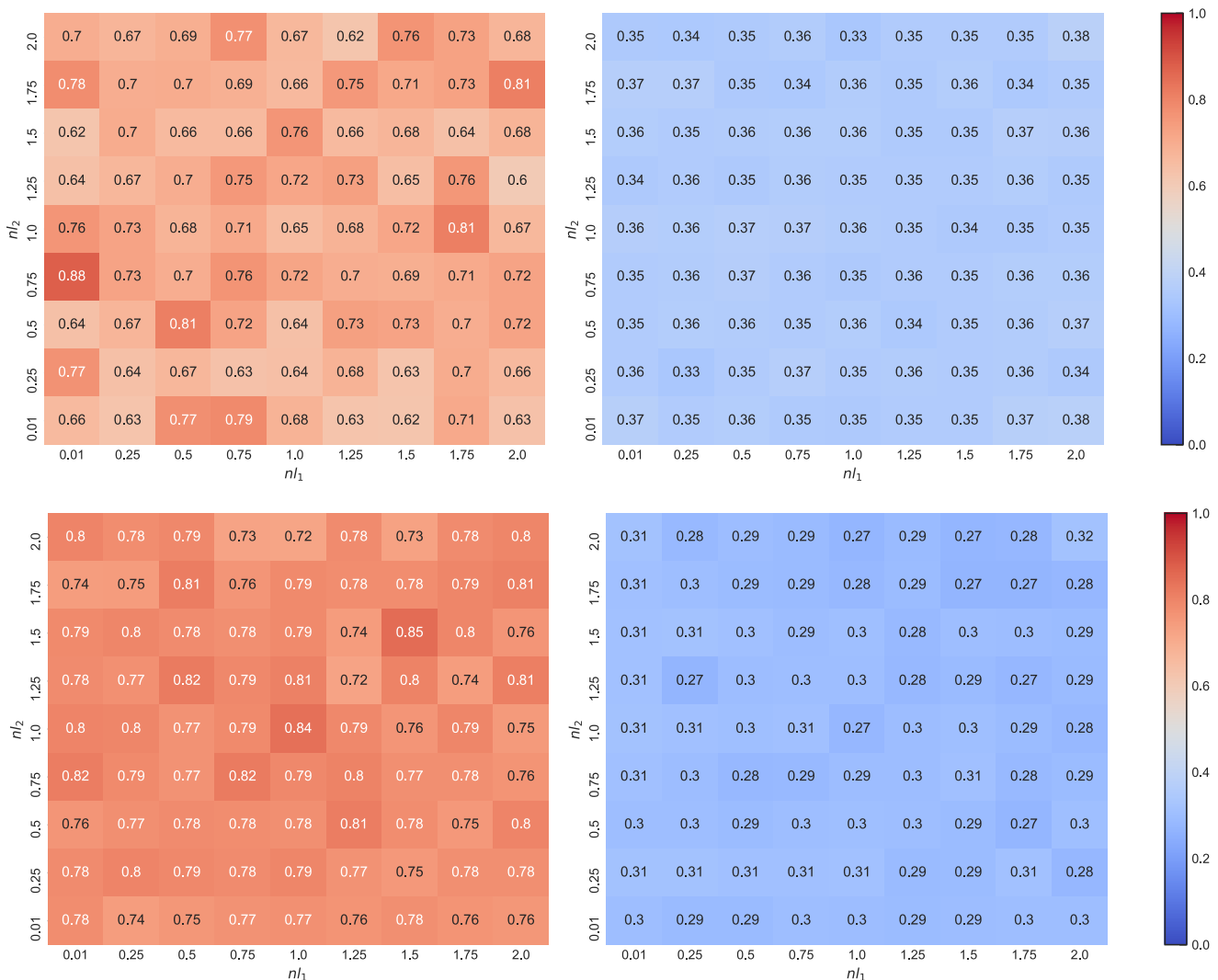


FIG. 7. TE (top row) and CCM (bottom row) causality of the Lorenz attractor for different degrees of nonlinearity. We compute the causalities for variations of λ_1 and λ_2 between 0.01 and 2, respectively. The left grids illustrate the original-causality while the right grid shows the surrogate-causality. All grid entries are averaged over 50 simulations under the standard configuration. The surrogate-causalities are averaged over $K = 10$ surrogate realizations.

Then, we merge the linear and nonlinear parts of the derivatives to construct the full governing equations,

$$\left(\frac{dx_j}{dt}\right) = \left(\frac{dx_j}{dt}\right)_{lin} + \left(\frac{dx_j}{dt}\right)_{nl}. \quad (23)$$

Finally, we assign coefficients to the individual terms and calibrate them to the data by using the gradient-descent based algorithm developed by Mariño and Míguez.³³

IV. RESULTS

In the following, we present the results of our analysis, which are divided into four categories: Evaluation of causality matrices, nonlinear strength variation, analysis of causal structures, and derivation of governing equations.

A. Evaluation of causality matrices

Our analysis of the Lorenz and Halvorsen systems indicates that the causality is predominantly driven by nonlinear properties. This is illustrated in Fig. 5, where the box plots show that all surrogate-based causalities measured by TE and CCM are significantly lower than the original causality. This is because the surrogate time series only exhibit the same linear properties as the original

time series while nonlinear effects are destroyed. We observe that a significant portion of TE and CCM can be attributed to nonlinear properties. As expected, we confirm that GC is indeed restricted to measuring linear causality as the original- and surrogate-GC are both on the same scale. The small deviations stem from the inaccuracies of the linear regression required for the calculation of GC. Analogously to Prichard and Theiler,³¹ we repeat the calculation where we use different random phases when calculating the surrogate-GC between two time series. Since the surrogate-GC almost diminishes, we conclude that GC—just as Pearson correlation—only depends on phase differences. Furthermore, our developed anti- and cross-causalities, which measure the causal flow from the linear properties of one time series to both the linear and nonlinear properties of another, vanish for all three inference methods. This further suggests that the causal flows are mainly dominated by nonlinearity.

To verify that our method only measures linear and nonlinear causality when the governing equations are fully linear and nonlinear, we performed the analysis for the models given in Eqs. (3) and (4). Figure 6 highlights the validity of our methods, as the fully linear model has predominantly linear causality because GC is significant and the original and surrogate TE and CCM have similar strengths. For the fully nonlinear model, we observe the opposite case, where GC is low and the surrogate TE and CCM are significantly lower than the original TE and CCM.

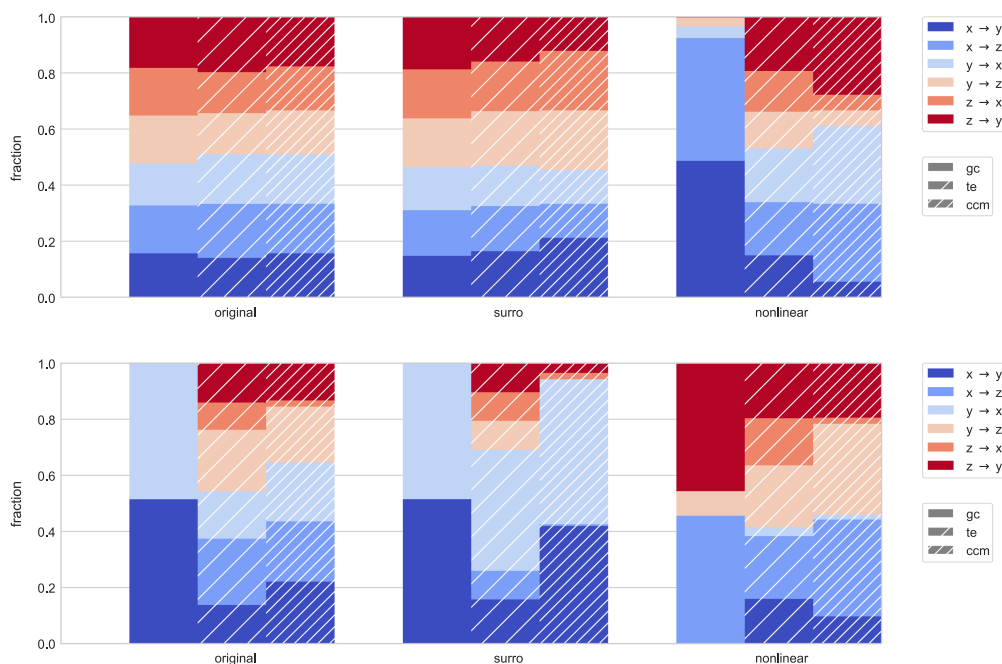


FIG. 8. Causality decomposition of the standard Lorenz (top row) and Halvorsen (bottom row) systems. For GC, TE, and CCM, we compute the original-, surrogate-, and nonlinear-causality, respectively. In order to obtain the contribution of each individual causal link to the causality of the whole system, we divide the causality of each link by the causality of the system. The contributions of the individual causal flows to the causality are mapped by color, while the different inference techniques are indicated by white stripes. The individual fractions are averaged over 50 simulations under the standard configuration. The surrogate-based causalities are averaged over $K = 10$ surrogate realizations.

B. Nonlinear strength variation

For the Lorenz and Halvorsen attractors, the analysis is repeated for variations in the degree of nonlinearity. While both systems diverge for nonlinearity degrees less or equal to 0, the upper bounds can be chosen arbitrarily as we do not observe significant changes to the attractor form. We conclude that the level of nonlinearity solely affects the scale of the attractors. Figure 2 illustrates the attractors for a selection of different parameter configurations. This behavior directly translates to the causality as indicated for the Lorenz system in Fig. 7. As expected, we find that the original causality is significantly larger than the surrogate causality for both TE and CCM across all degrees of nonlinearity. Furthermore, we observe that the grids show no visible gradient, which implies that the causality is independent of the degree of nonlinearity.

C. Analysis of causal structures

On a finer scale, we find that the causal structure of linear and nonlinear causality differs significantly for the Lorenz system, as illustrated in Fig. 9. We observe that the x and y pair is mainly driven by linear properties as it dominates the surrogate-causalities of GC and CCM—with both directions contributing equal amounts. In contrast, the surrogate-TE indicates that the direction x to y dominates the linear causality with a fraction of around 41%.

result is in line with the governing equations as the equation for x contains a linear contribution from y , while the equation for y contains a linear and nonlinear contribution from x . The rest of the system-causalities are more or less split evenly across the remaining flows.

In comparison, all flows in the Halvorsen attractor contribute approximately equally to the causality across all causality types and inference techniques, as depicted in Fig. 8. This causal structure is expected due to the circulant nature of the governing equations.

D. Derivation of governing equations

In order to verify our rationale, we apply it to the Lorenz and Halvorsen systems with their corresponding CCM-causal graphs, as computed from Eqs. (11) and (20), depicted in Fig. 9. The equations derived for the Lorenz system are

$$\begin{aligned} \frac{dx}{dt} &= y - x, \\ \frac{dy}{dt} &= x - xz - y, \\ \frac{dz}{dt} &= xy - z, \end{aligned} \tag{24}$$

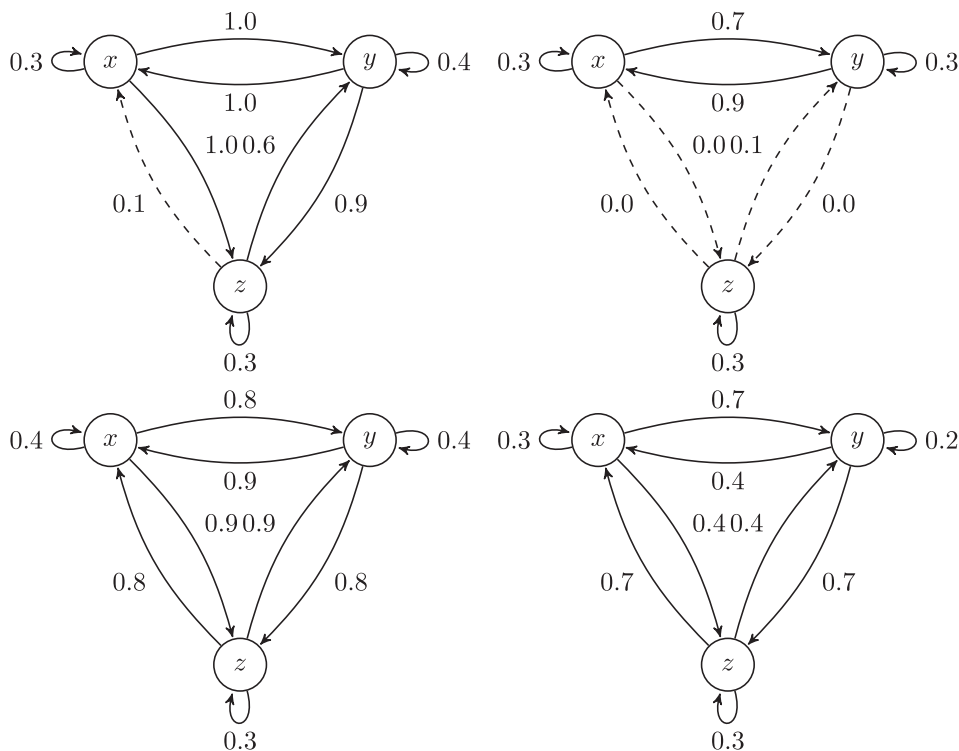


FIG. 9. CCM-causality graphs of the standard Lorenz (top row) and Halvorsen (bottom row) systems. The graphs depict the original (left) and the linear (right) CCM causality between the variables. The dashed lines indicate that the measured causality is not significant ($\theta < 0.1$). Note that the causalities in the loops are determined using the cross-CCM. The surrogate-based causal links on the right are averaged over $K = 10$ surrogate realizations.

TABLE I. Governing equations of the stock indices. This table depicts the derived governing equations for the six economies. The first column shows the time derivative of the economy while the second and third columns contain the linear and nonlinear terms before the COVID-19 outbreak, respectively. Analogously, the fourth and fifth columns contain the terms after the COVID-19 outbreak.

Economy	Before outbreak linear	Before outbreak nonlinear	After outbreak linear	After outbreak nonlinear
$\frac{dx_{eu}}{dt}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{eu} x_{px}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{px}$	$x_{jp}x_{px} + x_{us}x_{cn} + x_{us}x_{ee} + x_{us}x_{jp} + x_{cn}x_{ee} + x_{cn}x_{jp} + x_{ee}x_{jp} + x_{ee}x_{px} + x_{ee}x_{px} + x_{cn}x_{px} + x_{us}x_{px}$
$\frac{dx_{us}}{dt}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{cn} x_{px}$	$x_{eu} + x_{us} + x_{cn} + x_{jp} + x_{px}$	$x_{cn}x_{px} + x_{ee}x_{px} + x_{cn}x_{ee}$
$\frac{dx_{cn}}{dt}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{eu} x_{us}$	$x_{cn} + x_{ee} + x_{px}$	$x_{eu}x_{ee} + x_{us}x_{ee} + x_{eu}x_{us} + x_{eu}x_{px} + x_{ee}x_{px} + x_{us}x_{px}$
$\frac{dx_{ee}}{dt}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{jp} x_{px}$	$x_{eu} + x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{eu}x_{cn} + x_{us}x_{cn} + x_{jp}x_{px} + x_{us}x_{jp} + x_{cn}x_{jp} + x_{eu}x_{jp} + x_{eu}x_{us} + x_{eu}x_{px} + x_{cn}x_{px} + x_{us}x_{px}$
$\frac{dx_{jp}}{dt}$	$x_{cn} + x_{jp} + x_{px} + x_{eu}x_{ee}$	$x_{eu}x_{cn} + x_{us}x_{cn} + x_{us}x_{ee} + x_{cn}x_{jp} + x_{eu}x_{us} + x_{eu}x_{px} + x_{ee}x_{px} + x_{cn}x_{px} + x_{us}x_{jp}$	$x_{us} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{us}x_{ee} + x_{eu}x_{us} + x_{eu}x_{ee}$
$\frac{dx_{px}}{dt}$	$x_{eu} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{us} x_{px}$	$x_{eu} + x_{cn} + x_{ee} + x_{jp} + x_{px}$	$x_{us}x_{ee} + x_{eu}x_{us} + x_{eu}x_{ee}$

while the equations for the Halvorsen system are given by

$$\begin{aligned} \frac{dx}{dt} &= x - y - z - y^2, \\ \frac{dy}{dt} &= y - z - x - z^2, \\ \frac{dz}{dt} &= z - x - y - x^2. \end{aligned} \tag{25}$$

By comparing them to Eqs. (1) and (2), we find that our rationale reproduces the terms of the Lorenz and Halvorsen differential equations correctly. The calibration of the coefficients using the algorithm by Mariño and Míguez²³ yielded the correct coefficients $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$ for the Lorenz system and $a = 1.3$ for the Halvorsen system with errors less than $1e-4$, respectively.

These results are stable for thresholds $\theta < 0.2$. In order to ensure robustness, we repeat the analysis for different initial conditions and find that for a simulation length $T \geq 5000$ the causality inference and, hence, the equation derivation is stable.

In the following, we apply our rationale to a real-world system and derive the governing equations from the causal interactions between stock indices of six major economies: European Union, United States, China, Emerging Markets, Japan, and Pacific excluding Japan. The derived equations are shown in Table I, where we find that all economies except Japan have only one nonlinear term before the February 2020 COVID-19 pandemic outbreak. In contrast, the equations for the post-pandemic outbreak phase have at least three nonlinear terms in all economies, suggesting that nonlinearity has increased in the financial market. We find this result to be robust to changes in causal inference technique and thresholds $\theta < 0.2$. Furthermore, we would like to emphasize that we repeated the analysis,

where we remap the rank-ordered time series onto a Gaussian distribution. Since the results remain practically unchanged, we conclude that our results are mainly driven by dynamic nonlinearities.

This result suggests that the COVID-19 pandemic has led to a fundamental change in the global financial market, which seems to make sense in light of the equity rally that was detached from the real economy.³⁴ Looking forward, as indicated by Haluszczyński *et al.*,⁸ a large amount of nonlinearity in the market can potentially serve as an early indicator for financial crises.

Note that we do not assign coefficients to the individual terms of the equations as the calibration method by Mariño and Míguez³³ fails due to limited data and high dimensionality. Other equation derivation algorithms, such as Sparse Identification of Nonlinear Dynamics (SINDy),¹² also face this problem. SINDy is capable of generating equations with coefficients, but the equations diverge after a few simulation steps. Developing more sophisticated calibration methods to solve this problem is part of future research that is beyond the scope of this paper.

V. DISCUSSION

In this work, we analyzed the linear and nonlinear causal relations between variables in dynamical systems using different inference techniques and Fourier transform surrogates, which filter out the nonlinear properties of time series. We find for Lorenz and Halvorsen that nonlinearity is a key driver of causality and that nonlinear causality is independent of the strength of nonlinear terms in the governing equations. Furthermore, we developed a constructive and fully transparent rationale to derive the correct governing equations of the Lorenz and Halvorsen attractors directly from their causal structures—the resulting ease of interpretation is

the main advantage in comparison to black-box machine learning approaches. Finally, we applied our methods to stock indices from different economies and found that the outbreak of the COVID-19 pandemic triggered a structural change in the global financial markets.

This work can be extended in several directions. First, the provided framework can be deployed with further causal inference techniques and applied to other synthetic systems to confirm the universality of our results. Furthermore, new methods for calibrating the equation coefficients can be developed in order to address the problems of limited data and high dimensionality in real-world applications—this would enable precise predictions and the detection of unknown chaos and attractors.

ACKNOWLEDGMENTS

We would like to thank the DLR and Allianz Global Investors for providing data and computational resources.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Haochun Ma: Conceptualization (equal); Formal analysis (equal); Funding acquisition (equal); Methodology (equal); Software (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Alexander Haluszczynski:** Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Writing – review & editing (equal). **Davide Prosperino:** Formal analysis (equal); Software (equal); Visualization (equal); Writing – review & editing (equal). **Christoph R ath:** Conceptualization (equal); Methodology (equal); Supervision (equal); Writing – review & editing (equal).

DATA AVAILABILITY



The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- H. R. Brown and D. Lehmkuhl, "Einstein, the reality of space and the action–reaction principle," in *Einstein, Tagore and the Nature of Reality* (Routledge, 2016), pp. 27–54.
- E. Lorenz, "The butterfly effect," in *World Scientific Series on Nonlinear Science Series A* (World Scientific, 2000), Vol. 39, pp. 91–94.
- C. W. Granger, *Essays in Econometrics: Collected Papers of Clive W.J. Granger* (Cambridge University Press, 2001), Vol. 32.
- T. Schreiber, "Measuring information transfer," *Phys. Rev. Lett.* **85**, 461 (2000).
- G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *Science* **338**, 496–500 (2012).
- J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos* **28**, 075310 (2018).
- M. Paluř, V. Albrecht, and I. Dvoř ak, "Information theoretic test for nonlinearity in time series," *Phys. Lett. A* **175**, 203–209 (1993).
- A. Haluszczynski, I. Laut, H. Modest, and C. R ath, "Linear and nonlinear market correlations: Characterizing financial crises and portfolio optimization," *Phys. Rev. E* **96**, 062315 (2017).
- J. Hlinka, D. Hartman, M. Vejmelka, D. Novotn a, and M. Paluř, "Non-linear dependence and teleconnections in climate data: Sources, relevance, nonstationarity," *Clim. Dyn.* **42**, 1873–1886 (2014).
- J. L. Breeden and A. H ubler, "Reconstructing equations of motion from experimental data with unobserved variables," *Phys. Rev. A* **42**, 5817 (1990).
- T. Eisenhammer, A. H ubler, N. Packard, and J. S. Kelso, "Modeling experimental time series with ordinary differential equations," *Biol. Cybernet.* **65**, 107–112 (1991).
- S. L. Brunton, J. L. Proctor, and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," *Proc. Natl. Acad. Sci. U.S.A.* **113**, 3932–3937 (2016).
- B. C. Daniels and I. Nemenman, "Automated adaptive inference of phenomenological dynamical models," *Nat. Commun.* **6**, 1–8 (2015).
- K. Champion, B. Lusch, J. N. Kutz, and S. L. Brunton, "Data-driven discovery of coordinates and governing equations," *Proc. Natl. Acad. Sci. U.S.A.* **116**, 22445–22451 (2019).
- E. Hairer, S. P. N orsett, and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, 1993).
- E. N. Lorenz, "Deterministic nonperiodic flow," *J. Atmos. Sci.* **20**, 130–141 (1963).
- S. Vaidyanathan and A. T. Azar, "Adaptive control and synchronization of Halvorsen circulant chaotic systems," in *Advances in Chaos Theory and Intelligent Control* (Springer, 2016), pp. 225–247.
- V. P. Thoai, M. S. Kahkeshi, V. V. Huynh, A. Ouannas, and V.-T. Pham, "A nonlinear five-term system: Symmetry, chaos, and prediction," *Symmetry* **12**, 865 (2020).
- S. L. Bressler and A. K. Seth, "Wiener–Granger causality: A well established methodology," *Neuroimage* **58**, 323–329 (2011).
- L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for Gaussian variables," *Phys. Rev. Lett.* **103**, 238701 (2009).
- M. Mynter, "Evaluation and extension of the transfer entropy calculus for the measurement of information flows between futures time series during the COVID-19 pandemic," Master thesis (Ludwig-Maximilians-Universit at M unchen, 2021) (unpublished).
- S. Baur and C. R ath, "Predicting high-dimensional heterogeneous time series employing generalized local states," *Phys. Rev. Res.* **3**, 023215 (2021).
- M. Paluř and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Phys. Rev. E* **75**, 056211 (2007).
- J. M. McCracken and R. S. Weigel, "Convergent cross-mapping and pairwise asymmetric inference," *Phys. Rev. E* **90**, 062903 (2014).
- M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A* **45**, 3403 (1992).
- B. Cummins, T. Gedeon, and K. Spendlove, "On the efficacy of state space reconstruction methods in determining causality," *SIAM J. Appl. Dyn. Syst.* **14**, 335–381 (2015).
- L. Overbey and M. Todd, "Effects of noise on transfer entropy estimation for damage detection," *Mech. Syst. Signal Process.* **23**, 2178–2191 (2009).
- P. Krishna and A. K. Tangirala, "Inferring direct causality from noisy data using convergent cross mapping," in *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* (IEEE, 2019), pp. 1523–1528.
- C. R ath and R. Monetti, "Surrogates with random Fourier phases," in *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference* (World Scientific, 2009), pp. 274–285.
- C. R ath, M. Gliozzi, I. Papadakis, and W. Brinkmann, "Revisiting algorithms for generating surrogate time series," *Phys. Rev. Lett.* **109**, 144101 (2012).
- D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Phys. Rev. Lett.* **73**, 951 (1994).
- X. Wan, W. Wang, J. Liu, and T. Tong, "Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range," *BMC Med. Res. Methodol.* **14**, 1–13 (2014).
- I. P. Mari no and J. M iguez, "An approximate gradient-descent method for joint parameter estimation and synchronization of coupled chaotic systems," *Phys. Lett. A* **351**, 262–267 (2006).
- J. Cox, D. L. Greenwald, and S. C. Ludvigson, "What explains the COVID-19 stock market?" Technical Report, National Bureau of Economic Research, 2020.

RESEARCH ARTICLE | JUNE 12 2023

Efficient forecasting of chaotic systems with block-diagonal and binary reservoir computing

Haochun Ma; Davide Prosperino; Alexander Haluszczynski; Christoph R ath  

 Check for updates

Chaos 33, 063130 (2023)

<https://doi.org/10.1063/5.0151290>


View
Online


Export
Citation

CrossMark

AIP Advances

Why Publish With Us?

 25 DAYS average time to 1st decision	 740+ DOWNLOADS average per article	 INCLUSIVE scope
---	--	---

[Learn More](#)

 AIP
Publishing

Efficient forecasting of chaotic systems with block-diagonal and binary reservoir computing

Cite as: Chaos **33**, 063130 (2023); doi: 10.1063/5.0151290

Submitted: 20 March 2023 · Accepted: 12 May 2023 ·

Published Online: 12 June 2023



View Online



Export Citation



CrossMark

Haochun Ma,^{1,2} Davide Prosperino,^{1,2} Alexander Haluszczynski,² and Christoph Räth^{3,a)} 

AFFILIATIONS

¹Department of Physics, Ludwig-Maximilians-Universität, Schellingstraße 4, 80799 Munich, Germany

²Allianz Global Investors, risklab, Seidlstraße 24, 80335 Munich, Germany

³Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für KI Sicherheit, Wilhelm-Runge-Straße 10, 89081 Ulm, Germany

^{a)}Author to whom correspondence should be addressed: christoph.raeth@dlr.de

ABSTRACT

The prediction of complex nonlinear dynamical systems with the help of machine learning has become increasingly popular in different areas of science. In particular, reservoir computers, also known as echo-state networks, turned out to be a very powerful approach, especially for the reproduction of nonlinear systems. The reservoir, the key component of this method, is usually constructed as a sparse, random network that serves as a memory for the system. In this work, we introduce block-diagonal reservoirs, which implies that a reservoir can be composed of multiple smaller reservoirs, each with its own dynamics. Furthermore, we take out the randomness of the reservoir by using matrices of ones for the individual blocks. This breaks with the widespread interpretation of the reservoir as a single network. In the example of the Lorenz and Halvorsen systems, we analyze the performance of block-diagonal reservoirs and their sensitivity to hyperparameters. We find that the performance is comparable to sparse random networks and discuss the implications with regard to scalability, explainability, and hardware realizations of reservoir computers.

Published under an exclusive license by AIP Publishing. <https://doi.org/10.1063/5.0151290>

The application of reservoir computers to various fields in science and technology yields very promising and fast advancing results due to their capabilities in forecasting chaotic attractors, inferring unmeasured values in systems, and recognizing speech. While the construction of a reservoir computer is rather simple in comparison to other machine learning techniques, the architecture and functionality of them is in many regards still a black-box since at its core, the reservoir is usually still chosen as a random network. Thus, we replace the network with block-diagonal matrices dividing it into multiple smaller reservoirs, take out the randomness, and show that these alterations still deliver an equal quality of short- and long-term predictions. This architecture breaks with the common interpretation of the reservoir as a single network and may prove to be more scalable and easier to implement in hardware than their more complex variants while still performing as well.

I. INTRODUCTION

The analysis and modeling of complex dynamic systems is a key challenge across various disciplines in science, engineering, and

economics.¹ While machine learning approaches, like generative adversarial networks, can provide excellent predictions on dynamical systems,² difficulties with vast data requirements, the large number of hyperparameters, and lack of interpretability limit their usefulness in some scientific applications.³ However, it is required to fundamentally understand how, when, and why the models are working in order to prevent the risk of misinterpreting the results if deeper methodological knowledge is missing.⁴

In the context of complex system research, reservoir computers (RCs)⁵ have emerged for quantifying and predicting the spatiotemporal dynamics of chaotic nonlinear systems. They represent a special kind of recurrent neural networks (RNNs) and are often referred to as echo-state networks (ESNs).⁶ The core of the model is a fixed reservoir, which is a complex network with connections according to a predefined network topology. The input data are fed into the nodes of the reservoir and solely the weights of the readout layer, which transform the reservoir response to output variables, are subject to optimization via linear regression. This makes the learning extremely fast, comparatively transparent, and prevents the vanishing gradient problem of other RNN methods.⁷

The topology of a reservoir, or the arrangement of the nodes and connections within it, can have a significant impact on the

performance of a reservoir computing system.⁸ In current state-of-the-art models, the topology of the reservoir is often chosen randomly,⁹ with the hope that the resulting dynamics will be sufficiently complex to allow for good performance on a given task. However, this approach can be hit-or-miss,^{10,11} and it is impossible to know a priori how the topology of the reservoir will affect the performance of the system.

While Maass *et al.*¹² and Jaeger¹³ introduced ESNs with reservoirs being modeled as a random Erdős–Rényi network, Watts and Strogatz,¹⁴ Albert and Barabási,¹⁵ and others have shown that random networks are far from being common in physics, biology, or sociology. Instead, more complex networks like scale-free, small-world, or intermediate forms of networks^{16,17} are most often found in real-world applications. Further approaches to make reservoir computing more explainable have been made in recent years, with, e.g., Haluszczynski and R ath¹¹ comparing different network construction algorithms, Griffith *et al.*¹⁸ introducing very low connectivity networks, and Carroll and Pecora¹⁰ analyzing the effect of network symmetries on prediction performance. However, there still remain open questions about the functionality of reservoir computers, which need to be answered for developing new algorithms, for fine-tuning the system for specific applications, or building efficient hardware realizations of RCs.

In this work, we break with the interpretation of the reservoir as a single network by deliberately using block-diagonal matrices as reservoirs. This implies that we decompose the reservoir into multiple smaller reservoirs as outlined in Sec. II B. Furthermore, we use matrices of ones as the blocks, which take out the randomness of the network completely. We assess the ability of block-diagonal reservoirs for short- and long-term predictions by comparing the measures discussed in Sec. II C to the standard RC setup.

II. METHODS

We structure our methods section into three different parts: benchmark models, reservoir computing, and prediction performance measures.

A. Benchmark models

We perform our analyses on two synthetic example models, which exhibit chaotic behavior and are three-dimensional autonomous, dissipative flows.

1. Lorenz system

As in Pathak *et al.*¹⁹ and Lu *et al.*,²⁰ we use the Lorenz system, which was initially used for modeling atmospheric convection,²¹ as an example for replicating chaotic attractors using reservoir computing,

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y - x), \\ \frac{dy}{dt} &= x(\rho - z) - y, \\ \frac{dz}{dt} &= xy - \beta z,\end{aligned}\quad (1)$$

where the standard parameterization for chaotic behavior is $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$.

2. Halvorsen system

As in Herteux and R ath,²² we use the Halvorsen system for our analyses, which has a cyclic symmetry. While the nonlinear terms of the Lorenz system are mixed products of two different variables, the Halvorsen system²³ entails only non-mixed quadratic nonlinearities,

$$\begin{aligned}\frac{dx}{dt} &= ax - by - bz - y^2, \\ \frac{dy}{dt} &= ay - bz - bx - z^2, \\ \frac{dz}{dt} &= az - bx - by - x^2,\end{aligned}\quad (2)$$

where $a = 1.3$ and $b = 4$ are the standard parameter choice.

3. Simulating and splitting data

If not stated otherwise, we solve the differential equations of the synthetic system using the Runge–Kutta method²⁴ for $T = 70\,000$ steps and a discretization of $dt = 0.02$ in order to ensure a sufficient manifestation of the attractor. We discard the initial transient of $T = 50\,000$ steps and use the remaining steps for training $T_{\text{train}} = 10\,000$ and testing $T_{\text{test}} = 10\,000$ of the RCs. In order to get robust results, we vary the starting points on the attractor by using the rounded last point of one data sample as the starting point for the next. The initial starting points for the Lorenz and Halvorsen systems are $(-14, -20, 25)$ and $(-6.4, 0, 0)$, respectively. This setting is comparable to the ones used by Griffith *et al.*¹⁸ and Haluszczynski and R ath.¹¹ Figure 1 illustrates the attractors and trajectories of the simulated data.

B. Reservoir computing

A reservoir computer (RC)^{12,13,25,26} is an artificial recurrent neural network (RNN) that relies on a static internal network called *reservoir*. The term static means that, unlike other RNN approaches, the reservoir remains fixed once the network is constructed. The same is true for the input weights. Therefore, the RC system is computationally very efficient since the training process only involves optimizing the output layer. As a result, fast training and high model dimensionality are computationally feasible, making RC well suited for complex real-world applications.

1. Algorithm

The reservoir **A** is usually constructed as a sparse Erdős–R enyi random network²⁷ with dimensionality or number of nodes d . However, in this paper, we replace the network structure of the reservoir with a block-diagonal matrix of ones. This breaks with the widespread interpretation of the reservoir as a single network.

To feed the n -dimensional input data $\mathbf{u}(t)$ into reservoir **A**, a $d \times n$ input matrix \mathbf{W}_{in} is constructed, which defines how strongly each input dimension influences every single node. The elements of \mathbf{W}_{in} are chosen to be uniformly distributed random numbers within the interval $[-1, 1]$.

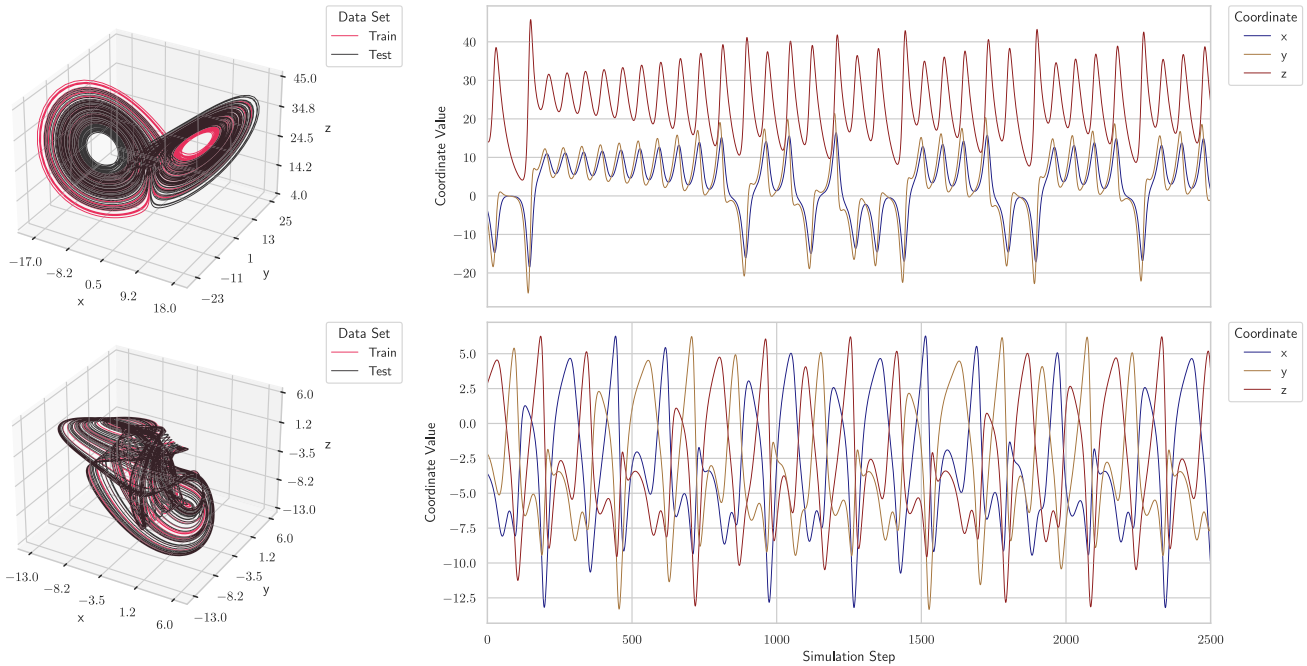


FIG. 1. Attractor (left) and coordinate trajectories (right) of the Lorenz (top) and Halvorsen (bottom) system. The left columns show the attractors of the two systems, where the first $T_{train} = 10\,000$ steps are used for training (red) and the subsequent $T_{test} = 10\,000$ steps are used for testing the prediction (black). The right columns illustrate the coordinate trajectories of both systems for the first $T = 2500$ steps. The parameters of the systems and their simulations are described in Sec. II A.

The dynamics of the RC system are contained in its d -dimensional reservoir states $\mathbf{r}(t)$. Being initially set to $r_i(0) = 0$ for all nodes, the time evolution can be defined using a recurrent formulation,

$$\mathbf{r}(t + 1) = f(\mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t)), \quad (3)$$

where f is a limited, nonlinear function—as is common, we use the hyperbolic tangent. Before the training process is started, the RC system should be initialized during a washout phase of t_w time steps in order to synchronize the reservoir states $\mathbf{r}(t)$ with the dynamics of the input signal $\mathbf{u}(t)$. Furthermore, in order to break potential problems arising from the anti-symmetry of the hyperbolic tangent, we use a quadratic readout as explained by Herteux and R ath.²² This means that the squared elements of the reservoir states are appended $\mathbf{r} \mapsto \{\mathbf{r}, \mathbf{r}^2\}$.

To obtain n -dimensional output from the (matrix of) reservoir states $\mathbf{r}(t)$, an output-mapping function \mathbf{W}_{out} is needed. This is accomplished by acquiring a sufficient number of reservoir states $\mathbf{r}(t_w, \dots, t_w + t_{T_{train}})$ and then choosing an output-mapping matrix \mathbf{W}_{out} such that the output of the reservoir is as close as possible to the known real data (matrix) $\mathbf{u}(t_w, \dots, t_w + t_{T_{train}})$. Then, the training can be executed by using Ridge regression,²⁸

$$\mathbf{W}_{out} = (\mathbf{r}^T \cdot \mathbf{r} + \beta \cdot \mathbf{I})^{-1} \mathbf{r}^T \cdot \mathbf{u}, \quad (4)$$

where β is the regularization constant that prevents overfitting and \mathbf{I} denotes a identity matrix. The predicted state $\mathbf{v}(t)$ can be obtained

by multiplying the output matrix with the reservoir state $\mathbf{r}(t)$,

$$\mathbf{v}(t) = \mathbf{W}_{out} \cdot \mathbf{r}(t). \quad (5)$$

After training, the predicted state $\mathbf{v}(t)$ can be fed back in the activation function as input instead of the actual data $\mathbf{u}(t)$ by combining Eqs. (3) and (5). The resulting recursive form of the equation for the reservoir states $\mathbf{r}(t)$ allows us to create predicted trajectories of arbitrary length,

$$\mathbf{r}(t + 1) = f(\mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{W}_{out} \cdot \mathbf{r}(t)). \quad (6)$$

2. Block-diagonal reservoir

The main focus of this work is to verify that a reservoir can be divided into multiple smaller reservoirs without limiting its prediction performance. Therefore, we choose our $d \times d$ dimensional reservoir topology to be a block-diagonal matrix with blocks \mathbf{J}_i , of size $b \times b$, where $i \in \{1, 2, \dots, \lfloor \frac{d}{b} \rfloor\}$. Each of the blocks essentially represents a separate smaller reservoir,

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{\lfloor \frac{d}{b} \rfloor} \end{pmatrix}. \quad (7)$$

As usual, the reservoir topology is rescaled to a target spectral radius ρ^* . Therefore, the spectral radius of the matrix \mathbf{J} needs to

be determined first, which is the largest absolute eigenvalue of the matrix,

$$\rho(\mathbf{J}) = \max\{|\lambda_1|, \dots, |\lambda_d|\}. \tag{8}$$

The time complexity of obtaining the eigenvalues of a d -dimensional matrix using bi-diagonalization is $\mathcal{O}(d^3)$.²⁹ We can speed up the calculation by a factor of

$$\frac{d^3}{b^3 \cdot \lfloor \frac{d}{b} \rfloor} \approx \left(\frac{d}{b}\right)^2, \tag{9}$$

since the eigenvalues of a block-diagonal matrix is the list of the eigenvalues of the blocks,

$$\rho(\mathbf{J}) = \max\left\{\rho(\mathbf{J}_1), \dots, \rho\left(\mathbf{J}_{\lfloor \frac{d}{b} \rfloor}\right)\right\}. \tag{10}$$

Then, the scaled reservoir \mathbf{A} , which is finally used in the RC, is given by

$$\mathbf{A} = \frac{\rho^*}{\rho(\mathbf{J})} \mathbf{J}. \tag{11}$$

In Secs. II B 3 and II B 4, we describe the different types of blocks \mathbf{J}_i that we study in this work: first, where the blocks are Erdős–Rényi networks, and second, where the blocks are matrices of ones.

3. Blocks of Erdős–Rényi networks

First, we choose the individual blocks \mathbf{J}_i to be Erdős–Rényi networks with a connection probability of $p = 0.02$.¹⁴ In our analyses, we distinguish between two cases:

1. *Individual blocks:* Each block \mathbf{J}_i is constructed separately with a different random seed.
2. *Equal blocks:* All the blocks are equal to each other and thus, the Erdős–Rényi network only needs to be constructed once,

$$\mathbf{J}_1 = \mathbf{J}_2 = \dots = \mathbf{J}_{\lfloor \frac{d}{b} \rfloor}.$$

Furthermore, this case delivers another increase in eigenvalue decomposition speed by a factor of $\lfloor \frac{d}{b} \rfloor$.

4. Blocks of matrices of ones

In order to take out the randomness of the reservoir, we construct it so that each block \mathbf{J}_i is a matrix full of ones. This has several special implications: first, we do not need to calculate the spectral radius of the reservoir $\rho(\mathbf{J})$ anymore, since it is equal to the block-size b ,

$$\mathbf{A} = \frac{\rho^*}{b} \mathbf{J}. \tag{12}$$

Furthermore, this reservoir architecture implies that in every iteration, each block \mathbf{J}_i acts as an averaging operator on the reservoir states, as explained in the following. This is comparable to average pooling layers of other machine learning techniques since the averaging reduces the dimensionality of the reservoir states and

“extracts” primarily features that are more robust.³¹ We denote the mean between the i th and j th row of reservoir state $\mathbf{r}(t)$ as

$$\bar{r}_{ij}(t) \equiv \frac{1}{j-i+1} \sum_i^j r_i(t). \tag{13}$$

Then, each block yields a vector of size $b \times 1$, which has equal values. For example, the first row of the multiplication $\mathbf{J} \cdot \mathbf{r}(t)$ reads

$$\left[\overbrace{1, \dots, 1}^b, \overbrace{0, \dots, 0}^{d-b} \right] \cdot \mathbf{r}(t) = \sum_{i=1}^b r_i(t) = b \cdot \bar{r}_{1:b}(t). \tag{14}$$

This is repeated exactly for the first b rows. Consequently, the reservoir multiplication $\mathbf{A} \cdot \mathbf{r}(t)$ from Eq. (3) yields

$$\mathbf{A} \cdot \mathbf{r}(t) = \frac{\rho^*}{b} \mathbf{J} \cdot \mathbf{r}(t) = \rho^* \begin{pmatrix} \bar{r}_{1:b}(t) \\ \bar{r}_{1:b}(t) \\ \vdots \\ \bar{r}_{(i-1) \cdot b + 1:i \cdot b}(t) \\ \bar{r}_{(i-1) \cdot b + 1:i \cdot b}(t) \\ \vdots \\ \bar{r}_{d-b+1:d}(t) \\ \bar{r}_{d-b+1:d}(t) \\ \vdots \end{pmatrix}. \tag{15}$$

Therefore, the reservoir multiplication contribution is identical for each block, which means that at each training step, the reservoir memory is the same for each block. This directly implies lower computational costs.

5. Implications for hardware reservoir computers

By separating the reservoir into multiple smaller reservoirs and by taking out the randomness we significantly reduce the complexity of the architecture, especially with regards to hardware implementations. Examples of the reservoir topologies and their respective spring-layouts according to the *Fruchterman–Reingold* force-directed algorithm³⁰ are illustrated in Fig. 2. We observe that the networks constructed fully or partly with Erdős–Rényi Networks have more complex interconnected network structures, while the networks with blocks of ones are ordered and have clear separate unified reservoirs.

This leads us to make the following assumptions on potential implications for hardware RCs:

- *Improved generalization:* Similar to ensemble methods³² from other machine learning methods, the use of multiple smaller reservoirs can lead to a more diverse representation of the inputs. This potentially reduces the risk of overfitting and improves generalization to unseen data.
- *Enhanced robustness:* The failure of one reservoir unit can be potentially compensated by the other units, which can still contribute to the processing of the inputs. This makes the system more robust to noise and errors, especially in hardware implementations.

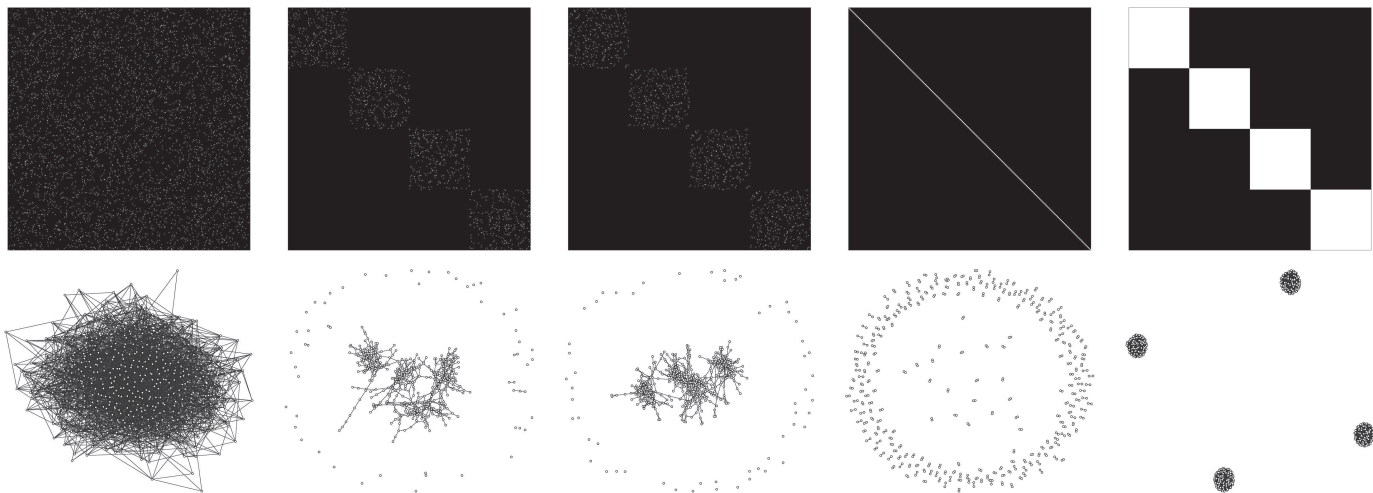


FIG. 2. Reservoir topologies (top) and corresponding spring-layouts³⁰ (bottom). The number of nodes is $d = 500$ for all reservoir topologies. In the top row, the white entries denote the presence of a network connection. In the bottom row, each node is represented by a white circle with black edges, while the connections are represented by black lines. From left to right, the illustrated network topologies are: (1) ordinary topology using a Erdős–Rényi graph, (2) block-diagonal topology with different Erdős–Rényi graphs of size $b = 125$ as blocks, (3) block-diagonal topology with equal Erdős–Rényi graphs of size $b = 125$ as blocks, (4) block-diagonal topology with matrices of ones of size $b = 2$ as blocks, and (5) block-diagonal topology with matrices of ones of size $b = 125$ as blocks.

- **Better scalability:** The computational and memory requirements can be reduced by parallelization, making the system more scalable and suitable for deployment on high-dimensional data.

Each of these assumptions requires different experimental setups—hence validating these potential implications is beyond the scope of this paper and is the subject of further research.

C. Measuring prediction performance

When forecasting nonlinear dynamical systems such as chaotic attractors, the goal of the predictions is not only to exactly replicate the actual short-time trajectory but also to reproduce the long-term statistical properties of the system called *climate*. This is important because by definition chaotic systems exhibit sensitive dependence on initial conditions and therefore small disturbances grow exponentially fast. Consequently, even if at first the short-term prediction is perfect, at some stage already numerical inaccuracies lead to the separation of the predicted and actual trajectories. However, for many applications, this is not a problem as long as the predicted trajectory still leads to the same attractor. In order to quantify this behavior, quantitative measures are needed that grasp the complex dynamics of the system. Therefore, we use the measures applied in the paper by Haluszczynski and R ath¹¹ and Haluszczynski.²⁶

1. Forecast horizon

To quantify the quality and duration of the short-term prediction of the trajectory, we use a fairly simple measure, which we call forecast horizon, as used by Haluszczynski and R ath.¹¹ For that, we track the number of time steps during which the predicted $\mathbf{v}(t)$ and the actual trajectory $\mathbf{v}_R(t)$ are matching. As soon as one of the

three coordinates exceeds certain deviation thresholds we consider the trajectories as not matching anymore. Throughout our study, we use

$$\tau = |\mathbf{v}(t) - \mathbf{v}_R(t)| > \delta, \tag{16}$$

where we define the thresholds as the standard deviation of the real data $\mathbf{v}_R(t)$,

$$\delta = \sigma(\mathbf{v}_R(t)). \tag{17}$$

The aim of this measure is that small fluctuations around the actual trajectory, as well as minor deviations do not exceed the threshold. A higher value means that the prediction is close to the “true” trajectory over a longer period of time and has not deviated yet, although the underlying system is chaotic.

2. Correlation dimension

To assess the structural complexity of an attractor, we calculate its correlation dimension, which measures the dimensionality of the space populated by the trajectory.³³ It belongs to the measures for fractal dimensionality, which have been proposed by Mandelbrot³⁴ in 1967. The correlation dimension is based on the correlation integral,

$$C(r) = \lim_{N \rightarrow \infty} \frac{1}{N^2} \sum_{i,j=1}^N \theta(r - |\mathbf{x}_i - \mathbf{x}_j|) = \int_0^r d^3 r' c(\mathbf{r}'), \tag{18}$$

where θ denotes the *Heaviside* function and $c(\mathbf{r}')$ is the standard correlation function. The integral represents the mean probability that two states in the phase space are close to each other at different time steps. This is the case if the distance between the two states is smaller than the threshold distance r .

TABLE I. Benchmark measures calculated on the predictions using the traditional reservoir architecture and on the test data. Using the traditional architecture with reservoir dimension $d = 500$ and target spectral radius $\rho^* = 0.1$, we vary the random seeds and attractor starting points, make predictions, and evaluate the quality of the predictions. This yields a total of $n = 10\,000$ different realizations of forecast horizons, correlation dimensions, and largest Lyapunov exponents. The average and standard deviation over these prediction measures for the respective systems are denoted in the columns “Lorenz traditional” and “Halvorsen traditional.” For each attractor starting point, we have a different test dataset. We calculate the correlation dimension and largest Lyapunov exponent of all test datasets and denote the average and standard deviation in the columns “Lorenz test” and “Halvorsen test.”

Measure/System	Lorenz traditional	Lorenz test	Halvorsen traditional	Halvorsen test
Forecast horizon τ	219 ± 10	∞	382 ± 12	∞
Correlation dimension ν	2.01 ± 0.09	2.02 ± 0.02	1.98 ± 0.07	1.99 ± 0.02
Lyapunov exponent λ_{\max}	0.86 ± 0.05	0.87 ± 0.03	0.72 ± 0.06	0.74 ± 0.03

The correlation dimension ν is then defined by the power-law relationship,

$$C(r) \propto r^\nu. \quad (19)$$

For self-similar, strange attractors, this relationship holds for a certain range of r , which needs to be properly calibrated. The calculation of the correlation dimension is done using the *Grassberger Procaccia* algorithm.³⁵ It is purely data-based and does not require any knowledge of the underlying governing equations of the system. One advantage of the correlation dimension over other fractal measures is that it can be calculated having a comparably small number of data points available. In the context of this work, mainly the relative comparison among various predictions and actual trajectories is of interest and, therefore, the accuracy of the absolute values is not the highest priority.

3. Lyapunov exponent

Besides the fractal dimensionality, the statistical climate of an attractor is also characterized by its temporal complexity represented by the *Lyapunov* exponents.³⁶ They describe the average rate of divergence of nearby points in the phase space, and thus measure sensitivity with respect to initial conditions. There is one exponent for each dimension in the phase space. If the system has at least one positive Lyapunov exponent, it is classified as chaotic. The magnitudes of λ_i quantify the time scale for which the system becomes unpredictable.³⁷ Since at least one positive exponent is the requirement for being classified as chaotic, it is sufficient for the purposes in this work to calculate only the largest Lyapunov exponent λ_{\max} :

$$d(t) = Ce^{\lambda_{\max} t}, \quad (20)$$

where $d(t)$ denotes the distance of two initially nearby states in phase space and the constant C is the normalization constant at the initial separation. Thus, instead of determining the full Lyapunov spectrum, we only need to find the largest one as it describes the overall system behavior to a large extent. Here, we use the *Rosenstein* algorithm.³⁸ As mentioned for the correlation dimension, mainly a relative comparison is of interest in order to characterize states of the system rather than determine the exact absolute values. Again, for this measure, no model or knowledge of the underlying governing equations is required.

D. Benchmarks

In order to evaluate whether the predictions using the modified RC architecture are on par with the traditional setup, we run multiple predictions for different reservoir dimensions, target spectral radii, attractor starting points, and random seeds.

We do not observe significant changes for network dimensions $d \geq 400$ and find a target spectral radius of $\rho^* = 0.1$ to have the best overall prediction results. This is consistent to the findings by Haluszczyński and R ath.¹¹

Thus, using the traditional architecture with reservoir dimension $d = 600$ and target spectral radius $\rho^* = 0.1$, we vary the random seeds and attractor starting points, make predictions, and evaluate the quality of the predictions. This yields a total of $n = 10\,000$ different realizations of forecast horizons, correlation dimensions, and largest Lyapunov exponents. The average and standard deviation over these prediction measures for the respective systems are denoted in the columns “Lorenz traditional” and “Halvorsen traditional” in Table I. We use these values as benchmarks for our analysis so that we can compare the prediction performance of our modified architecture to the traditional RC setup.

Furthermore, for each attractor starting point, we have a different test dataset. We calculate the correlation dimension and largest Lyapunov exponent of all test datasets and denote the average and standard deviation in the columns “Lorenz test” and “Halvorsen test” in Table I. These values can be seen as the “true” correlation dimension and largest Lyapunov exponent.

III. RESULTS

In the following, we present the results of our analyses for variations of different parameters to demonstrate the robustness of the modified architecture. The varied parameters are:

- *Network dimension d* : We vary the network dimension between $d \in \{400, 450, \dots, 600\}$ as these are sensible values in RC research. We specifically choose multiples of 50 in order to have a decent number $\lfloor \frac{d}{b} \rfloor$ of block-sizes b .
- *Block-size b* : We vary the block-size b and set it to all divisors of the network dimension d . We exclude the divisor $b = 1$ as it essentially takes out the reservoir. Also, we do not use $b = d$ since it represents again a single network as in the traditional architecture.

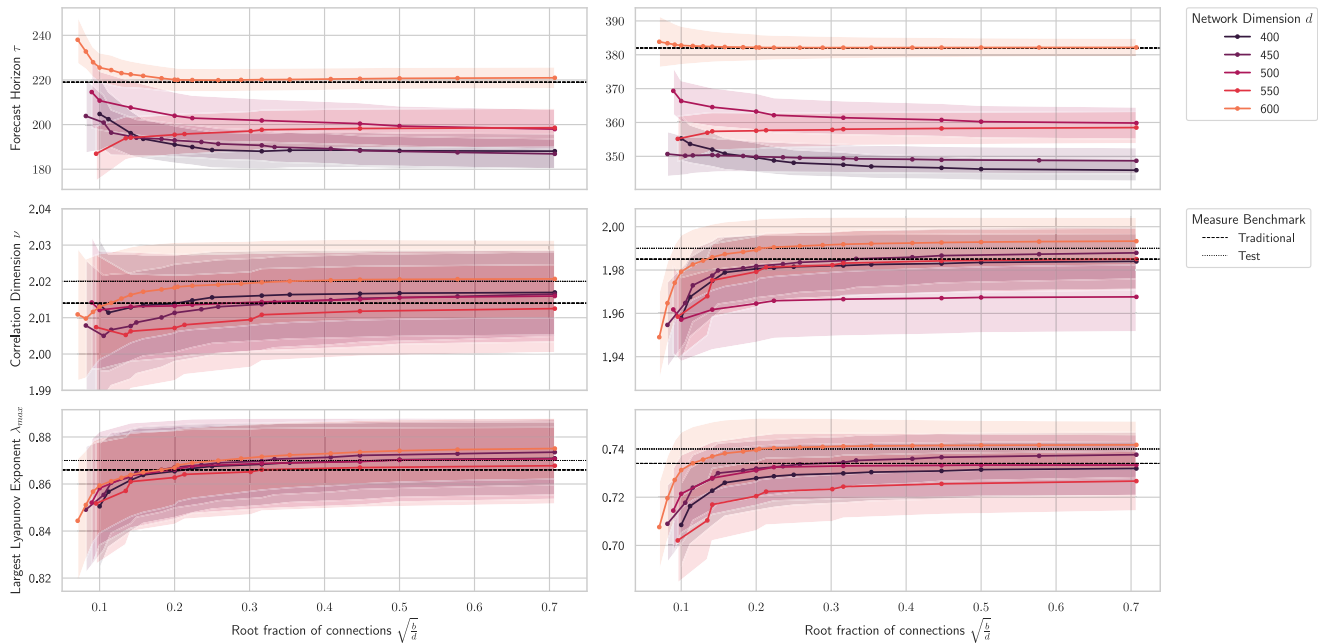


FIG. 3. Prediction measures for the Lorenz (left column) and Halvorsen systems (right column) for different random seeds of individual Erdős–Rényi blocks. The displayed prediction measures from top to bottom are: forecast horizon τ , correlation dimension ν , and the largest Lyapunov exponent λ_{max} . The differently colored lines represent different network dimensions with the corresponding shadowed area denoting the standard deviation over the variations. The dashed and dotted black horizontal lines represent the prediction measure benchmarks specified in Table I. The dashed line (lower) represents the average prediction performance of the traditional RC architecture, while the dotted line (higher) represents the average correlation dimension and largest Lyapunov exponent of the test data.

- **Target spectral radius ρ^* :** We vary the spectral radius from $\rho^* \in \{0.1, 0.2, \dots, 2.0\}$ and find that, similarly to the traditional architecture, the target spectral radius of $\rho^* = 0.1$ yields the most robust prediction results. Henceforth, we set $\rho^* = 0.1$ as default.
- **Attractor starting points:** We choose 500 different starting points on the attractor as explained in Sec. II A 3.
- **Random seeds:** We choose 100 different random seeds across all components of the RC architecture which have randomness: input weights \mathbf{W}_{in} and the reservoir \mathbf{A} for block-diagonal Erdős–Rényi networks.

In order to make the figures easier to visualize, we calculate the fraction of connection and use the root of it as the x axis,

$$\sqrt{\frac{b}{d}}. \tag{21}$$

Thus, the higher the fraction, the bigger the blocks and the number of connections in the network. This is necessary because the number of divisors for each network size d is different and the divisors are not equally spaced.

A. Blocks of Erdős–Rényi networks

As mentioned before, we distinguish between two cases for the Erdős–Rényi blocks. First, where all the blocks are individual networks and second, where all blocks are equal.

- **Individual blocks:** We find that for both the Lorenz and the Halvorsen systems, all prediction measures are close to the respective benchmark values in Table I and even surpass them for a network size of 600. Generally, we observe that small block-sizes have a worse long-term prediction quality with regard to the correlation dimension and the largest Lyapunov exponent. However, they appear to have a better short-term forecast horizon. The standard deviation over the variation of random seeds is comparable to the benchmarks. The results are illustrated in Fig. 3. Furthermore, we find the variation for different input weights and starting points to be similarly robust.
- **Equal blocks:** Similar results can be observed for the equal blocks (Fig. 4)—however, the standard deviation is slightly lower. This can be explained by the reduced level in randomness since only one block is randomly constructed.

Generally, we find that the prediction performances stabilize for $\sqrt{\frac{b}{d}} > 0.3$ for both individual and equal blocks. As explained in Eq. (9), this speeds up the calculation of the reservoir spectral radius by a factor of ≈ 123 for individual blocks and ≈ 412 for equal blocks.

Furthermore, we observe that the modified architecture outperforms traditional RC with regard to long-term predictions while it performs slightly worse for short-term predictions. This can be inferred by looking at the correlation dimensions and largest Lyapunov exponents, which are higher than the respective values of

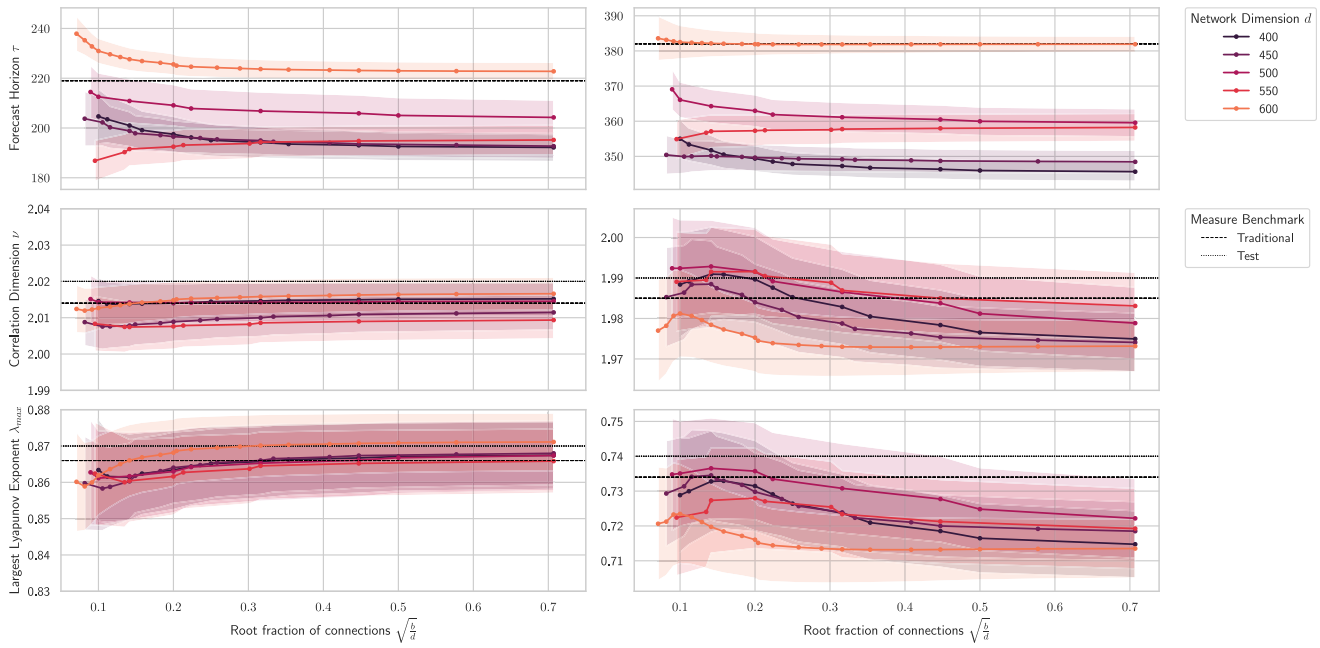


FIG. 4. Prediction measures for the Lorenz (left column) and Halvorsen systems (right column) for different random seeds of equal Erdős–Rényi blocks. The setup of this figure is similar to Fig. 3.

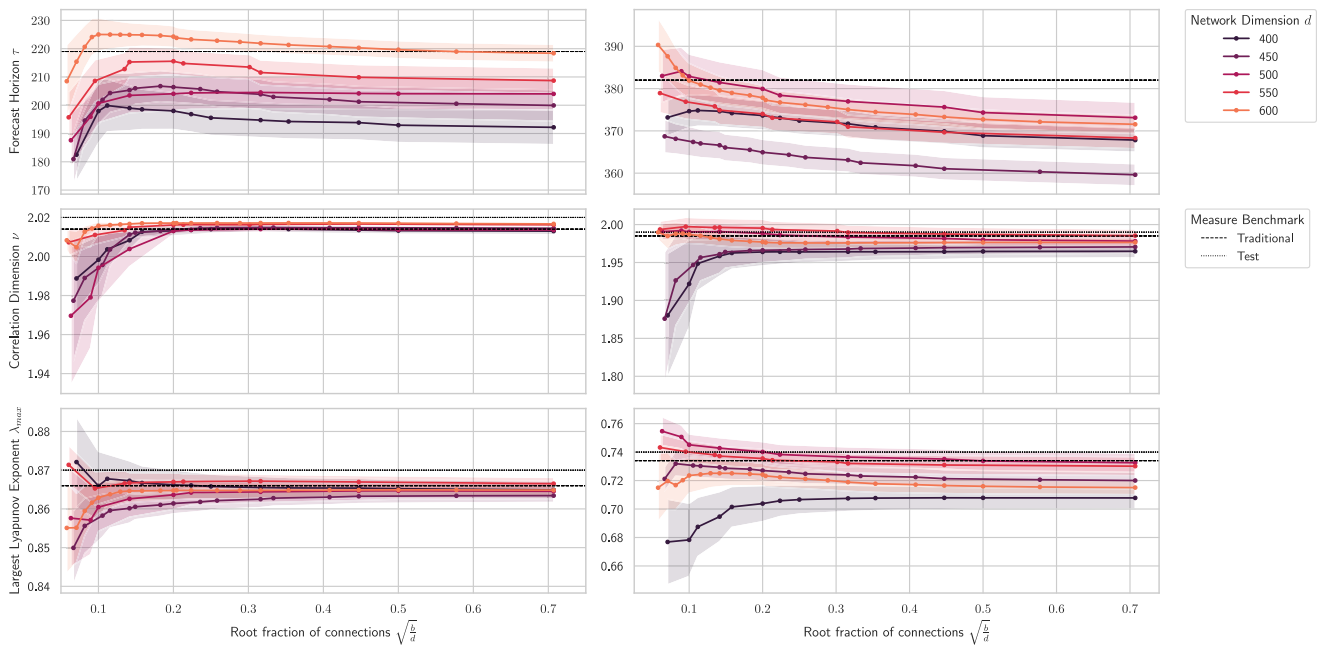


FIG. 5. Prediction measures for the Lorenz (left column) and Halvorsen systems (right column) for different input weights. The setup of this figure is similar to Fig. 3.

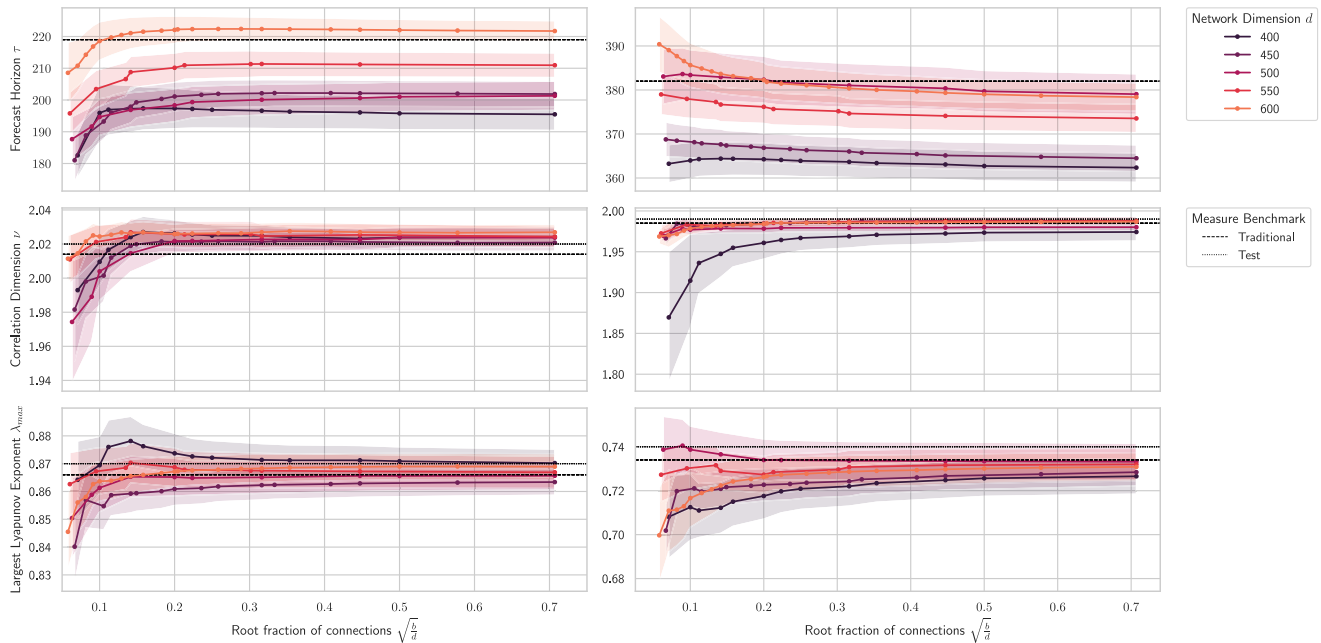


FIG. 6. Prediction measures for the Lorenz (left column) and Halvorsen systems (right column) for different attractor starting points. The setup of this figure is similar to Fig. 3.

the traditional RC architecture and closer to the “true” values of the test data. A comparable short-term prediction performance can be achieved by increasing the network dimension to $d = 600$. There we see that the forecast horizon is able to match the value of the traditional architecture and can even slightly outperform it for small block-sizes of $\sqrt{\frac{b}{a}} \approx 0.05$. In this specific case, the calculation speed of the spectral radius is increased by a factor of 160 000 and 3 200 000 for individual and equal blocks, respectively.

This behavior holds true for both the Lorenz and Halvorsen systems.

B. Blocks of matrices of ones

For the blocks of matrices of ones, the randomness of the reservoirs is taken out. Thus, we focus on the remaining randomness, which is present in the input weights \mathbf{W}_{in} , and the variation of the attractor starting point.

- **Input weights \mathbf{W}_{in} :** We find that for both the Lorenz and the Halvorsen systems, all prediction measures are close to the respective benchmark values and even surpass them for some network sizes. Generally, we observe that the performance is slightly worse than using blocks of Erdős–Rényi networks and the benchmarks. As expected, the standard deviation over the variation of input weights is comparable to the benchmarks and lower than for randomly constructed reservoirs. The results are illustrated in Fig. 5.

- **Attractor starting points:** A similar behavior can be observed for the variation in attractor starting points. However, we observe that network dimensions $d = 600$ even outperform the benchmarks for some block-sizes. The results are illustrated in Fig. 6.

In general, we find that the prediction quality for using blocks of ones as the reservoir is stable for a variation in input weights and attractor starting points. For certain block-sizes, the modified architecture is able to surpass the benchmarks for both short- and long-term predictions. Finding the best performing instance of these reservoirs can be done in a few iterations by varying the block-size for a large enough network dimension. Since the computationally expensive task of calculating the spectral radius of the reservoir is not necessary in this setup, fine-tuning the RC architecture with a parameter scan is fast and scalable.

IV. CONCLUSION AND OUTLOOK

In this paper, we introduce an alternative approach to constructing reservoir computers by replacing the reservoir, which is traditionally a single random network, with a block-diagonal matrix. This implies that the reservoir can be composed of multiple smaller reservoirs, which breaks with the common understanding of the reservoir as a single network.

Furthermore, we remove the randomness of the reservoir by using matrices of ones for the individual blocks.

We evaluate the short- and long-term prediction performance of block-diagonal reservoirs for two nonlinear chaotic systems: the Lorenz and Halvorsen systems. For that, we use three measures:

forecast horizon, correlation dimension, and the largest Lyapunov exponent. We find that—overall—the quality of the predictions is comparable to classical random networks. Although the block-diagonal reservoirs tend to perform slightly worse than the traditional architecture on average, some block-diagonal reservoirs with appropriate size of the blocks perform as well and sometimes even better than the conventional network reservoirs.

We find the result to be robust over variations in network dimensions, block-sizes, target spectral radii, attractor starting points, input weights, and random seeds.

This modified reservoir architecture not only has immediate large benefits regarding the computational effort but also the great potential for simple and fast hardware implementations of reservoir computers becomes obvious. Following this line of research is subject to further research.

We discover many interesting lines of future research. Further directions we find promising to take a look at are: understanding whether block-diagonal reservoirs improve generalization, enhance robustness, or increase scalability.

Current and future work is dedicated to the investigation of these questions—not the least because the answers to them will shed new light on the complexity of the underlying dynamical system.

ACKNOWLEDGMENTS

We would like to acknowledge the DLR for providing code and computational resources.

AUTHOR DECLARATIONS

Conflict of Interest

The authors have no conflicts to disclose.

Author Contributions

Haochun Ma: Conceptualization (equal); Formal analysis (equal); Investigation (equal); Methodology (equal); Software (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Davide Prosperino:** Formal analysis (equal); Software (equal); Writing – original draft (equal); Writing – review & editing (equal). **Alexander Haluszczynski:** Formal analysis (equal); Validation (equal); Visualization (equal); Writing – original draft (equal); Writing – review & editing (equal). **Christoph R ath:** Conceptualization (equal); Supervision (equal); Writing – original draft (equal); Writing – review & editing (equal).

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

- ¹S. L. Brunton and J. N. Kutz, *Data-Driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2022).
- ²A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, and A. A. Bharath, “Generative adversarial networks: An overview,” *IEEE Signal Process. Mag.* **35**, 53–65 (2018).

- ³J. Zhang, Y. Wang, P. Molino, L. Li, and D. S. Ebert, “Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models,” *IEEE Trans. Visualiz. Comp. Graphics* **25**, 364–373 (2018).
- ⁴R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, “Explainable machine learning for scientific insights and discoveries,” *IEEE Access* **8**, 42200–42216 (2020).
- ⁵G. Tanaka, T. Yamane, J. B. H eroux, R. Nakane, N. Kanazawa, S. Takeda, H. Numata, D. Nakano, and A. Hirose, “Recent advances in physical reservoir computing: A review,” *Neural Netw.* **115**, 100–123 (2019).
- ⁶D. Prokhorov, “Echo state networks: Appeal and challenges,” in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005* (IEEE, 2005), Vol. 3, pp. 1463–1466.
- ⁷S. Hochreiter, “The vanishing gradient problem during learning recurrent neural nets and problem solutions,” *Int. J. Uncertain. Fuzziness Knowledge-Based Syst.* **6**, 107–116 (1998).
- ⁸A. Haluszczynski, J. Aumeier, J. Herteux, and C. R ath, “Reducing network size and improving prediction stability of reservoir computing,” *Chaos* **30**, 063136 (2020).
- ⁹G. Holzmann, “Reservoir computing: A powerful black-box framework for non-linear audio processing,” in *International Conference on Digital Audio Effects (DAFx)* (Citeseer, 2009).
- ¹⁰T. L. Carroll and L. M. Pecora, “Network structure effects in reservoir computers,” *Chaos* **29**, 083130 (2019).
- ¹¹A. Haluszczynski and C. R ath, “Good and bad predictions: Assessing and improving the replication of chaotic attractors by means of reservoir computing,” *Chaos* **29**, 103143 (2019).
- ¹²W. Maass, T. Natschl ager, and H. Markram, “Real-time computing without stable states: A new framework for neural computation based on perturbations,” *Neural Comput.* **14**, 2531–2560 (2002).
- ¹³H. Jaeger, “The ‘echo state’ approach to analysing and training recurrent neural networks—with an erratum note,” German National Research Center for Information Technology GMD Technical Report, Bonn, Germany, No. 148, p. 13 (2001).
- ¹⁴D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’ networks,” *Nature* **393**, 440–442 (1998).
- ¹⁵R. Albert and A.-L. Barab asi, “Statistical mechanics of complex networks,” *Rev. Mod. Phys.* **74**, 47 (2002).
- ¹⁶A. D. Broido and A. Clauset, “Scale-free networks are rare,” *Nat. Commun.* **10**, 1017 (2019).
- ¹⁷M. Gerlach and E. G. Altmann, “Testing statistical laws in complex systems,” *Phys. Rev. Lett.* **122**, 168301 (2019).
- ¹⁸A. Griffith, A. Pomerance, and D. J. Gauthier, “Forecasting chaotic systems with very low connectivity reservoir computers,” *Chaos* **29**, 123108 (2019).
- ¹⁹J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, and E. Ott, “Using machine learning to replicate chaotic attractors and calculate Lyapunov exponents from data,” *Chaos* **27**, 121102 (2017).
- ²⁰Z. Lu, B. R. Hunt, and E. Ott, “Attractor reconstruction by machine learning,” *Chaos* **28**, 061104 (2018).
- ²¹E. N. Lorenz, “Deterministic nonperiodic flow,” *J. Atmos. Sci.* **20**, 130–141 (1963).
- ²²J. Herteux and C. R ath, “Breaking symmetries of the reservoir equations in echo state networks,” *Chaos* **30**, 123142 (2020).
- ²³S. Vaidyanathan and A. T. Azar, “Adaptive control and synchronization of Halvorsen circulant chaotic systems,” in *Advances in Chaos Theory and Intelligent Control* (Springer, 2016), pp. 225–247.
- ²⁴E. Hairer, S. P. N orsett, and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, 1993).
- ²⁵H. Jaeger and H. Haas, “Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication,” *Science* **304**, 78–80 (2004).
- ²⁶A. Haluszczynski, “Prediction and control of nonlinear dynamical systems using machine learning,” Ph.D. dissertation (Ludwig-Maximilians-Universit at M unchen, 2021).
- ²⁷P. Erd os, A. R enyi *et al.*, “On the evolution of random graphs,” *Publ. Math. Inst. Hung. Acad. Sci.* **5**, 17–60 (1960), https://static.renyi.hu/~p_erdos/1960-10.pdf

- ²⁸A. E. Hoerl and R. W. Kennard, "Ridge regression: Applications to nonorthogonal problems," *Technometrics* **12**, 69–82 (1970).
- ²⁹C. C. Paige, "Bidiagonalization of matrices and solution of linear equations," *SIAM J. Numer. Anal.* **11**, 197–209 (1974).
- ³⁰S. G. Kobourov, "Spring embedders and force directed graph drawing algorithms," [arXiv:1201.3011](https://arxiv.org/abs/1201.3011) (2012).
- ³¹D. Yu, H. Wang, P. Chen, and Z. Wei, "Mixed pooling for convolutional neural networks," in *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24–26, 2014, Proceedings 9* (Springer, 2014), pp. 364–375.
- ³²T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1* (Springer, 2000), pp. 1–15.

- ³³P. Grassberger and I. Procaccia, "Measuring the strangeness of strange attractors," in *The Theory of Chaotic Attractors* (Springer, 2004), pp. 170–189.
- ³⁴B. Mandelbrot, "How long is the coast of Britain? Statistical self-similarity and fractional dimension," *Science* **156**, 636–638 (1967).
- ³⁵P. Grassberger, "Generalized dimensions of strange attractors," *Phys. Lett. A* **97**, 227–230 (1983).
- ³⁶A. Wolf, J. B. Swift, H. L. Swinney, and J. A. Vastano, "Determining Lyapunov exponents from a time series," *Physica D* **16**, 285–317 (1985).
- ³⁷R. Shaw, "Strange attractors, chaotic behavior, and information flow," *Z. Naturforsch. A* **36**, 80–112 (1981).
- ³⁸M. T. Rosenstein, J. J. Collins, and C. J. De Luca, "A practical method for calculating largest Lyapunov exponents from small data sets," *Physica D* **65**, 117–134 (1993).



OPEN

A novel approach to minimal reservoir computing

Haochun Ma¹, Davide Prosperino¹ & Christoph R ath²✉

Reservoir computers are powerful machine learning algorithms for predicting nonlinear systems. Unlike traditional feedforward neural networks, they work on small training data sets, operate with linear optimization, and therefore require minimal computational resources. However, the traditional reservoir computer uses random matrices to define the underlying recurrent neural network and has a large number of hyperparameters that need to be optimized. Recent approaches show that randomness can be taken out by running regressions on a large library of linear and nonlinear combinations constructed from the input data and their time lags and polynomials thereof. However, for high-dimensional and nonlinear data, the number of these combinations explodes. Here, we show that a few simple changes to the traditional reservoir computer architecture further minimizing computational resources lead to significant and robust improvements in short- and long-term predictive performances compared to similar models while requiring minimal sizes of training data sets.

The prediction of complex dynamic systems is a key challenge across various disciplines in science, engineering, and economics¹. While machine learning approaches, like generative adversarial networks, can provide sensible predictions², difficulties with vast data requirements, the large number of hyperparameters, and lack of interpretability limit their usefulness in some scientific applications³. However, it is required to fundamentally understand how, when, and why the models are working to prevent the risk of misinterpreting the results if deeper methodological knowledge is missing⁴.

In the context of complex systems research, reservoir computers (RCs)^{5,6} have emerged for predicting the dynamics of chaotic systems. The core of the model is a fixed reservoir, which is usually constructed randomly^{7–9}. The input data is fed into the nodes of the reservoir and solely the weights of the readout layer, which transform the reservoir response to output variables, are subject to optimization via linear regression. This makes the learning extremely fast and comparatively transparent. However, this approach can be hit-or-miss, and it is hardly possible to know a priori how the topology of the reservoir will affect the performance^{10–12}.

Recent research has emerged on algorithms which do not require randomness. They are built around regressions¹³ on large libraries of linear and nonlinear combinations constructed from the data observations and their time lags, such as next generation reservoir computers (NG-RCs)¹⁴ or sparse identification of nonlinear dynamics (SINDy)¹⁵. These algorithms are built around nonlinear vector autoregression (NVAR)¹⁶ and the mathematical fact that a powerful universal approximator can be constructed by using an RC with a linear activation function^{17,18}.

The model we present in this paper is based on the same mathematical principles — but instead of getting rid of the traditional reservoir architecture altogether, we take an intermediate step and make only a few simple changes: we restructure the input weights so that all coordinate combinations are fed separately into the reservoir. Additionally, we remove the randomness of the reservoir by replacing it with a block-diagonal matrix of blocks of ones. Instead of introducing the nonlinearity in the activation function, we add higher orders of the reservoir states in the readout.

Using the example of synthetic, chaotic systems, and in particular the Lorenz system, we show that these alterations lead to excellent short- and long-term predictions that significantly outperform traditional RC, NG-RC, and SINDy. While prediction performance is often evaluated visually, we use three quantitative measures: the largest Lyapunov exponent, the correlation dimension, and the forecast horizon. We also validate the robustness of our results by using multiple attractor starting points, different training data sizes and discretizations.

¹Department of Physics, Ludwig-Maximilians-Universit at, Schellingstra e 4, 80799 Munich, Germany. ²Deutsches Zentrum f ur Luft- und Raumfahrt (DLR), Institut f ur KI Sicherheit, Wilhelm-Runge-Stra e 10, 89081 Ulm, Germany. ✉email: christoph.raeth@dlr.de

Results

In this work, we show how small changes to the traditional RC architecture can significantly improve its prediction capability of chaotic systems especially for low data requirements. Therefore, similar to Gauthier et al.,¹⁴ we use the minimal data setup for the Lorenz system with a discretization of $dt=0.025$ and $T_{train}=400$ training data points.

The minimal possible architecture would be a spectral radius $\rho^*=0$ and block-size $b=1$, for which our RC reduces to the case described by Gonon and Ortega.¹⁷ Here, we do not have a reservoir and directly feed the input data to the readout and perform a Ridge regression. While we find this parametrization to be capable of reasonable predictions, a few minor alterations increase the performance significantly.

The standard RC architecture used in this work has block-size $b=3$, spectral radius $\rho^*=0.1$, and a nonlinearity degree $\eta=2$. This equals 36 variables per coordinate. The results of this setup are illustrated in Fig. 1.

In order to obtain robust results we repeat the analysis for 1000 different starting points on the attractor and compare the prediction performance to the other models. In Fig. 2 we see that the novel RC architecture

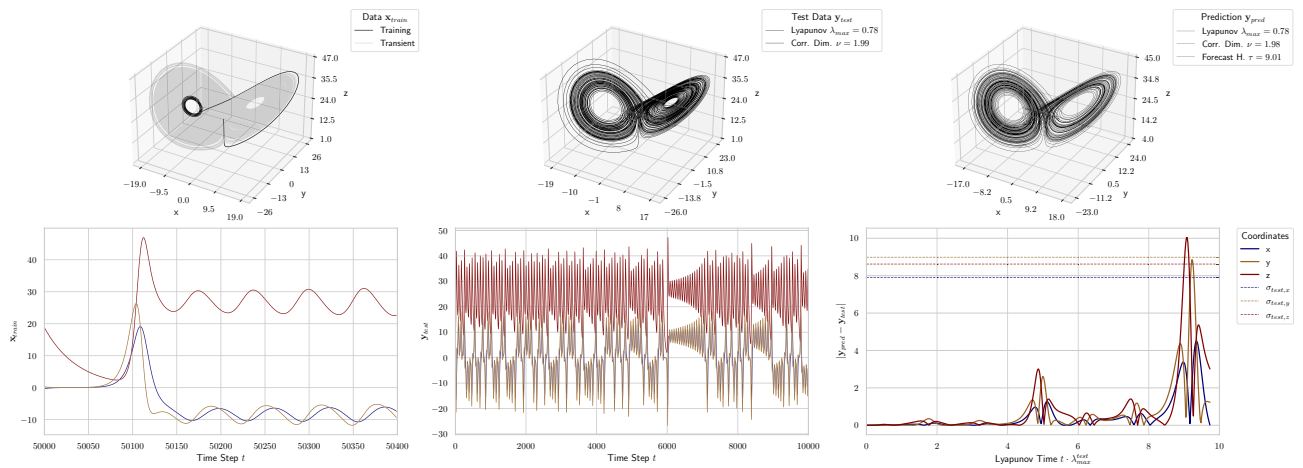


Figure 1. Prediction on a minimal training data set of the Lorenz system. The first column shows the attractor (top) and the trajectories (bottom) of the 400 training data points (and the discarded transient). The second column shows the attractor and the trajectories of the test data. The third column shows the attractor (top) and the absolute prediction error (bottom) of the prediction. The dashed lines indicate the standard deviations of the three components of the test data.

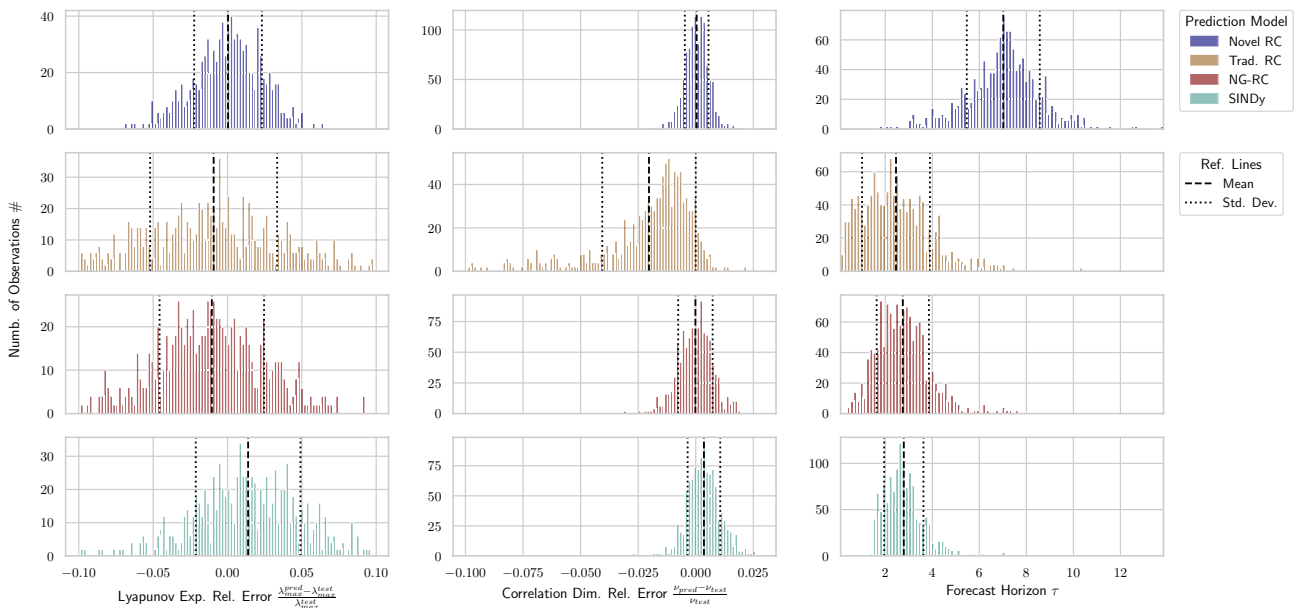


Figure 2. Prediction measures (columns) of different models (rows) for 1000 different starting points on the Lorenz attractor ($dt=0.025, T_{train}=400$). For the correlation dimension and the Lyapunov exponent we calculate the relative error to the respective test data. The mean and standard deviation of each distribution is denoted by a dashed and dotted black line, respectively.

significantly outperforms them with regards to short-term predictions with an average forecast horizon of ~ 7.0 Lyapunov times — this is ~ 2.5 times more than the averages of the other models. The long-term prediction is also slightly better as the average relative errors of the correlation dimension and the Lyapunov exponent are $\sim 3.5 \cdot 10^{-4}$, respectively — this is ~ 9.0 and ~ 39.7 times smaller than the averages of other models. The traditional RC has generally more widely distributed errors due to its randomness.

We verify the robustness of our novel RC to variations in discretization and length of training data. In Fig. 3 we observe that it is quite robust and as expected, performs significantly better than comparable models especially with regards to short-term prediction. Here, we only see a decline in prediction performance for coarse discretizations $dt > 0.045$. The robustness of the long-term prediction is similar to traditional RC and SINDy. Interestingly, we see a decline in performance of NG-RC for larger training lengths $T_{train} > 700$ and finer discretizations $dt < 0.02$. Furthermore, we find our model to be reasonably robust to changes in hyperparameters and noise up to a signal-to-noise ratio of ~ 38 dB.

Furthermore, we analyze the prediction performance of our model on different chaotic systems, which have different nonlinear behavior. We choose the models so that we can understand the inner workings of our RC better. For example, the Halvorsen system has only quadratic nonlinearities with no interacting coordinates and hence the input matrix only needs the first three blocks (which represent the distinct coordinates). Another example to point out is the Rabinovich-Fabrikant system, which has cubic nonlinearities. Here, we see that a nonlinearity degree of $\eta \geq 3$ is necessary for a reasonable prediction. The model parameters and the prediction measures for the different systems are illustrated in Table 1.

Discussion

In this work, we present a novel RC architecture that outperforms comparable methods in terms of short- and long-term predictions while requiring similarly minimal training datasets and computational power. The architecture is modified by restructuring the input weights and reservoir such that combinations of input data coordinates are fed separately into the reservoir. Therefore, we use a block-diagonal matrix of ones as the reservoir, which acts as an averaging operator for the reservoir states at each update step. Similar to average pooling layers in other machine learning methods, this can be interpreted as a way to primarily “extract” features that are more robust¹⁹. It also takes out the randomness of traditional RC. Instead of using a nonlinear activation function to create the reservoir states, we capture the nonlinearity of the data in the readout layer by appending higher orders of the reservoir states before the Ridge regression. We find that these changes lead to a significant improvement in the short- and long-term predictions of chaotic systems in comparison to models such as the traditional RC, NG-RC, and SINDy. In order to evaluate the prediction performance, we compute the largest Lyapunov exponent, the correlation dimension, the correlation dimension, and the prediction horizon.

This work can be extended in many directions. For example, the generation of the reservoir states can be explored to understand what the RC actually learns. In our modified architecture, the states are constructed

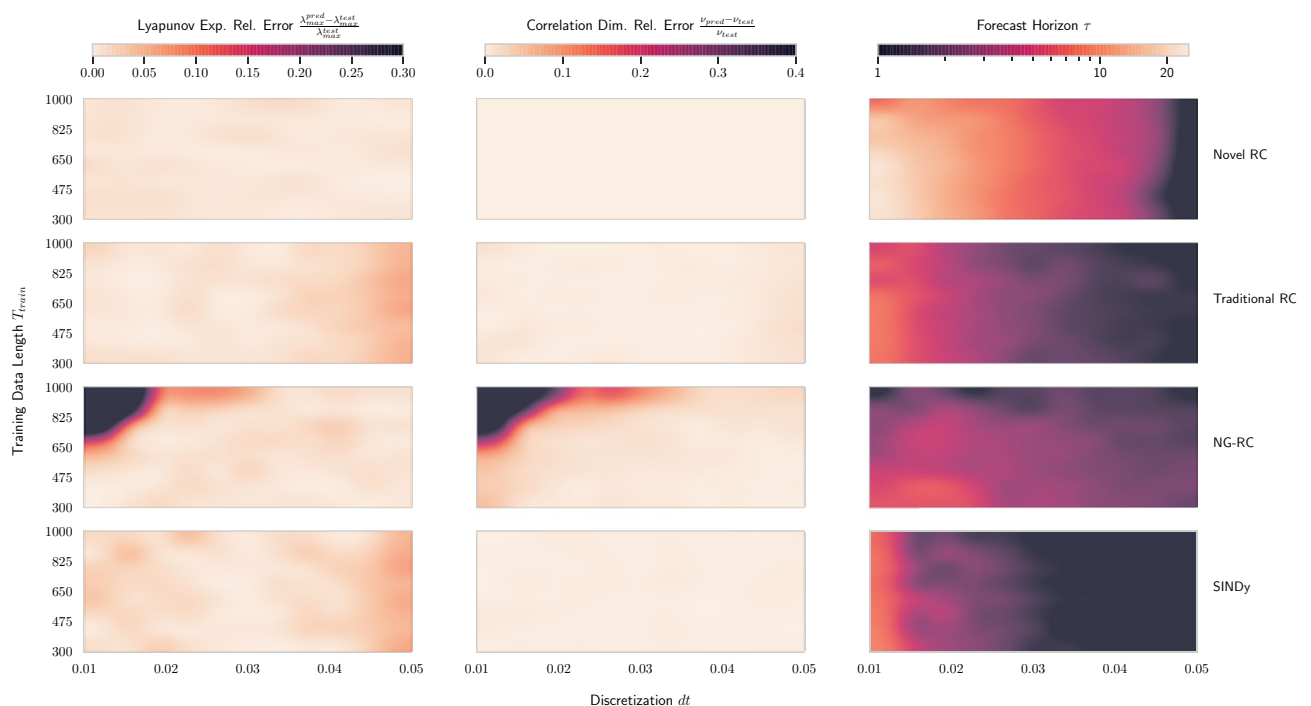


Figure 3. Prediction measures (columns) of different models (rows) for different discretizations and lengths of training data of the Lorenz system. We vary the discretization (x-axis) and the length (y-axis) of the training data between (0.01, 0.05) and (300, 1000), respectively. For the correlation dimension and the Lyapunov exponent we calculate the relative error to the respective test data. Note that the forecast horizon has a logarithmic color scale. Each value in the heatmaps is the average over 100 variations of attractor starting points.

System	Training Data		Novel Architecture			Forecast Horizon τ			
	T_{train}	dt	b	ρ^*	η	Novel	Trad.	NG-RC	SINDy
Halvorsen	300	0.01	3	0.1	2	498±34	231±47	249±27	335±37
Rabi.-Fabr.	300	0.01	3	0.1	3	261±23	168±36	89±12	107±11
Aizawa ⁽⁴³⁾	300	0.01	3	0.1	4	193±16	131±27	76±9	65±7
Dadras-Momeni ⁽⁴⁴⁾	300	0.01	3	0.1	2	423±25	228±41	259±19	248±21
Rössler ⁽⁴⁵⁾	300	0.01	3	0.1	2	781±51	301±72	332±40	401±55
Four wing ⁽⁴⁶⁾	300	0.01	3	0.1	2	1497±39	1135±68	659±28	712±31
Chen ⁽⁴⁷⁾	300	0.01	3	0.1	2	922±41	880±72	750±36	812±41

Table 1. Minimal setup for different chaotic systems. We vary the parameters of the training data and the RC architecture to find the minimal setup for different chaotic systems. To do this, we compute the relative errors of the Lyapunov exponent and the correlation dimension for 100 different attractor starting points. The minimum setup is defined as the setup where the average relative errors of the Lyapunov exponent and the correlation dimension are both $< 10^{-2}$. This ensures that the long-term climate of the chaotic system is reliably reproduced. In this table, we denote the parameters of the data setup (columns 1–2) and RC architecture (columns 3–5). The last 3 columns denote the mean and standard deviations of the forecast horizon for the different prediction models. The governing equations can be found in the respective references.

by mixing the average of the past data with the new data with different “proportions”. Therefore, methods for constructing the reservoir states, such as the exponentially weighted moving average (EWMA) of the data²⁰, should be explored. Related to this, the design of the readout is also an interesting topic to look into. Similarly to NG-RC and SINDy, nonlinear functions could be applied and appended to the reservoir states in order to capture more complex structures in the data.

Another study can be conducted on how the elimination of randomness from RC-like models affects their capabilities, e.g., information processing capacity²¹ or multifunctionality²².

Furthermore, the applicability to high-dimensional and highly nonlinear data can be analyzed and compared with models relying on large feature libraries, such as NG-RC and SINDy. Since the number of variables scales less rapidly in our architecture, it would be relevant to see how much computational power can be saved, especially for hardware RCs.

Moreover, the model can be tested on real-world examples from different disciplines to produce reliable short- and long-term predictions, especially in cases where training data is scarce and expensive.

Methods

Reservoir computers. A reservoir computer (RC)^{5,23,24} is an artificial recurrent neural network (RNN) that relies on a static network called *reservoir*. The term static means that, unlike other RNN approaches, the reservoir remains fixed once the network is constructed. The same is true for the input weights. Therefore, the RC is computationally very efficient since the training process only involves optimizing the output layer. As a result, fast training and high model dimensionality are computationally feasible, making RC well suited for complex real-world applications.

In the following we describe the individual components of the architecture and the modifications that we propose. To make the following section more understandable we introduce them in a high-level summary:

1. *Input weights:* the input weights \mathbf{W}_{in} are designed so that each combination of the coordinates of the data is fed into the reservoir separately.
2. *Reservoir:* the reservoir \mathbf{A} is chosen as a block-diagonal matrix consisting of matrices of ones with size b .
3. *Reservoir states:* we do not use a nonlinear activation function in order to construct the reservoir states $\mathbf{r}(t)$. Hence the iterative update equation reduces to:

$$\mathbf{r}(t + 1) = \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t), \tag{1}$$

where $\mathbf{u}(t)$ denotes the training data at time t .

4. *Readout:* instead of only inserting the squared reservoir states²⁵, our generalized states $\tilde{\mathbf{r}}$ contain all orders up to a nonlinearity degree η :

$$\tilde{\mathbf{r}} = \{\mathbf{r}, \mathbf{r}^2, \dots, \mathbf{r}^{\eta-1}, \mathbf{r}^\eta\}. \tag{2}$$

5. *Training and Prediction:* as in traditional RCs, we stack the training data \mathbf{u} and the corresponding reservoir states $\tilde{\mathbf{r}}$ to matrices \mathbf{U} and $\tilde{\mathbf{R}}$ respectively. We then solve the equation $\mathbf{W}_{out} \cdot \tilde{\mathbf{R}} = \mathbf{U}$ by using Ridge regression²⁶ resulting in:

$$\mathbf{W}_{out} = \mathbf{U} \cdot \tilde{\mathbf{R}}^T (\tilde{\mathbf{R}} \cdot \tilde{\mathbf{R}}^T + \beta \mathbf{I})^{-1}, \tag{3}$$

where $\beta=10^{-5}$ is the regularization constant that prevents overfitting and \mathbf{I} denotes the identity matrix. The prediction procedure of the reservoir states also stays the same (with the adjusted update equation):

$$\mathbf{r}(t + 1) = \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{W}_{out} \cdot \tilde{\mathbf{r}}(t). \tag{4}$$

Note that the reservoir \mathbf{A} only acts on the “simple” reservoir state \mathbf{r} , while the second summand acting on \mathbf{W}_{out} is $\tilde{\mathbf{r}}$ containing all the nonlinear powers. The predicted time series $\mathbf{y}(t)$ can then be obtained by the multiplication:

$$\mathbf{y}(t) = \mathbf{W}_{out} \cdot \tilde{\mathbf{r}}(t). \tag{5}$$

Input weights. In order to feed the input data $\mathbf{u}(t)$ into the reservoir, an input weights matrix \mathbf{W}_{in} is defined, which determines how strongly each coordinate influences every single node of the reservoir network. In a traditional RC, the elements of \mathbf{W}_{in} are chosen to be uniformly distributed random numbers within the interval $[-1, 1]$.

In our novel framework we do not choose the elements randomly, but follow a structured approach. Firstly, in order to remove the randomness, for a block-size of b we take b equally spaced values between $[1, 0]$.

$$\mathbf{w} = (w_1, w_2, \dots, w_b)^T = \left(1, \frac{b-2}{b-1}, \dots, \frac{1}{b-1}, 0\right)^T \tag{6}$$

To avoid non-invertible matrices for ridge regression, we take the square root values of all weights $\mathbf{w} = (\sqrt{w_1}, \dots, \sqrt{w_b})^T$. Then we specifically structure the input matrix so that the different combinations of input data coordinates, also called *features*, are fed separately into the reservoir. In the case of a 3-dimensional system with coordinates $\mathbf{u}(t) = (x, y, z)^T(t)$ and a nonlinearity order $\eta=2$, the input matrix (multiplication) looks like:

$$\mathbf{W}_{in} \cdot \mathbf{u}(t) = \begin{pmatrix} \mathbf{w} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{w} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{w} \\ \mathbf{w} & \mathbf{w} & \mathbf{0} \\ \mathbf{w} & \mathbf{0} & \mathbf{w} \\ \mathbf{0} & \mathbf{w} & \mathbf{w} \end{pmatrix} \cdot \mathbf{u}(t) = \begin{pmatrix} x \\ y \\ z \\ x+y \\ x+z \\ y+z \end{pmatrix} (t) \otimes \mathbf{w} \tag{7}$$

where \otimes denotes the tensor product and hence each block represents one feature f . For n -dimensional data the feature space contains $n_f = 2^n - 2$ elements. Thus, the dimension of the reservoir is $d = n_f \cdot b$.

Reservoir. The core of an RC, the reservoir \mathbf{A} , is usually constructed as a sparse Erdős-Rényi random network²⁷ with number of nodes d . As for the choice of input weights, we choose the reservoir in such a way that each feature remains separate. Therefore, we use a block diagonal matrix \mathbf{J} consisting of ones \mathbf{J} with block size b . Thus, each block \mathbf{J}_i can be directly mapped to a particular feature:

$$\mathbf{J} = \begin{pmatrix} \mathbf{J}_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{n_f} \end{pmatrix} \mapsto \begin{pmatrix} \mathbf{J}_x & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{J}_y & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{J}_{y+z} \end{pmatrix}. \tag{8}$$

Similar to a traditional RC, we scale the spectral radius $\rho(\mathbf{J})$ of the reservoir to a target spectral radius ρ^* . While the computation of the spectral radius is usually a computationally expensive task²⁸ that scales with $\mathcal{O}(d^3)$, the computation is no longer necessary for block diagonal matrices of ones \mathbf{J} . This is because the eigenvalues of the matrix are equal to the block size b . Thus, the rescaled reservoir is given by:

$$\mathbf{A} \equiv \frac{\rho^*}{\rho(\mathbf{J})} \mathbf{J} \longrightarrow \frac{\rho^*}{b} \mathbf{J}. \tag{9}$$

Our default target spectral radius is $\rho^*=0.1$.

Reservoir states. As in traditional RC, we use a recurrent update equation to capture the dynamics of the system in the so-called reservoir states $\mathbf{r}(t)$. This would normally require a bounded nonlinear activation function $g(\cdot)$ that captures the nonlinear properties of the data. The activation function (usually the hyperbolic tangent) is applied on an element-by-element basis.

However, as mentioned earlier, we shift the nonlinearity entirely to the readout. Therefore, the time evolution of the states is iteratively determined via:

$$\mathbf{r}(t + 1) = g(\mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t)) \tag{10}$$

$$\longrightarrow \mathbf{A} \cdot \mathbf{r}(t) + \mathbf{W}_{in} \cdot \mathbf{u}(t). \tag{11}$$

Due to our choice of architecture, the reservoir states for each feature can be obtained separately:

$$\mathbf{r}(t) = \begin{pmatrix} \mathbf{r}_x \\ \mathbf{r}_y \\ \mathbf{r}_z \\ \mathbf{r}_{x+y} \\ \mathbf{r}_{x+z} \\ \mathbf{r}_{y+z} \end{pmatrix} (t) = \begin{pmatrix} \mathbf{r}_{x,1} \\ \mathbf{r}_{x,2} \\ \vdots \\ \mathbf{r}_{x,b} \\ \mathbf{r}_{y,1} \\ \vdots \\ \mathbf{r}_{y+z,b} \end{pmatrix} (t). \tag{12}$$

Hence, we can take all reservoir states belonging to one feature $\mathbf{r}_f(t)$ — we call them *feature states* — and analyze them separately. This helps us to understand that the reservoir acts as an averaging operator on the feature states:

$$\mathbf{A}_f \cdot \mathbf{r}_f(t) = \left(\frac{\rho^*}{b} \sum_{i=1}^b \mathbf{r}_{f,i}(t) \right) \cdot \mathbf{I}_b = \rho^* \cdot \bar{\mathbf{r}}_f(t), \tag{13}$$

where \mathbf{I}_b is a vector of ones of size b . Thus, in each iteration step, the feature states are “normalized” to the average of the past feature states $\bar{\mathbf{r}}_f(t)$ and a varying strength (determined by the input weight) is added to the new feature data $f(t)$:

$$\mathbf{r}_f(t + 1) = \rho^* \cdot \bar{\mathbf{r}}_f(t) + \mathbf{w} \cdot f(t). \tag{14}$$

where f can be replaced by any other feature without loss of validity. The average, or “memory”, of each feature is tracked in the last row of the feature states since $w_b=0$. Furthermore, this implies that target spectral radius ρ^* determines how strongly the memory of the data is kept in each iteration step.

Readout. While a quadratic readout, i.e., the squared reservoir states \mathbf{r}^2 , is often added to a traditional RC to break the symmetry of the activation function²⁵, we need the readout to capture the nonlinearity of the data. Therefore, we add even higher orders of nonlinearity to the so-called generalized states $\tilde{\mathbf{r}}$. For a given degree of nonlinearity η they look like the following:

$$\tilde{\mathbf{r}} = \{ \mathbf{r}, \mathbf{r}^2, \dots, \mathbf{r}^{\eta-1}, \mathbf{r}^\eta \}. \tag{15}$$

Hence, for a degree of nonlinearity η and a block-size of b , the number of elements in the readout, which is also the number of variables to be optimized, is:

$$n_{out} = (2^\eta - 2) \cdot \eta \cdot b = \left(\sum_{k=1}^{\eta} \binom{\eta}{k} - 1 \right) \cdot \eta \cdot b, \tag{16}$$

which we rewrite to binomial coefficients for better comparison. For high-dimensional data with high nonlinearity, the number of variables to be optimized is much smaller than for comparable predictive models such as NG-RC¹⁴ or SINDy¹⁵. This is because NG-RC and SINDy require combinations with recurrences. Therefore, the size of their feature space for a nonlinearity degree is η (at least):

$$n_f = \sum_{k=1}^{\eta} \binom{\eta+k-1}{k} = \binom{\eta+\eta}{\eta} - 1, \tag{17}$$

which grows much faster for larger n and η than the expression for n_{out} in Eq. 16.

Prediction performance measures. When forecasting nonlinear dynamical systems, the goal is not only to exactly replicate the short-time trajectory, but also to reproduce the long-term statistical properties, or climate, of the system.

Correlation dimension. To assess the structural complexity of an attractor, we calculate its correlation dimension ν , which measures the fractal dimensionality of the space populated by its trajectory^{29,30}. The correlation dimension is implicitly defined by the power-law relationship based on the correlation integral:

$$C(r) = \int_0^r d^n r' c(\mathbf{r}') \longrightarrow C(r) \propto r^\nu \tag{18}$$

where n is the dimension of the data and $c(\mathbf{r}')$ is the standard correlation function. The integral represents the mean probability that two states in the phase space are close to each other at different time steps. This is the case if the distance between the two states is smaller than the threshold distance r . For self-similar, strange attractors, this power-law relationship holds for a certain range of r , which can be calibrated using the *Grassberger-Procaccia* algorithm³¹. The benefits of this measure are that it is purely data-based, it only needs a small number of data points, and it does not require any knowledge of the underlying governing equations of the system.

Lyapunov exponents. Besides its fractal dimensionality, the statistical climate of an attractor is also characterized by its temporal complexity represented by the *Lyapunov* exponents³². They describe the average rate of

divergence of nearby points in the phase space, and thus measure sensitivity with respect to initial conditions³³. There is one exponent for each dimension in the phase space. If the system has at least one positive Lyapunov exponent, it is classified as chaotic. Thus, it is sufficient for the purposes in this work to calculate only the largest Lyapunov exponent λ_{max} :

$$d(t) = C \cdot e^{\lambda_{max} \cdot t} \quad (19)$$

where $d(t)$ denotes the distance of two initially nearby states in phase space and the constant C is the normalization constant at the initial separation. Thus, instead of determining the full Lyapunov spectrum, we only need to find the largest one as it describes the overall system behavior to a large extent. Therefore we use the *Rosenstein* algorithm³⁴.

Forecast horizon. To quantify the quality and duration of the short-term prediction of the trajectory we use the forecast horizon τ ¹². It tracks the number of time steps for which the absolute error between each coordinate of the predicted $y_{pred}(t)$ and test $y_{test}(t)$ data does not exceed the standard deviation of the test data $\sigma(y_{test}(t))$:

$$|y_{pred}(t) - y_{test}(t)| < \sigma(y_{test}(t)). \quad (20)$$

We express the forecast horizon in units of the Lyapunov time by multiplying it with the discretization and maximum Lyapunov exponent of the test data $\tau \cdot dt \cdot \lambda_{max}^{test}$. This measure is intended to evaluate how long a prediction can reproduce the actual trajectory before the chaotic nature of the system leads to an exponential divergence.

Dynamical systems. We apply our model to a number of synthetic chaotic systems. In our analyses, we focus on the following three due to their specific properties in terms of nonlinearity.

Lorenz. As it is common in RC research^{35,36} we use the Lorenz system which was initially proposed for modeling atmospheric convection³⁷:

$$\begin{aligned} \frac{dx}{dt} &= \sigma \cdot (y - x) \\ \frac{dy}{dt} &= x \cdot (\rho - z) - y \\ \frac{dz}{dt} &= x \cdot y - \beta \cdot z, \end{aligned} \quad (21)$$

where the standard parametrization for chaotic behavior is $\sigma=10$, $\rho=28$, and $\beta=8/3$.

Halvorsen. As in Hertreux and R ath²⁵ we use the Halvorsen system³⁸ for our analyses, which has a cyclic symmetry and, unlike to the Lorenz system, only has nonlinearities without interaction of coordinates:

$$\begin{aligned} \frac{dx}{dt} &= a \cdot x - b \cdot y - b \cdot z - y^2 \\ \frac{dy}{dt} &= a \cdot y - b \cdot z - b \cdot x - z^2 \\ \frac{dz}{dt} &= a \cdot z - b \cdot x - b \cdot y - x^2, \end{aligned} \quad (22)$$

where $a=1.3$ and $b=4$ are the standard parameters.

Rabinovich–Fabrikant. In order to test whether our model works also for systems entailing cubic nonlinearities, we analyze the Rabinovich–Fabrikant system³⁹:

$$\begin{aligned} \frac{dx}{dt} &= y \cdot (z - 1 + x^2) + \gamma \cdot x \\ \frac{dy}{dt} &= x \cdot (3 \cdot z + 1 - x^2) + \gamma \cdot y \\ \frac{dz}{dt} &= -2 \cdot z \cdot (\alpha + x \cdot y), \end{aligned} \quad (23)$$

where $\alpha=0.14$ and $\gamma=0.1$ are the standard parameters.

Simulating and splitting data. Since we compare our model with NG-RC and SINDy, we use the same data setup as the original works^{14,15}. Hence, we solve the differential equations of the systems using the *Runge-Kutta* method⁴⁰ with a discretization of $dt=0.025$ in order to ensure a sufficient manifestation of the attractor.

We discard the initial transient of $T_{transient}=50000$, use the next $T_{train}=400$ steps for training, then skip $T_{skip}=10000$ steps, and use the remaining $T_{test}=10000$ for testing the prediction. Hence in total we simulate $T=70400$ steps.

To get robust results, we also vary the starting points on the attractor by using the rounded last point of one data sample as the starting point for the next. The initial starting points for the Lorenz, Halvorsen, and Rabinovich-Fabrikant systems are $(-14, -20, 25)$, $(-6.4, 0, 0)$, and $(-0.4, 0.1, 0.7)$, respectively.

Comparable prediction models. We compare our novel RC to other models designed for predicting dynamical systems. We briefly describe them in the following.

Traditional reservoir computer. For the traditional RC architecture we choose an Erdős-Rényi network of dimension $d=600$ with a target spectral radius ρ^* and a quadratic readout. This equals 1200 variables per coordinate to be optimized. In order to get robust results we repeat the prediction for 1000 realizations and take the average of the prediction measures.

Next generation reservoir computer. The next generation reservoir computer (NG-RC) developed by Gauthier et al.¹⁴ is a so-called nonlinear vector autoregression (NVAR) machine and thus, does not require a reservoir. It solely needs the feature vector, which consists of time-delay observations of the data and nonlinear functions of these observations. The resulting output weights can be used to construct the governing equations of the data. We use the standard setting with time delays $k=2$ and skips $s=1$.

Sparse identification of nonlinear dynamics. Sparse identification of nonlinear dynamics (SINDy)¹⁵ discovers the underlying dynamical system of data by learning its governing equations through sparse regression. It is similar to NG-RC, but uses an iterative approach to filter only relevant features. We use the standard parametrization and the official Python package PySINDy^{41,42}.

Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on reasonable request.

Received: 1 May 2023; Accepted: 1 August 2023

Published online: 10 August 2023

References

1. S. L. Brunton and J. N. Kutz, *Data-driven Science and Engineering: Machine Learning, Dynamical Systems, and Control* (Cambridge University Press, 2022)
2. Creswell, A. et al. Generative adversarial networks: An overview. *IEEE Signal Process. Mag.* **35**, 53 (2018).
3. Zhang, J., Wang, Y., Molino, P., Li, L. & Ebert, D. S. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE Trans. Vis. Comput. Graph.* **25**, 364 (2018).
4. Roscher, R., Bohn, B., Duarte, M. F. & Garcke, J. Explainable machine learning for scientific insights and discoveries. *IEEE Access* **8**, 42200 (2020).
5. Jaeger, H. The “echo state” approach to analysing and training recurrent neural networks—with an erratum note. *Bonn Ger. Ger. Natl. Res. Center Inf. Technol. GMD Tech. Rep.* **148**, 13 (2001).
6. D. Prokhorov, Echo state networks: appeal and challenges, in *Proc. 2005 IEEE International Joint Conf. on Neural Networks, 2005.*, Vol. 3 (IEEE, 2005) pp. 1463–1466
7. Broido, A. D. & Clauset, A. Scale-free networks are rare. *Nat. Commun.* **10**, 1017 (2019).
8. Gerlach, M. & Altmann, E. G. Testing statistical laws in complex systems. *Phys. Rev. Lett.* **122**, 168301 (2019).
9. G. Holzmann, Reservoir computing: a powerful black-box framework for nonlinear audio processing, in *International Conf. on Digital Audio Effects (DAFx)* (Citeseer, 2009)
10. J. Platt, H. Abarbanel, S. Penny, A. Wong, and R. Clark, Robust forecasting through generalized synchronization in reservoir computing, in *AGU Fall Meeting Abstracts*, Vol. 2021 (2021) pp. NG25B–0522
11. Tanaka, G. et al. Recent advances in physical reservoir computing: A review. *Neural Netw.* **115**, 100 (2019).
12. Haluszczynski, A. & R ath, C. Good and bad predictions: Assessing and improving the replication of chaotic attractors by means of reservoir computing. *Chaos Interdisc. J. Nonlinear Sci.* **29**, 103143 (2019).
13. Bollt, E. On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd a3b2 show [feature]. *Chaos Interdisc. J. Nonlinear Sci.* **31**, 013108 (2021).
14. Gauthier, D. J., Bollt, E., Griffith, A. & Barbosa, W. A. Next generation reservoir computing. *Nat. Commun.* **12**, 5564 (2021).
15. Brunton, S. L., Proctor, J. L. & Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proc. Natl. Acad. Sci.* **113**, 3932 (2016).
16. Weise, C. L. The asymmetric effects of monetary policy: A nonlinear vector autoregression approach. *J. Money Credit Bank.* **85**, 25468 (1999).
17. Gonon, L. & Ortega, J.-P. Reservoir computing universality with stochastic inputs. *IEEE Trans. Neural Netw. Learn. Syst.* **31**, 100 (2019).
18. Hart, A. G., Hook, J. L. & Dawes, J. H. Echo state networks trained by tikhonov least squares are $l_2(\mu)$ approximators of ergodic dynamical systems. *Phys. D Nonlinear Phenom.* **421**, 132882 (2021).
19. D. Yu, H. Wang, P. Chen, and Z. Wei, Mixed pooling for convolutional neural networks, in *Rough Sets and Knowledge Technology: 9th International Conf., RSKT 2014, Shanghai, China, October 24–26, 2014, Proc. 9* (Springer, 2014) pp. 364–375
20. Hunter, J. S. The exponentially weighted moving average. *J. Qual. Technol.* **18**, 203 (1986).
21. Dambre, J., Verstraeten, D., Schrauwen, B. & Massar, S. Information processing capacity of dynamical systems. *Sci. Rep.* **2**, 1 (2012).
22. Flynn, A., Tsachouridis, V. A. & Amann, A. Multifunctionality in a reservoir computer. *Chaos Interdisc. J. Nonlinear Sci.* **31**, 013125 (2021).
23. Maass, W., Natschläger, T. & Markram, H. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Comput.* **14**, 2531 (2002).
24. Jaeger, H. & Haas, H. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science* **304**, 78 (2004).
25. J. Herteux and C. R ath, Breaking symmetries of the reservoir equations in echo state networks. *Chaos Interdisc. J. Nonlinear Sci.* **30**, 123142 (2020)

26. Hoerl, A. E. & Kennard, R. W. Ridge regression: Applications to nonorthogonal problems. *Technometrics* **12**, 69 (1970).
27. Erdos, P. *et al.* On the evolution of random graphs. *Publ. Math. Inst. Hung. Acad. Sci* **5**, 17 (1960).
28. Paige, C. C. Bidiagonalization of matrices and solution of linear equations. *SIAM J. Numer. Anal.* **11**, 197 (1974).
29. P. Grassberger and I. Procaccia, Measuring the strangeness of strange attractors, in *The Theory of Chaotic Attractors* (Springer, 2004) pp. 170–189
30. Mandelbrot, B. How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science* **156**, 636 (1967).
31. Grassberger, P. Generalized dimensions of strange attractors. *Phys. Lett. A* **97**, 227 (1983).
32. Wolf, A., Swift, J. B., Swinney, H. L. & Vastano, J. A. Determining lyapunov exponents from a time series. *Phys. D Nonlinear Phenom.* **16**, 285 (1985).
33. Shaw, R. Strange attractors, chaotic behavior, and information flow. *Z. Naturforschung A* **36**, 80 (1981).
34. Rosenstein, M. T., Collins, J. J. & De Luca, C. J. A practical method for calculating largest lyapunov exponents from small data sets. *Phys. D Nonlinear Phenom.* **65**, 117 (1993).
35. Pathak, J., Lu, Z., Hunt, B. R., Girvan, M. & Ott, E. Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data. *Chaos Interdisc. J. Nonlinear Sci.* **27**, 121102 (2017).
36. Lu, Z., Hunt, B. R. & Ott, E. Attractor reconstruction by machine learning. *Chaos Interdisc. J. Nonlinear Sci.* **28**, 061104 (2018).
37. Lorenz, E. N. Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130 (1963).
38. S. Vaidyanathan and A. T. Azar, Adaptive control and synchronization of halvorsen circulant chaotic systems, In *Advances in chaos theory and intelligent control* (Springer, 2016) pp. 225–247
39. Rabinovich, M. I., Fabrikant, A. L. & Tsimring, L. S. Finite-dimensional spatial disorder. *Soviet Phys. Usp.* **35**, 629 (1992).
40. E. Hairer, S. P. Nørsett, and G. Wanner, *Solving Ordinary Differential Equations I* (Springer, 1993) <https://doi.org/10.1007/978-3-540-78862-1>
41. B. de Silva, K. Champion, M. Quade, J.-C. Loiseau, J. Kutz, and S. Brunton, Pysindy: A python package for the sparse identification of nonlinear dynamical systems from data. *J. Open Source Softw.* **5**, 2104 (2020) <https://doi.org/10.21105/joss.02104>
42. A. A. Kaptanoglu, B. M. de Silva, U. Fasel, K. Kaheman, A. J. Goldschmidt, J. Callaham, C. B. Delahunt, Z. G. Nicolaou, K. Champion, J.-C. Loiseau, J. N. Kutz, and S. L. Brunton, Pysindy: A comprehensive python package for robust sparse system identification. *J. Open Source Softw.* **7**, 3994 (2022) <https://doi.org/10.21105/joss.03994>
43. Aizawa, Y. *et al.* Stagnant motions in hamiltonian systems. *Progr. Theor. Phys. Suppl.* **98**, 36 (1989).
44. Dadras, S. & Momeni, H. R. A novel three-dimensional autonomous chaotic system generating two, three and four-scroll attractors. *Phys. Lett. A* **373**, 3637 (2009).
45. Rossler, O. An equation for hyperchaos. *Phys. Lett. A* **71**, 155 (1979).
46. Qi, G., Chen, G., van Wyk, M. A., van Wyk, B. J. & Zhang, Y. A four-wing chaotic attractor generated from a new 3-d quadratic autonomous system. *Chaos Solitons Fractals* **38**, 705 (2008).
47. Chen, G. & Ueta, T. Yet another chaotic attractor. *Int. J. Bifurc. Chaos* **9**, 1465 (1999).

Author contributions

H.M conducted the research and wrote the main manuscript text. C.R. and D.P. assisted in the research. All authors reviewed the manuscript.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to C.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023

Linear and nonlinear causality in financial markets

Haochun Ma,^{1,2} Davide Prosperino,^{1,2} Alexander Haluszczynski,² and Christoph Räth³

¹Ludwig-Maximilians-Universität, Department of Physics, Schellingstraße 4, 80799 Munich, Germany

²Allianz Global Investors, risklab, Seidlstraße 24, 80335, Munich, Germany

³Deutsches Zentrum für Luft- und Raumfahrt (DLR), Institut für KI Sicherheit, Wilhelm-Runge-Straße 10, 89081 Ulm, Germany

(*Electronic mail: christoph.raeth@dlr.de)

Identifying and quantifying co-dependence between financial instruments is a key challenge for researchers and practitioners in the financial industry. Linear measures such as the Pearson correlation are still widely used today, although their limited explanatory power is well known. In this paper we present a much more general framework for assessing co-dependencies by identifying and interpreting linear and nonlinear causalities in the complex system of financial markets. To do so, we use two different causal inference methods, transfer entropy and convergent cross-mapping, and employ Fourier transform surrogates to separate their linear and nonlinear contributions. We find that stock indices in Germany and the U.S. exhibit a significant degree of nonlinear causality and that correlation, while a very good proxy for linear causality, disregards nonlinear effects and hence underestimates causality itself. The presented framework enables the measurement of nonlinear causality, the correlation-causality fallacy, and motivates how causality can be used for inferring market signals, pair trading, and risk management of portfolios. Our results suggest that linear and nonlinear causality can be used as early warning indicators of abnormal market behavior, allowing for better trading strategies and risk management.

Within the complex system of financial markets, understanding the intricate ties between assets is crucial. Although the Pearson correlation has been a standard measure for these relationships, its linear approach might not fully represent the entire spectrum of causality. This study employs sophisticated causal inference algorithms and methods to differentiate between linear and nonlinear causal contributions. By examining major stock indices from Germany and the U.S., we uncover profound and possibly nonlinear linkages. More than presenting a new approach, this research indicates a significant shift in our perception and quantification of financial market behaviors. Such insights hold promise for refining market predictions, optimizing trading strategies, and improving portfolio risk management.

acades in foreign exchange markets. Such insights challenge the adequacy of linear dependency metrics. Addressing this, Haluszczynski *et al.*⁷ segregated linear from nonlinear mutual information contributions using Fourier transform surrogates, aiming to quantify nonlinear correlations among financial assets. The authors demonstrated that the integration of nonlinear correlations into portfolio construction led to an increase in investment performance. A pressing query is the continued reliance on the Pearson correlation⁸ as a causality proxy, given the intricate nature of causality measurement within dynamic systems. Granger's initial study in the 1960s⁹ addressed the difference between causality and correlation, leading to the development of more advanced causal inference tools. This ranged from information-theoretic tools¹⁰ to state-space reconstruction models¹¹. While causal inference has mainly focused on determining causality¹², the study of its linear versus nonlinear characteristics has not been performed in detail. Beginning work has been performed by Paluš and Vejmelka¹³ and Hlinka *et al.*¹⁴, who focused on mutual information to detect nonlinear dynamics in time series and evaluated nonlinearity contributions in climate connectivity.

I. INTRODUCTION

The field of econophysics is garnering heightened attention in the physics domain, offering a novel lens to conventional financial methodologies¹. This emerging perspective draws from statistical physics tools, spanning signal processing, agent-based market frameworks, and random matrix theory². Understanding the co-dependence of financial assets is paramount across various finance sectors, especially when quantifying portfolio-associated risks³. This development has seen industry practitioners keenly monitor the evolution of co-dependence metrics. Predominantly, mutual dependencies of financial instruments are characterized via the Pearson correlation of their return time series. However, there is increasing research underscoring the nonlinear characteristics of these series⁴. Notably, Mantegna and Stanley⁵ showed the power law scaling dynamics of financial indices' probability distributions, while Ghashghaie *et al.*⁶ pinpointed turbulent cas-

In this paper, we analyze causality in financial markets by separating linear and nonlinear contributions to causality using Fourier transform surrogates. To do so, we use two different causal inference techniques and apply them to historical stock data of the German DAX and the U.S. Dow-Jones index. We also identify causality-based statistical properties of financial data and motivate how linear and nonlinear causality can be separated and measured. We find that while correlation is a good proxy for linear causality, nonlinear effects are disregarded and thus, significant amount of nonlinear causality is neglected. This is potentially dangerous when practitioners evaluate the risk of a portfolio only using correlation. Therefore, we propose a simple integration of causality measures into market signal inference, pair trading, and portfolio construction routines and show that they yield superior results.

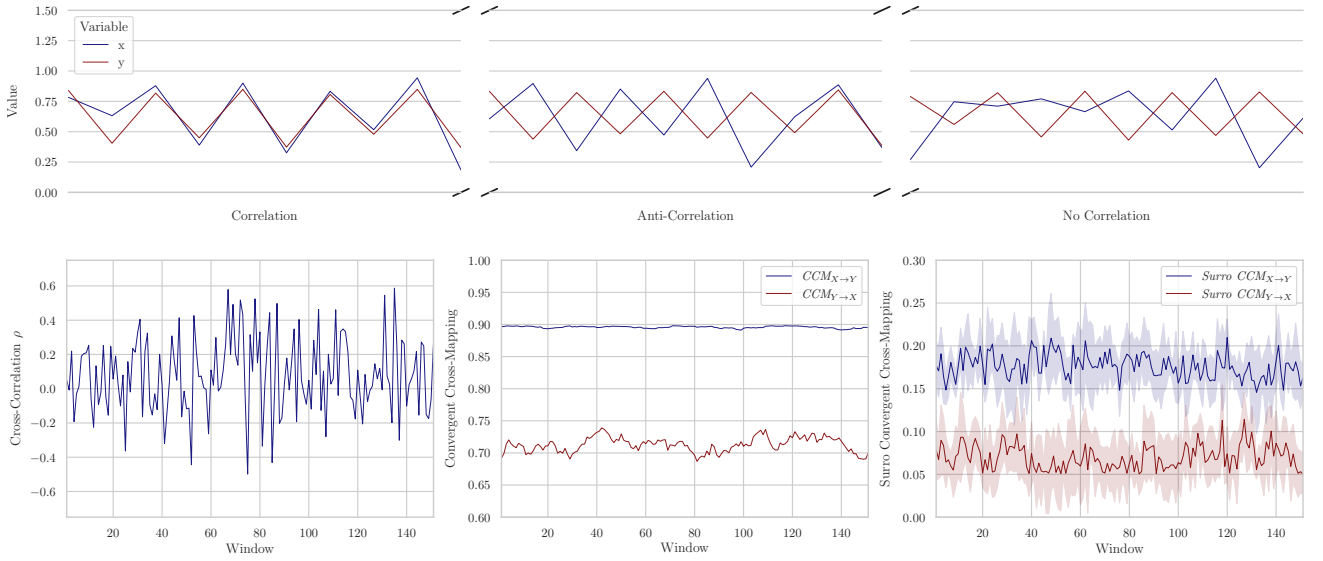


FIG. 1. Mirage correlations and causality. The top row shows different regimes of the coupled difference system defined in Equation 1. It appears that the variables are correlated in the first regime, anti-correlated in the second, and lose all coherence in the third. The bottom row shows the rolling correlation (left), causality (center), and linear causality (right). The causality is measured using Convergent Cross Mapping (CCM). While the correlation alternates between periods of positive, negative, and zero correlation, the causality in both directions stays stable over time. This also holds true for the linear causality. When comparing the measurements to the governing equations, we see that causality offers a more stable and accurate representation of the co-dependence between the two variables than correlation does.

II. METHODS

We structure our methods section in three different parts: data, causality measures, linear and nonlinear decomposition, and financial frameworks.

A. Data

In this section, we describe the data used for this study. Before we apply our framework to real-world data, we demonstrate it on a synthetic example. Additionally, we use rolling windows in order to evaluate our analysis dynamically.

1. Coupled Difference

A simple example of a system that displays chaotic behavior is the coupled difference as introduced in¹⁵. This system was also employed by Sugihara *et al.*¹¹ to illustrate *Convergent Cross Mapping* (CCM), a causality inference method integral to this study. It is defined by the following two equations:

$$\begin{aligned} x(t+1) &= x(t) \cdot [r_x - r_x \cdot x(t) - \beta_{y \rightarrow x} \cdot y(t)] \\ y(t+1) &= y(t) \cdot [r_y - r_y \cdot x(t) - \beta_{x \rightarrow y} \cdot x(t)], \end{aligned} \quad (1)$$

where the standard parameters are: $r_x = 3.8$, $r_y = 3.5$, $\beta_{y \rightarrow x} = 0.02$, and $\beta_{x \rightarrow y} = 0.1$. We selected this system due to its exhibition of so-called *mirage correlations*, which means that

variables may be positively coupled for long periods but can spontaneously become uncorrelated or decoupled. This can lead to problems when fitting models or inferring causality from observational data¹¹.

2. Financial Data

For our real-world analysis, we select a subset of stocks from the DAX and Dow-Jones indices that represent the 30 highest capitalized and thus most influential companies in Germany and the U.S., respectively. Beginning on January 19, 1973, our data consists of the daily closing prices of all stocks that were in the index through April 20, 2022, to provide a consistent universe of stocks over the entire period. This yields a total of $N_{DAX} = 11$ and $N_{DJ} = 17$ time series with 12785 data points. We would like to note that the survival bias¹⁶ is negligible for our analysis.

To ensure stationary time series, we convert the stock prices p_t to logarithmic returns:

$$x_t = \log p_t - \log p_{t-1}.$$

The time horizon of our data is long enough to examine a number of important market events—starting with the global recession of the early 1980s, it also includes Black Monday (October 19, 1987), when stock markets around the world collapsed for the first time since World War II. From 1997 to 2001, markets were characterized by excessive speculation and the overvaluation of many technology companies, which

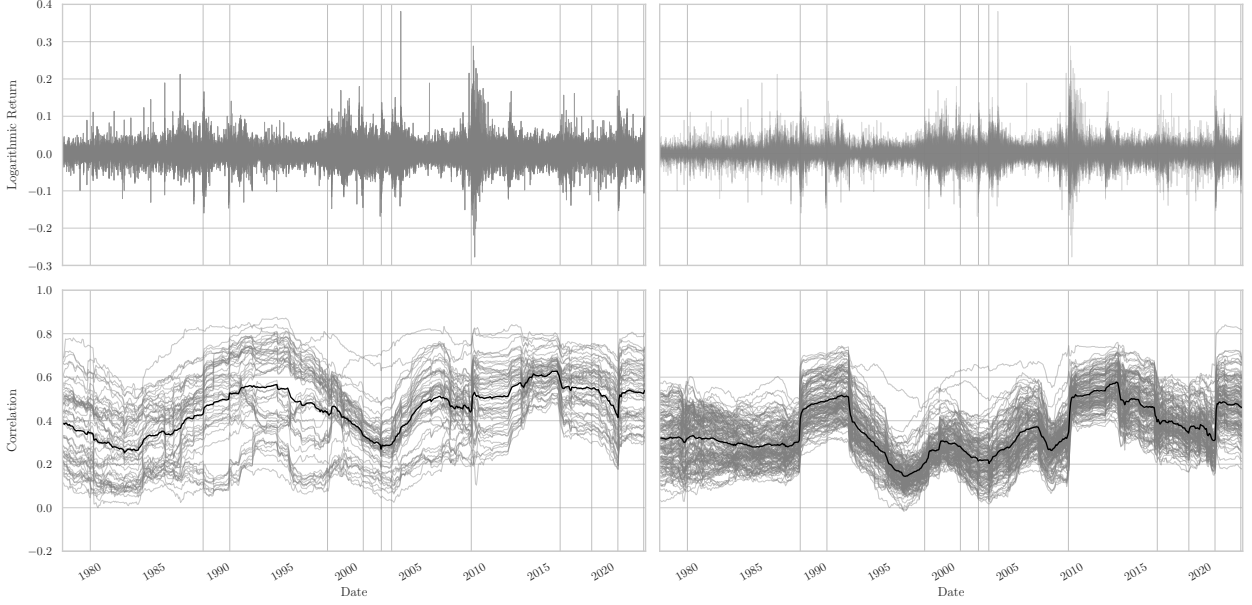


FIG. 2. Historical Stock Returns and Correlation. The top row shows the logarithmic returns of the historical stock data of the German DAX (left) and the U.S. Dow-Jones (right) index, respectively. Each line represents the logarithmic return of one stock over time. The bottom row shows the pairwise correlations between the stocks. Each line represents the correlation between two stocks over time. The black line shows the average correlation inside the index. The vertical lines represent important economic or political events.

led to the *dot-com bubble*¹⁷. The bubble burst in 2002 with substantial price declines in July and September. Finally, our data includes the 2007/2008 subprime mortgage crisis, when the market declined from its all-time high in October 2007 and crashed after the collapse of Lehman Brothers on September 15, 2008. As a result of slowing growth of the GDP of China and the Greek debt default, investors sold shares globally between 2015 and 2016. The data further includes the so-called *Volmageddon* on February 5, 2018, where a large sell-off in the U.S. stock market led to a spike in implied market volatility¹⁸. Finally, the data includes the impact of the COVID-19 pandemic, which, among other events, triggered a sudden global stock market crash on February 20, 2020. In addition, our period under review also includes a number of important global political events. These include the fall of the Berlin Wall on November 9, 1989, which triggered the collapse of the Soviet Union, the attacks of September 11, 2001, and the Russian invasion of the Ukraine on February 24, 2022.

3. Rolling Windows

To obtain dynamically evolving results, we divide the data into overlapping rolling windows¹⁹ and compute our measures for each interval following the approach by Haluszczyński *et al.*⁷. We use a sliding window of $T_w = 1000$ trading days, which corresponds to roughly four years of data. The gap or stride between successive intervals is set to $\delta T = 20$ trading days, roughly amounting to a month. As such, the w -th interval is represented as:

$$\mathbf{x}^{(w)} = (x_{1+(w-1)\delta T}, \dots, x_{T_w+(w-1)\delta T}), \quad (2)$$

which yields a total of $w = 594$ overlapping windows. A (causality) measure $\psi(\mathbf{x}, \mathbf{y}) \mapsto \mathbb{R}$, which maps two time series to a real number, is thus transformed into a vector $\Psi \in \mathbb{R}^w$.

B. Causality Measures

We select two techniques that represent prominent categories currently used in causal inference²⁰—however, it is important to note that our framework is applicable to any method capable of detecting nonlinear causality.

1. Pearson Correlation

Before describing the causal inference methods, we introduce the *Pearson correlation*²¹. We use it as a benchmark since it is still widely popular in the financial industry due to its simple calculation and interpretability⁸. It quantifies the strength and direction of the linear relationship between two variables. It is computed as follows:

$$\rho(\mathbf{x}, \mathbf{y}) \equiv \frac{\sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})}{\sqrt{\sum_{t=1}^T (x_t - \bar{x})^2} \sqrt{\sum_{t=1}^T (y_t - \bar{y})^2}}, \quad (3)$$

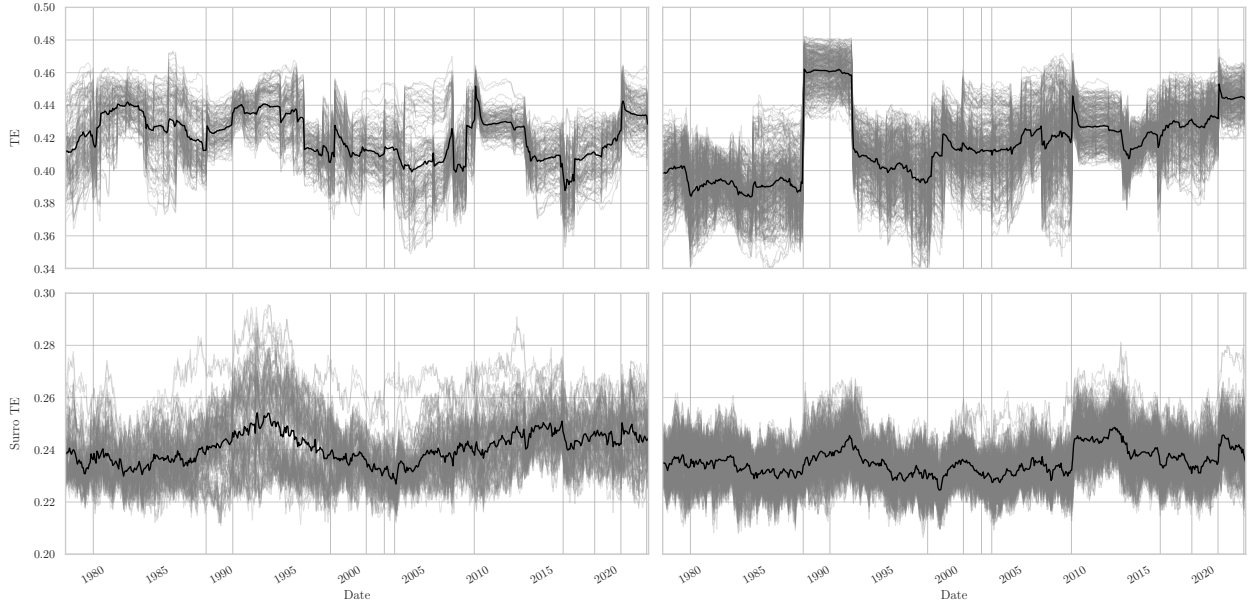


FIG. 3. Transfer Entropy. The first row shows the historical TE of stocks within the German DAX (left) and the U.S. Dow-Jones (right) indices, respectively. Each line represents one direction of the TE between two stocks over time. The bottom row illustrates the corresponding surrogate TE. The vertical lines represent important economic or political events.

where x_t denotes the stock returns at time t and $\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$ signifies their expected value. The correlation is normalized and bounded to the interval $[-1, 1]$ and thus allows direct comparisons across pairwise correlations between different stocks. As shown by Bonett and Wright²² a sample size of $T \leq 56$ is sufficient to estimate the measure reliably.

2. Transfer Entropy

Transfer Entropy (TE) is a powerful information-theoretic measure introduced by Schreiber¹⁰ which has gained popularity in the field of causal inference, particularly in the analysis of time series data. TE provides a way to quantify the directed flow of information between variables, which allows assessing causal relationships in a probabilistic framework. The TE from X to Y is defined as:

$$TE_{X \rightarrow Y} = H(Y_{t+1}, Y_t) + H(Y_t, X_t) - H(Y_{t+1}, Y_t, X_t) - H(Y_t),$$

where $H(Y_{t+1}, Y_t)$, $H(Y_t, X_t)$, $H(Y_{t+1}, Y_t, X_t)$, and $H(Y_t)$ are the joint and marginal entropies of the respective variables. To facilitate comparison between different estimations of TE, we apply the subsequent normalization:

$$TE_{X \rightarrow Y} = \frac{H(Y_{t+1}, Y_t) + H(Y_t, X_t) - H(Y_{t+1}, Y_t, X_t) - H(Y_t)}{\sqrt{H(Y_{t+1}, Y_t) \cdot H(X_{t+1}, X_t)}}. \quad (4)$$

The normalization to $[0, 1]$ stems from our understanding of TE as an asymmetric causal measure. This interpretation

aligns with the concept of covariance, which, when rescaled, results in the normalized form, the aforementioned Pearson correlation²¹.

We would like to point out that the calculation of empirical probability densities p and hence information-theoretic measures raise unexpected difficulties exceeding the scope of this work. While it is common to use histograms with equally distributed bins to estimate densities, Mynter²³ showed that this method potentially leads to biases since the estimation is dependent on the partition details—hence, finding a robust estimator is non-trivial. However, for the purpose of our research, we find that equally distributed bins perform reasonably well. Furthermore, it is worth mentioning that TE might capture false causalities depending on the dimension of conditioning¹³.

3. Convergent Cross Mapping

Convergent Cross Mapping (CCM) is an influential technique utilized for causal inference within the realm of complex dynamical systems¹¹. It aims to reveal causal connections between variables by reconstructing the dynamics that underlie them. CCM operates on the premise that variables with causal links will exhibit similar dynamical behavior, leading to a notion referred to as *shadowing*.

The underlying idea is based on Takens' theorem, which states that the entire state space can be reconstructed from a single embedded coordinate of the system, also called a *shadow manifold*²⁴. Due to transitivity, two coordinates within a system can then be mapped to each other by neigh-

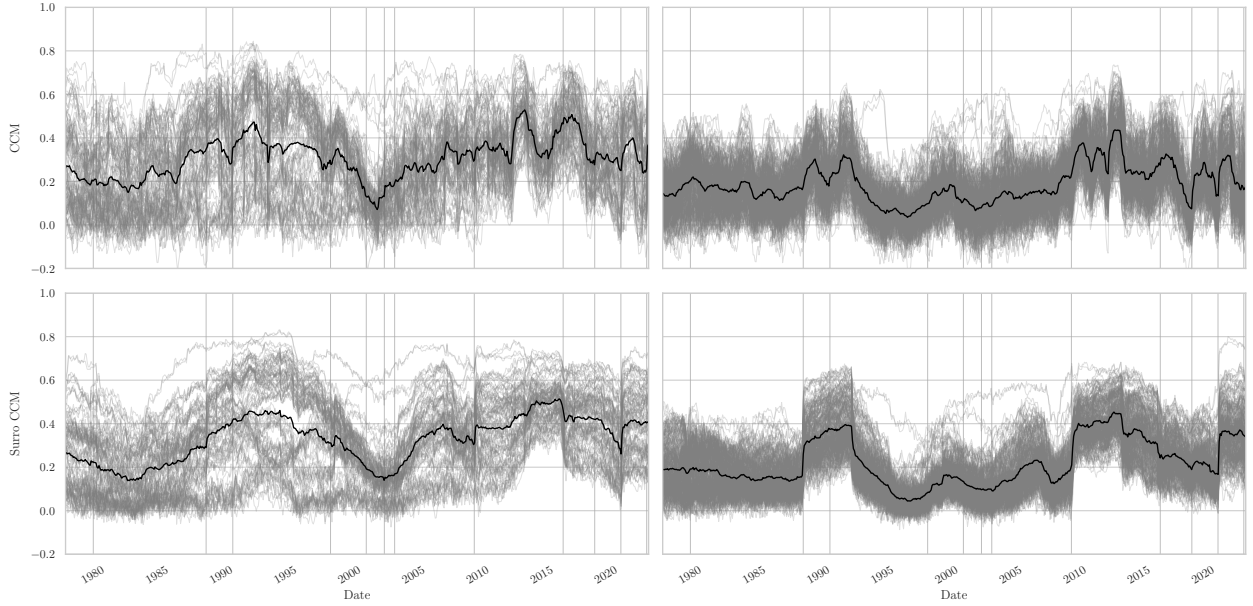


FIG. 4. Convergent Cross Mapping. The setup of this figure is analogous to Fig. 3.

boring states in their respective shadow manifolds—this allows for cross prediction. The quality of the prediction, evaluated using the Pearson correlation, quantifies the strength of the causal relationship. The algorithm of CCM can be outlined as follows:

1. **Time Delay Embedding** Embed the time series data of X and Y into higher-dimensional spaces using the embedding dimension κ and time delay τ .
2. **Library Construction** Create a library of vectors from the reconstructed state space \mathbf{X} , denoted as \mathcal{L}_X , and a library of vectors from the reconstructed state space \mathbf{Y} , denoted as \mathcal{L}_Y .
3. **Nearest Neighbor Selection** For each vector $\mathbf{X}(i)$ in the shadow manifold \mathcal{M}_X , find its nearest neighbor in \mathcal{M}_Y , denoted as $\mathbf{Y}(j)$. Similarly, for each vector $\mathbf{Y}(k)$ in \mathcal{M}_Y , find its nearest neighbor in \mathcal{M}_X , denoted as $\mathbf{X}(l)$.
4. **Cross Mapping** Assess the predictability of X based on Y by comparing the distances between the vector pairs $\mathbf{X}(i)$ and $\mathbf{Y}(j)$, and the vector pairs $\mathbf{Y}(k)$ and $\mathbf{X}(l)$. A statistical measure, such as the correlation coefficient ρ , can be used to quantify the predictability.
5. **Convergence Analysis** Repeat the cross mapping procedure for different library lengths. Evaluate the correlation as a function of the number of points used and assess the convergence of the results. The convergence of the cross mapping indicates the presence of a causal relationship between X and Y .

In the original application of CCM, convergence typically requires visual inspection. However, we've implemented a more

systematic approach using expanding windows. For a given vector of correlations ρ of size n , we calculate the standard deviation within each window. Convergence is determined if the standard deviation consistently decreases, eventually falling below a predefined threshold θ . If convergence is achieved, the mean of the last s values is calculated to smooth any outliers. Conversely, if convergence is not reached, the causality measure in CCM is set to zero. This process is mathematically expressed as:

$$CCM_{X \rightarrow Y} \equiv \begin{cases} \frac{1}{n} \sum_{i=1}^s \rho_{n-s+i} & \text{if } \rho \text{ converges} \\ 0 & \text{otherwise} \end{cases} \in [-1, 1]. \quad (5)$$

This process automates the evaluation of CCM causality for various connections within a system at a reasonable speed. To standardize the measure and render it comparable with other non-directional causal inference methods, the correlation distance, denoted as $d = \sqrt{2(1-\rho)}$, can be employed.

CCM's effectiveness in identifying causal relationships within time series data is affected by multiple aspects. The presence of noise or missing values in the data can alter the outcomes²⁵, and the choice of appropriate embedding dimensions κ and time delays τ is subject to the characteristics of the specific dataset²⁶. For example, the optimal value for τ can be determined by finding the first local minimum in the *Mutual Information* (MI) respective to τ . Additionally, the *False Nearest Neighbor* (FNN) algorithm can help finding the smallest embedding dimension that maintains the attractor's structure, ensuring that neighboring points in the original time series stay neighbors in the embedded version²⁷.

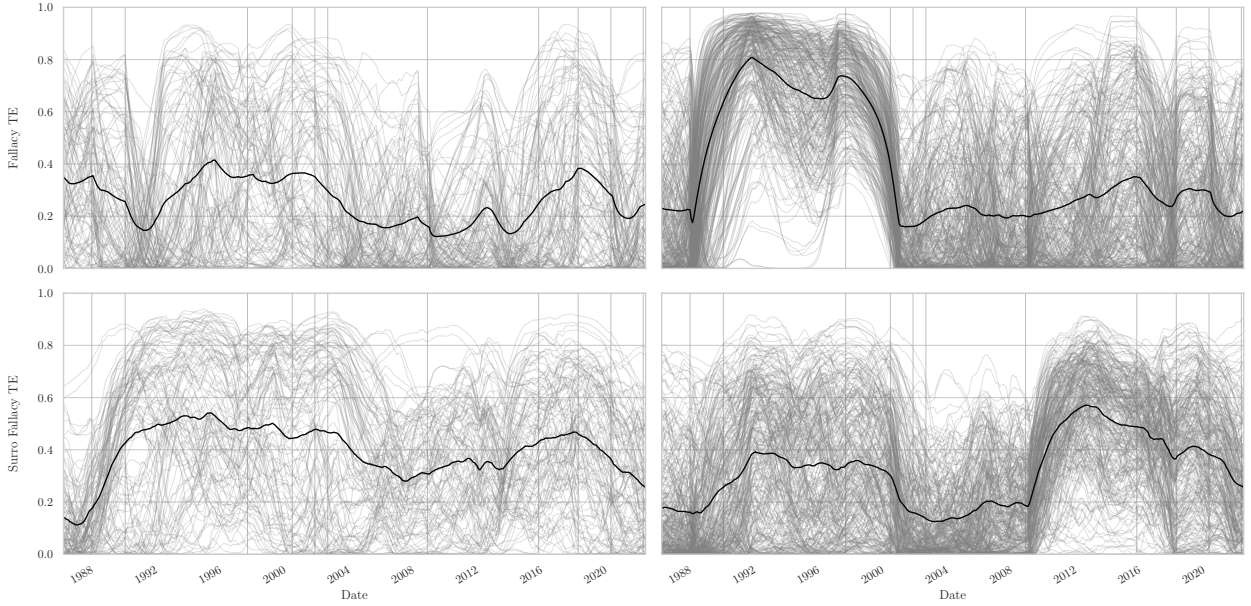


FIG. 5. Fallacy Transfer Entropy. The first row shows the historical fallacy TE of stocks within the German DAX (left) and the U.S. Dow-Jones (right) indices, respectively. Each line represents one direction of the fallacy TE between two stocks over time. The bottom row illustrates the corresponding surrogate TE. The setup of this figure is analogous to Fig. 3.

4. Limits of Causality Measures

We would like to emphasize our recognition of the limitations associated with the causal inference techniques we are presenting, as well as the broader challenges inherent in causal inference. Nonetheless, the purpose of this paper is to utilize these methods as a means to demonstrate a framework for dissecting causality into linear and nonlinear components within the context of finance. It is important to note that this paper does not delve into assessing the accuracy of these methods in capturing genuine causal relationships, nor does it explore the robustness of the methods themselves. Despite their drawbacks, these two methodologies have shown successful applications across various real-world scenarios²⁸. For in-depth analyses on TE and CCM, we recommend referring to Overbey and Todd²⁹ and Krishna and Tangirala³⁰, respectively.

Additionally, we acknowledge that TE and CCM operate with reconstructed spaces and have theoretical vulnerabilities when applied to variables within an attractor²⁸. Nevertheless, the analysis conducted in this paper relies on simulated data rather than a purely theoretical foundation. For a comprehensive discussion on the efficacy of state-space reconstruction methods in establishing causality, we direct interested readers to Cummins, Gedeon, and Spendlove³¹.

C. Linear and Nonlinear Decomposition

To decompose the causal relationships within time series systems into components originating from linear and nonlinear drivers, we employ surrogate techniques based on the *Fourier*

Transform (FT). Employing these surrogates on (causality) measures, we devise methodologies to systematically capture the quantitative breakdown of linear and nonlinear influences.

1. Fourier Transform Surrogates

FT surrogates destroy the nonlinear characteristics of a time series x while keeping the linear ones unaffected³². The algorithm to generate FT surrogates is described as follows³³:

1. **Fourier Transform:** Given a real-valued time series $\mathbf{x} = \{x_1, x_2, \dots, x_N\}$, compute its Fourier transform $\mathbf{F}(\mathbf{x})$ using the *Fast Fourier Transform* (FFT) algorithm³⁴.

$$\mathbf{F}(\mathbf{x}) = \text{FFT}(\mathbf{x})$$

2. **Phase Randomization:** Preserve the amplitudes but randomize the phases of the Fourier coefficients. This can be done by multiplying the complex Fourier coefficients by a random phase factor $e^{i\phi}$, where ϕ is uniformly distributed over the interval $[0, 2\pi]$. The phase-randomized Fourier Transform $\mathbf{F}'(\mathbf{x})$ is given by:

$$F'_k = |F_k| \cdot e^{i\phi_k}, \quad \phi_k \in [0, 2\pi]$$

3. **Inverse Fourier Transform:** Compute the inverse FT of the phase-randomized coefficients to obtain the surrogate time series $\tilde{\mathbf{x}}$:

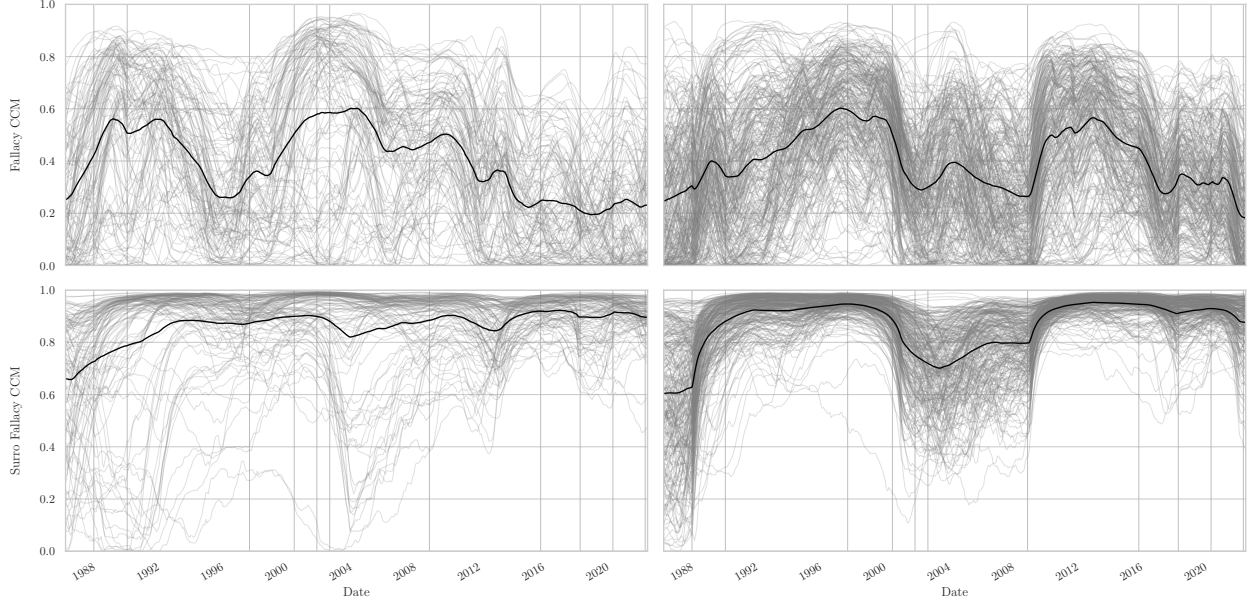


FIG. 6. Fallacy Convergent Cross Mapping. The setup of this figure is analogous to Fig. 5.

$$\tilde{\mathbf{x}} = IFFT(\mathbf{F}'(\mathbf{x}))$$

By keeping the amplitudes of the original data and only randomizing the phases, the resulting surrogates maintain the power spectral density of the original time series but break the higher-order statistical dependencies.

To enhance the reliability of our findings, we average metrics derived from surrogate time series over various instances K of random phases. The surrogate of time series \mathbf{x} , when subjected to the random phases of realization k , is denoted as $\tilde{\mathbf{x}}^{(k)}$.

2. Linear and Nonlinear Measures

In order to evaluate how much of a (causal) measure is attributed to linear or nonlinear effects, we adopt a specific approach that involves the calculation of measures on surrogate time series. Within the context of this research, we focus on a bivariate measure, denoted as $\psi(\mathbf{x}, \mathbf{y})$, which is a function mapping two time series to a real number. This function's purpose is to capture the relationship between the two time series in numerical terms. The corresponding surrogate or linear measure is defined as the average over K surrogate realizations of both time series:

$$\tilde{\psi}(\mathbf{x}, \mathbf{y}) \equiv \frac{1}{K} \sum_{k=1}^K \psi(\tilde{\mathbf{x}}^{(k)}, \tilde{\mathbf{y}}^{(k)}). \quad (6)$$

Here, the superscript k indicates that we add the same random phases to both time series within a single realization. This

choice ensures that phase differences remain unaffected, preserving specific properties such as the Pearson correlation³⁵. To ensure robustness we repeat the calculation for $K = 50$ surrogate realizations.

3. Nested Measures

As aforementioned, employing rolling windows transforms the measure ψ into a vector. This transition allows for the investigation of interrelations between two measures through a third expression:

$$\psi_{ir} \equiv \rho(\psi_1, \psi_2). \quad (7)$$

Particularly, we can utilize the Pearson correlation ρ to study the relationship between the original measure and its corresponding surrogate, expressed as:

$$\rho(\psi, \tilde{\psi}). \quad (8)$$

This method also allows for expressing the coefficient of determination using the Pearson correlation, as mentioned in³⁶:

$$R^2 = \rho^2 \in [0, 1], \quad (9)$$

This enables us to quantify the extent of the measure attributable to linear influences, more precisely, the fraction of the variability in the measure ψ that can be explained from the surrogate measure $\tilde{\psi}$. What remains then emanates from nonlinear characteristics:

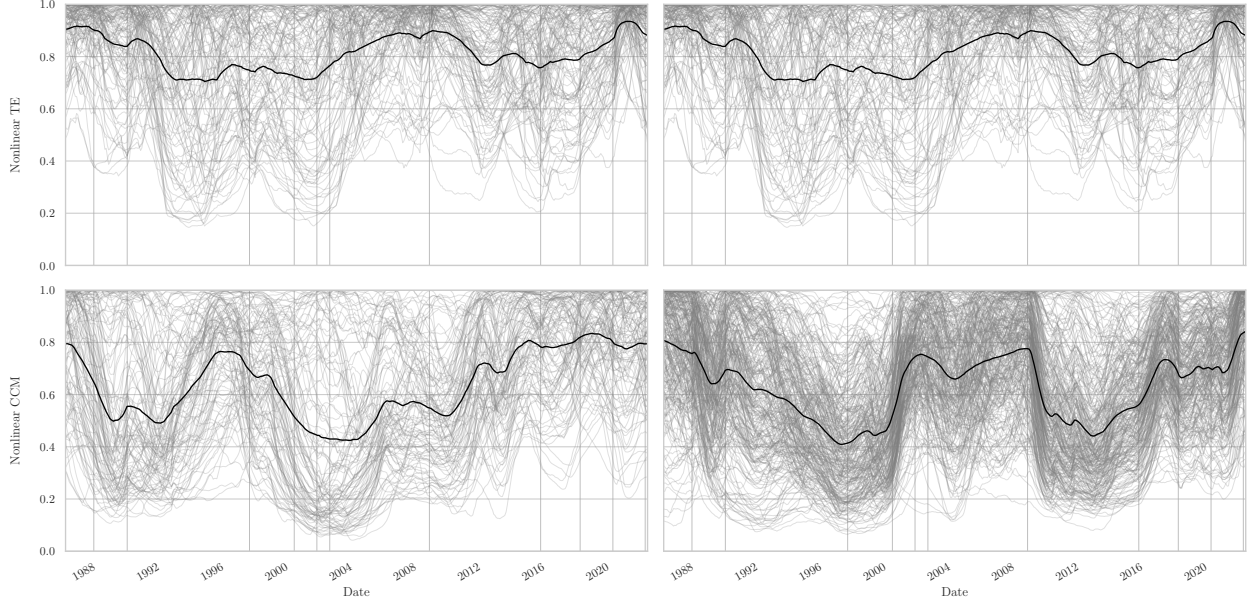


FIG. 7. Nonlinear causality. The first row shows the historical nonlinear TE of stocks within the German DAX (left) and the U.S. Dow-Jones (right) indices, respectively. Each line represents one direction of the TE between two stocks over time. The bottom row illustrates the nonlinear CCM. The vertical lines represent important economic or political events.

$$\psi_{nl} \equiv 1 - \rho^2(\psi, \tilde{\psi}). \quad (10)$$

Furthermore, there's an application to the exploration of the correlation-causality fallacy³⁷. This involves determining how much of the causality is explained by correlation:

$$\psi_{fall} \equiv \rho^2(\psi, \rho), \quad (11)$$

serving as a gauge of the causal relationship that can be explained by correlation. Specifically, this measure for the fallacy can be applied to the surrogate measure in order to evaluate how much of the linear causality is captured by correlation:

$$\psi_{fall,lin} \equiv \rho^2(\tilde{\psi}, \rho). \quad (12)$$

D. Financial Frameworks

Here, we introduce two financial frameworks and demonstrate how causality can be easily integrated, while simultaneously enhancing performance.

1. Pair Trading

Pair trading is a popular and widely utilized strategy in quantitative finance that aims to capitalize on relative price movements between two closely related assets³⁸. This strategy is

grounded in the concept of mean reversion, which assumes that over time, the prices of assets that are historically correlated tend to revert to their historical average relationship. The basic premise is to find two stocks that are highly correlated. When they deviate from this correlation (i.e., one stock moves up while the other moves down or vice versa), we take a *long* position in the underperforming stock and a *short* position in the outperforming stock, expecting them to revert to their historical correlation³⁹. Thus, a basic form of the strategy involves the following steps:

1. **Correlation Calculation:** We calculate the rolling historical and the short-term correlation between two stocks
2. **Signal Generation:** When the current correlation ρ_t deviates from its historical mean by a certain threshold, a trading signal is generated. A common approach is to use the *z-score* z of the spread, which measures the number of standard deviations by which the current correlation deviates from its historical mean:

$$z_t = \frac{\rho_t - \bar{\rho}_{hist}}{\sigma_{\rho_{hist}}}, \quad (13)$$

where $\bar{\rho}_{hist}$ and $\sigma_{\rho_{hist}}$ denote the mean and standard deviation of the historical correlation, respectively.

3. **Trade Execution:** When the *z-score* crosses a predefined threshold (e.g., above a positive threshold for a long trade or below a negative threshold for a short

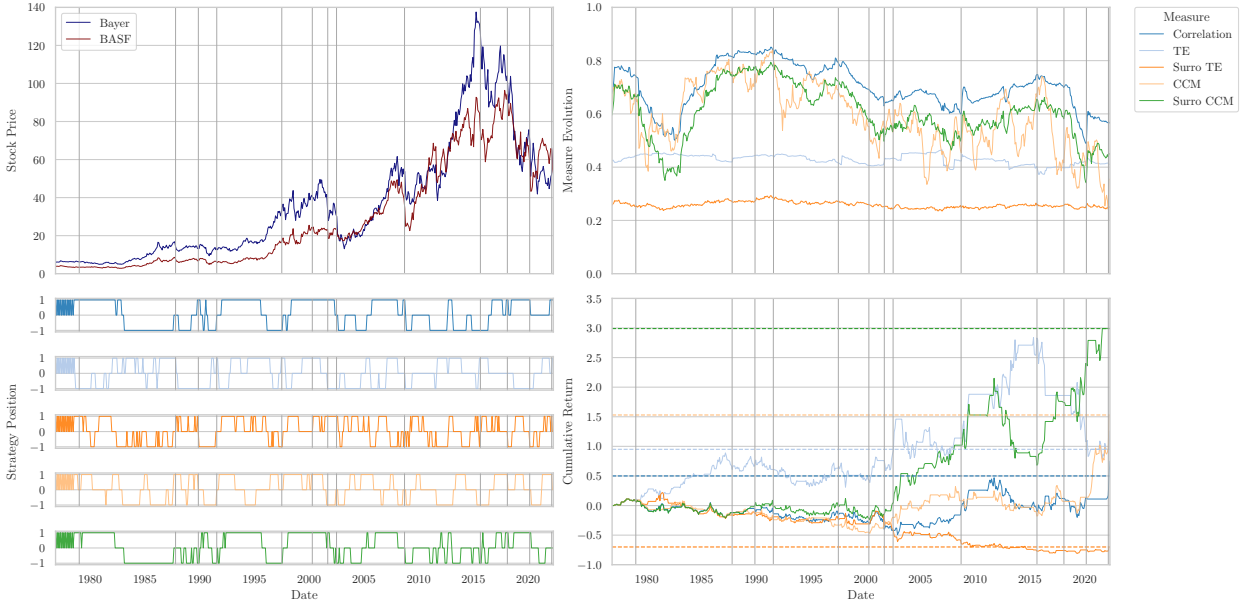


FIG. 8. Pair Trading. The stock prices of two companies from the DAX (Bayer and BASF) are displayed in the top left figure. The top right figure presents the co-dependence measures over time, with each color corresponding to a specific co-dependence measure that is included in the legend on the right-hand side. The bottom left chart illustrates the strategy positions over time, with long position in Bayer and short position in BASF indicated by 1, the opposite indicated by -1 , and no investment indicated by 0. The graph in the lower right corner illustrates the cumulative return achieved by the strategy over time. The dotted horizontal lines mark the strategy's most recent cumulative return value. The vertical lines indicate notable economic or political events.

trade), a trade is initiated. A long trade involves buying the underperforming asset and simultaneously shorting the overperforming asset. We set the threshold at $z_t \pm 1.5$.

- Profit Taking:** The strategy aims to profit from the mean reversion process. As the spread narrows and returns to its historical mean, the positions are unwound, resulting in a profit.

We would like to note that we are aware of the simplifications of the strategy and that for practical use more fine-tuning is necessary. However, we find the parametrization of the strategy to be sufficient for illustrative purposes. For our purposes we exchange the historical Pearson correlation with the TE and CCM respectively.

2. Portfolio Optimization

In the world of finance, *Markowitz Portfolio Theory* (MPT), developed by Harry Markowitz in 1952, is a cornerstone concept for investors and financial analysts⁴⁰. This theory revolutionized the way investors think about constructing portfolios. It is based on a fundamental premise: rational investors seek to maximize their portfolio's expected return while minimizing its risk. The key insight here is that an asset's risk and return should not be evaluated in isolation but rather in the context of the entire portfolio.

The expected return of a portfolio is calculated as a weighted sum of the expected returns of its individual assets:

$$E(R_p) = \sum_{i=1}^n w_i \cdot E(R_i), \quad (14)$$

where $E(R_p)$ is the expected return of the portfolio, w_i is the weight of asset i in the portfolio, and $E(R_i)$ is the expected return of asset i . Even though historic returns do not indicate future performance, it is common to use the historical mean as a proxy for the expected returns³⁹.

The portfolio's variance is a measure of its risk. It considers not only the individual asset variances but also the correlation between assets. The formula for portfolio variance is:

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i \cdot w_j \cdot \sigma_i \cdot \sigma_j \cdot \rho_{ij}, \quad (15)$$

where σ_p^2 is the variance of the portfolio, w_i and w_j are the weights of assets i and j in the portfolio, and σ_{ij} is the covariance between assets i and j . We can replace the correlation with a causality measure ψ or use the sign of the correlation if the measure ψ is normalized to $[0, 1]$:

$$\sigma_p^2 = \sum_{i=1}^n \sum_{j=1}^n w_i \cdot w_j \cdot \sigma_i \cdot \sigma_j \cdot \psi_{ij} \cdot \text{sgn}(\rho_{ij}), \quad (16)$$

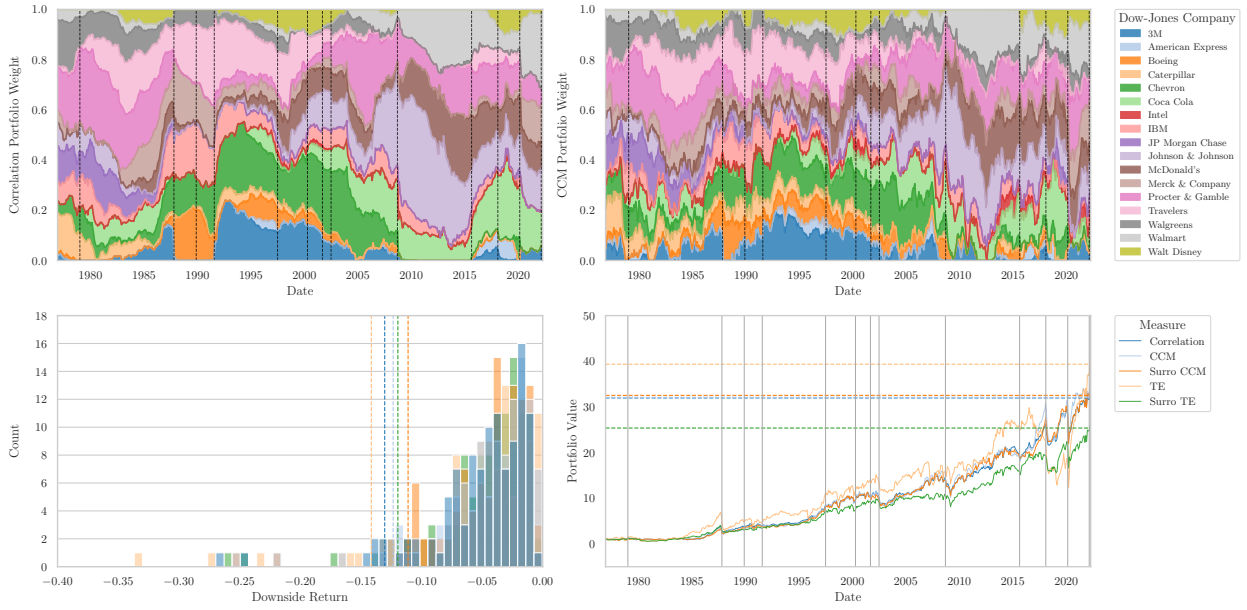


FIG. 9. Minimum Risk Portfolio Optimization. The top row displays the optimized Minimum Risk Portfolio weights over time using both the correlation (on the left) and CCM (on the right) as co-dependence measures. Each colored area represents a stock from the Dow-Jones, which is mapped in the legend to the right. The dotted vertical lines depict significant economic or political events. In the bottom row, the left figure illustrates the distributions of the downside returns when using different co-dependence measures. The vertical lines depict the VaR at $\alpha = 1\%$ level. The plot to the right displays the portfolio's value over time. The vertical lines denote significant economic or political occurrences. The dotted horizontal lines denote the portfolio's most recent value. Each color corresponds to a particular codependence measure, which is mapped in the right-hand side legend.

where $sgn(\cdot)$ denotes the Sign function.

A popular measure of the riskiness of historical portfolio performance is *Value-at-Risk* (VaR), which quantifies the potential loss in value of an investment or portfolio over a specified time horizon at a α^{41} confidence level. A $1 - \alpha$ VaR = x means that there is a α chance that the portfolio will lose more than x . Unlike standard deviation, VaR measures tail risk and does not assume a normal distribution, which is particularly important for risk management purposes. We use the default value of $\alpha = 1\%$.

Two portfolios of great importance within MPT are the *Minimum Risk Portfolio* and the *Maximum Sharpe Ratio Portfolio*. These portfolios play a crucial role in portfolio analysis and optimization:

- **Minimum Risk:** The Minimum Risk Portfolio represents the portfolio with the lowest possible risk for a given set of assets. Mathematically, it can be formulated as an optimization problem. The solution to this problem provides the weights of assets in the Minimum Risk Portfolio:

$$\begin{aligned} &\text{Minimize } \sigma_p^2 \\ &\text{Subject to } E(R_p) = \text{target return} \\ &\sum_{i=1}^n w_i = 1 \\ &w_i \geq 0 \quad \text{for all } i \end{aligned}$$

- **Maximum Sharpe Ratio:** The Maximum Sharpe Ratio Portfolio represents the portfolio that offers the highest risk-adjusted return. The Sharpe Ratio (S) measures this risk-adjusted performance:

$$S = \frac{E(R_p - R_f)}{\sigma_p} \tag{17}$$

To find the Maximum Sharpe Ratio Portfolio, we maximize the Sharpe Ratio by adjusting the asset weights. Mathematically:

$$\begin{aligned} &\text{Maximize } S \\ &\text{Subject to } \sum_{i=1}^n w_i = 1 \\ &w_i \geq 0 \quad \text{for all } i \end{aligned}$$

We illustrate a simple way to incorporate causality measures into portfolio construction through the utilization of these two portfolios. As a result, we regularly adjust the portfolio by optimizing its weightings with the mentioned algorithms to align it with the prevailing market conditions. To achieve this, we apply the rolling causality measures as previously demonstrated in this paper. Consequently, we can assess the advantages of using causality measures as the co-dependency metric for the portfolio, examining both its performance and risk management implications.

III. RESULTS

In the following we present the results of our analyses, which we structure into three subsections. As motivated by Figure 1, we observe that for complex and chaotic systems it is difficult to measure the co-dependence of variables through correlations as they can exhibit different regimes of positive, negative, and no correlation even though they are guided by exactly the same governing equations. This is illustrated by the rolling window analysis of the correlation, which is unrobust and changes significantly over time. Hence in order to measure their co-dependence reliably, another measure is needed. Causality measures, such as CCM, are a valuable technique to measure the causality of two variables in both directions and provide stable results over time. Furthermore, by using FT surrogates, we can separate the causality in linear and nonlinear contributions which helps to understand the intricate nature of the co-dependence. The Figure shows that the separation of causality is stable over different windows and also plausible when compared to the governing Equations 1.

A. Historical Causality

To demonstrate the practical applicability of our framework, we have employed it in an analysis of major German and U.S. stock indices. The data and the dynamic correlation patterns are visually depicted in Figure 2. Notably, these correlations undergo significant shifts during and after pivotal economic and political events. This phenomenon can be attributed to the changing behavior of investors and other market participants in response to these impactful occurrences. Furthermore, this effect extends to our investigation of causality measures, as demonstrated in Figures 3 and 5. These figures reveal that linear and nonlinear causality measures, such as Transfer Entropy (TE) and Convergent Cross Mapping (CCM), exhibit analogous responses to these events.

Specifically, when examining TE, it becomes apparent that TE is highly responsive to these events, displaying sharp fluctuations. In contrast, surrogate TE remains relatively stable and does not react as drastically. Conversely, surrogate CCM appears to respond more strongly than regular CCM, displaying significant jumps similar to the observed patterns in correlation. One of the most striking examples of this behavior is observed during Black Monday in 1987, where we witness substantial increases in correlation, TE, and surrogate

CCM, particularly in the context of U.S. stocks. Two other significant events that exhibit similar patterns are the global financial crisis in 2009 and the COVID-19 pandemic in 2020. These observations suggest that these events triggered structural shifts in the market, which is reasonable given their profound impacts on the global economy. An intriguing observation is that TE experiences more pronounced fluctuations compared to surrogate TE during these events, while the opposite is observed for CCM. This suggests that the linear dynamics in the stock markets were more profoundly influenced, possibly due to investors simultaneously adjusting their stock positions in response to the market crashes.

B. Correlation-Causality Fallacy and Nonlinear Causality

Upon examination of Figure 5, it becomes evident that both the original and surrogate Transfer Entropy (TE) exhibit a moderate correlation. Notably, there is an intriguing exception during the period spanning from approximately 1990 to 2002 in the U.S. stock market, where a substantial portion, approximately 75%, of TE can be attributed to correlation. This spike coincided with the rise and eventual burst of the dotcom bubble, suggesting that it might have served as an indicator of abnormal market behavior during this period.

One of the most significant findings from this analysis is the observation that fallacy of surrogate Convergent Cross Mapping (CCM) is remarkably high, around 90%, in both the German and U.S. stock indices, as depicted in Figure 6. This suggests that correlation effectively acted as a suitable proxy for linear causality for the majority of the past few decades. However, in periods where this fallacy diminishes, such as the aftermath of the dotcom bubble in 2002 and the onset of the global financial crisis in 2008, relying solely on correlation as a measure of co-dependence significantly underestimates portfolio risk, as nonlinear effects cannot be disregarded. This effect is even more pronounced when examining the fallacy of the original CCM, where we also observe a substantial drop during these phases.

To gauge the extent of nonlinear contributions to our causality measures, we delve into the analysis of how much of the causality can be accounted for by its surrogate. In Figure 7, we observe the evolution of nonlinear causality over time, noting that nonlinear TE and CCM exhibit similar but not identical behaviors. Both measures reveal heightened levels of nonlinearity during the period between the dotcom bubble burst and the commencement of the global financial crisis. In contrast, before and after this period, we observe phases with less nonlinearity. This indicates that these two major economic events should be assessed differently, as the dotcom bubble led to increased nonlinearity in its aftermath, while the global financial crisis, precipitated by the U.S. housing market crisis, ushered in a phase of more linear market behavior. Particularly for CCM, this behavior is quite drastic, with jumps exceeding 20%. In conclusion, our analysis suggests that nonlinear causality can be a valuable tool for anticipating and evaluating financial impacts, provided it is continually monitored and assessed in the context of evolving market dynamics.

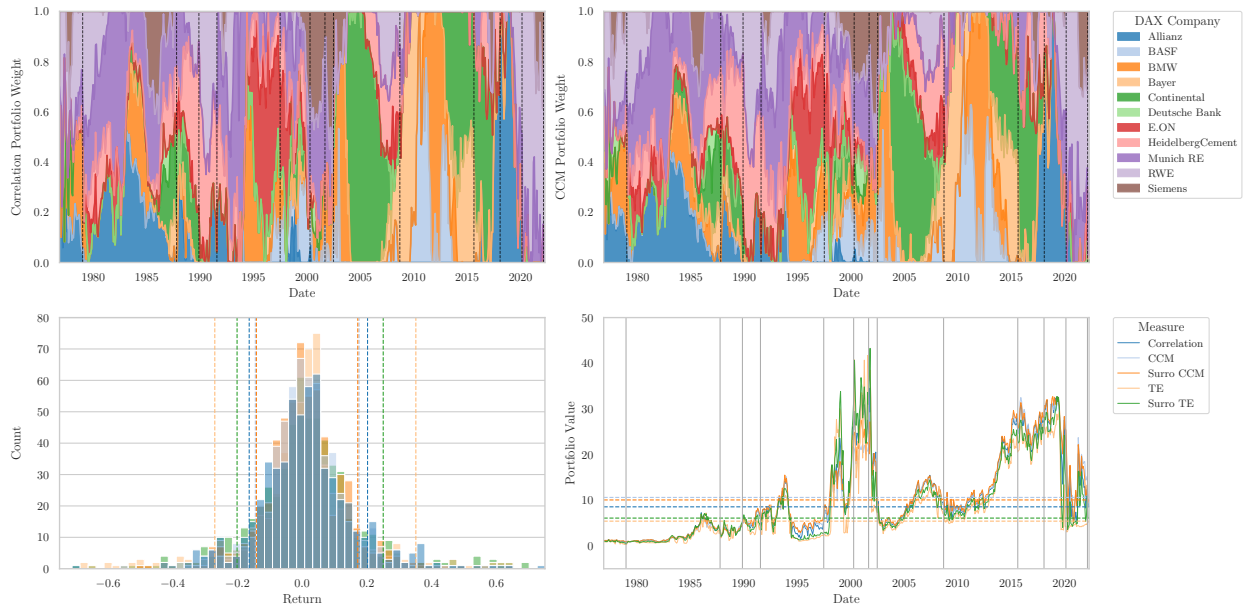


FIG. 10. Maximum Sharpe Ratio Portfolio Optimization. The setup of this figure is analogous to Fig. 9. The top row displays the optimized Maximum Sharpe Ratio Portfolio weights over time using both the correlation (on the left) and CCM (on the right) as co-dependence measures. Each colored area represents a stock from the DAX, which is mapped in the legend to the right. The dotted vertical lines depict significant economic or political events. In the bottom row, the left figure illustrates the distributions of the returns when using different co-dependence measures. The vertical lines depict the standard deviations of the returns. The plot to the right displays the portfolio’s value over time. The vertical lines denote significant economic or political occurrences. The dotted horizontal lines denote the portfolio’s most recent value. Each color corresponds to a particular co-dependence measure, which is mapped in the right-hand side legend.

C. Pair Trading and Portfolio Optimization

To effectively apply causality measures in practical financial scenarios, we present two common financial frameworks where the interdependence between assets plays a pivotal role. The first concept we explore is pair trading, a logical choice given its reliance on the idea that two assets tend to revert to a default correlation, and deviations from this norm can be profitably exploited. In Figure 8, we use two German stocks from the chemical industry, Bayer and BASF, to illustrate how causality measures can be seamlessly integrated. It’s noteworthy that even though the differences in the evolution of co-dependence measures are relatively similar, over time, these subtle distinctions significantly impact trading performance. Of particular interest is the fact that the trading strategy employing surrogate Convergent Cross Mapping (CCM) outperforms the one utilizing correlation by a substantial margin, approximately six times, despite the measures’ apparent similarity. Additionally, we observe that Transfer Entropy (TE) and CCM perform better than correlation, while surrogate TE lags behind and even delivers negative returns. This straightforward example underscores the potential of a causality-based pair trading strategy.

As previously highlighted, relying solely on correlation can potentially lead to an underestimation of risk, a perilous scenario when managing a portfolio. In Figure 9, we employ stocks from the U.S. Dow-Jones index and minimize risk by

dynamically optimizing the portfolio weights on a monthly basis. It becomes evident that the allocations of a portfolio using correlation and CCM exhibit visible disparities over time. This divergence is reflected in the portfolio’s downside returns and overall performance. Notably, we observe that a portfolio employing surrogate Transfer Entropy (TE), CCM, and surrogate CCM achieves a superior 1% Value at Risk (VaR) while slightly enhancing portfolio performance.

Similarly, in the context of optimizing the Sharpe Ratio, as depicted in Figure 10, the inclusion of causality measures results in a more favorable risk-return profile. When optimizing the stocks of the German DAX index, we note a reduction in portfolio standard deviation and an increase in portfolio value over time, particularly when employing original and surrogate CCM.

IV. CONCLUSION AND OUTLOOK

The present study has addressed the issue of identifying and quantifying co-dependence among financial instruments, which continues to be a paramount challenge for both researchers and practitioners in the financial industry. While traditional linear measures like the Pearson correlation have maintained their prominence, this paper has introduced a novel framework aimed at analyzing both linear and nonlinear causal relationships within financial markets. To achieve this, we have employed two distinct causal inference method-

ologies, namely Transfer Entropy and Convergent Cross-Mapping, and have utilized Fourier transform surrogates to disentangle their respective linear and nonlinear contributions.

Our findings have unveiled that stock indices in Germany and the U.S. exhibit a substantial degree of nonlinear causality, a phenomenon that has largely eluded previous investigations. It is important to recognize that while correlation, exemplified by the Pearson correlation coefficient, serves as an excellent proxy for linear causality, it falls short in capturing the intricate nonlinear dynamics that underlie financial markets. Consequently, relying solely on correlation can lead to an underestimation of causality itself.

The framework introduced in this study not only facilitates the quantification of nonlinear causality but also sheds light on the perilous "correlation-causality fallacy." By delving into the nuances of causality, we have motivated how these insights can be harnessed for practical applications, including inferring market signals, implementing pair trading strategies, and enhancing the management of portfolio risk.

One of the insights derived from our findings underscores the role that both linear and nonlinear causality can play as early warning indicators for unusual market dynamics. Furthermore, our results suggest that a straightforward incorporation of these causality measures into strategies, such as pair trading and portfolio optimization, can yield better outcomes compared to a reliance solely on Pearson correlation. This understanding can significantly empower traders and risk managers, enabling them to craft more effective trading strategies and to adopt a more proactive approach to risk mitigation.

Looking ahead, the implications of our findings extend to various facets of financial research and practice. Further exploration of nonlinear causality may uncover new dimensions of financial market interactions, potentially leading to the development of innovative trading algorithms and risk management tools. Additionally, the integration of causality measures into existing financial models and frameworks holds the promise of enhancing their predictive accuracy and robustness.

In conclusion, this paper has introduced a comprehensive framework for disentangling linear and nonlinear causality within financial markets. The revelation of substantial nonlinear causality and the recognition of the limitations of traditional correlation measures underline the importance of taking a more nuanced approach to co-dependency analysis. The insights gained from this study have the potential to enhance the way we perceive and navigate the intricacies of financial markets, contributing to more informed decision-making, better risk management practices, and more financial stability.

ACKNOWLEDGMENTS

We would like to thank the DLR and AllianzGI for providing data and computational resources.

¹F. Jovanovic, R. N. Mantegna, and C. Schinckus, "When financial economics influences physics: The role of econophysics," Available at SSRN 3294548 (2018).

- ²G.-J. Wang, C. Xie, S. Chen, J.-J. Yang, and M.-Y. Yang, "Random matrix theory analysis of cross-correlations in the us stock market: Evidence from Pearson's correlation coefficient and detrended cross-correlation coefficient," *Physica A: statistical mechanics and its applications* **392**, 3715–3730 (2013).
- ³R. N. Mantegna and H. E. Stanley, *Introduction to econophysics: correlations and complexity in finance* (Cambridge university press, 1999).
- ⁴G.-J. Wang, C. Xie, and H. E. Stanley, "Correlation structure and evolution of world stock markets: Evidence from Pearson and partial correlation-based networks," *Computational Economics* **51**, 607–635 (2018).
- ⁵R. N. Mantegna and H. E. Stanley, "Scaling behaviour in the dynamics of an economic index," *Nature* **376**, 46–49 (1995).
- ⁶S. Ghoshghaie, W. Breyman, J. Peinke, P. Talkner, and Y. Dodge, "Turbulent cascades in foreign exchange markets," *Nature* **381**, 767–770 (1996).
- ⁷A. Haluszczynski, I. Laut, H. Modest, and C. R ath, "Linear and nonlinear market correlations: Characterizing financial crises and portfolio optimization," *Physical Review E* **96**, 062315 (2017).
- ⁸J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," in *Noise reduction in speech processing* (Springer, 2009) pp. 1–4.
- ⁹C. W. Granger, *Essays in econometrics: collected papers of Clive WJ Granger*, Vol. 32 (Cambridge University Press, 2001).
- ¹⁰T. Schreiber, "Measuring information transfer," *Physical review letters* **85**, 461 (2000).
- ¹¹G. Sugihara, R. May, H. Ye, C.-h. Hsieh, E. Deyle, M. Fogarty, and S. Munch, "Detecting causality in complex ecosystems," *science* **338**, 496–500 (2012).
- ¹²X. Ge and A. Lin, "Dynamic causality analysis using overlapped sliding windows based on the extended convergent cross-mapping," *Nonlinear Dynamics* **104**, 1753–1765 (2021).
- ¹³M. Paluř and M. Vejmelka, "Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections," *Physical Review E* **75**, 056211 (2007).
- ¹⁴J. Hlinka, D. Hartman, M. Vejmelka, D. Novotn a, and M. Paluř, "Non-linear dependence and teleconnections in climate data: sources, relevance, nonstationarity," *Climate dynamics* **42**, 1873–1886 (2014).
- ¹⁵A. L. Lloyd, "The coupled logistic map: a simple model for the effects of spatial heterogeneity on population dynamics," *Journal of Theoretical Biology* **173**, 217–230 (1995).
- ¹⁶S. J. Brown, W. Goetzmann, R. G. Ibbotson, and S. A. Ross, "Survivorship bias in performance studies," *The Review of Financial Studies* **5**, 553–580 (1992).
- ¹⁷J. B. DeLong and K. Magin, "A short note on the size of the dot-com bubble," (2006).
- ¹⁸H. P. Krishnan, A. Bennington, H. P. Krishnan, and A. Bennington, "The vix "volmageddon", with exchange-traded notes destabilizing the market," *Market Tremors: Quantifying Structural Risks in Modern Financial Markets*, 83–119 (2021).
- ¹⁹E. Zivot, J. Wang, E. Zivot, and J. Wang, "Rolling analysis of time series," *Modeling financial time series with S-Plus®*, 299–346 (2003).
- ²⁰J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science* **28**, 075310 (2018).
- ²¹K. Pearson, "VII. note on regression and inheritance in the case of two parents," *proceedings of the royal society of London* **58**, 240–242 (1895).
- ²²D. G. Bonett and T. A. Wright, "Sample size requirements for estimating Pearson, Kendall and Spearman correlations," *Psychometrika* **65**, 23–28 (2000).
- ²³M. Mynter, "Evaluation and extension of the transfer entropy calculus for the measurement of information flows between futures time series during the covid-19 pandemic," (2021).
- ²⁴F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80* (Springer, 2006) pp. 366–381.
- ²⁵D. M onster, R. Fusaroli, K. Tyl en, A. Roepstorff, and J. F. Sherson, "Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence," *Future Generation Computer Systems* **73**, 52–62 (2017).
- ²⁶S. Wallot and D. M onster, "Calculation of average mutual information (ami) and false-nearest neighbors (fnn) for the estimation of embedding parameters of multidimensional time series in matlab," *Frontiers in psychology* **9**,

- 1679 (2018).
- ²⁷M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Physical review A* **45**, 3403 (1992).
- ²⁸J. M. McCracken and R. S. Weigel, "Convergent cross-mapping and pairwise asymmetric inference," *Physical Review E* **90**, 062903 (2014).
- ²⁹L. Overbey and M. Todd, "Effects of noise on transfer entropy estimation for damage detection," *Mechanical Systems and Signal Processing* **23**, 2178–2191 (2009).
- ³⁰P. Krishna and A. K. Tangirala, "Inferring direct causality from noisy data using convergent cross mapping," in *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)* (IEEE, 2019) pp. 1523–1528.
- ³¹B. Cummins, T. Gedeon, and K. Spendlove, "On the efficacy of state space reconstruction methods in determining causality," *SIAM Journal on Applied Dynamical Systems* **14**, 335–381 (2015).
- ³²C. R ath and R. Monetti, "Surrogates with random fourier phases," in *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference* (World Scientific, 2009) pp. 274–285.
- ³³C. R ath, M. Gliozzi, I. Papadakis, and W. Brinkmann, "Revisiting algorithms for generating surrogate time series," *Physical review letters* **109**, 144101 (2012).
- ³⁴E. O. Brigham, *The fast Fourier transform and its applications* (Prentice-Hall, Inc., 1988).
- ³⁵D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Physical review letters* **73**, 951 (1994).
- ³⁶E. Kasuya, "On the use of r and r squared in correlation and regression," Tech. Rep. (Wiley Online Library, 2019).
- ³⁷M. Maziarz, "A review of the granger-causality fallacy," *The journal of philosophical economics: Reflections on economic and social issues* **8**, 86–105 (2015).
- ³⁸G. Vidyamurthy, *Pairs Trading: quantitative methods and analysis*, Vol. 217 (John Wiley & Sons, 2004).
- ³⁹J. C. Hull, "Options, futures and other derivatives," (2019).
- ⁴⁰H. M. Markowitz, "Foundations of portfolio theory," *The journal of finance* **46**, 469–477 (1991).
- ⁴¹D. Duffie and J. Pan, "An overview of value at risk," *Journal of derivatives* **4**, 7–49 (1997).
- ⁴²P. B erard, G. Besson, and S. Gallot, "Embedding riemannian manifolds by their heat kernel," *Geometric & Functional Analysis GAFA* **4**, 373–398 (1994).
- ⁴³A. A. Tsonis, E. R. Deyle, H. Ye, and G. Sugihara, "Convergent cross mapping: theory and an example," in *Advances in Nonlinear Geosciences* (Springer, 2018) pp. 587–600.
- ⁴⁴L. Barnett, A. B. Barrett, and A. K. Seth, "Granger causality and transfer entropy are equivalent for gaussian variables," *Physical review letters* **103**, 238701 (2009).
- ⁴⁵T. Schreiber and A. Schmitz, "Improved surrogate data for nonlinearity tests," (1999), arXiv:chao-dyn/9909041 [chao-dyn].
- ⁴⁶D. Prichard and J. Theiler, "Generating surrogate data for time series with several simultaneously measured variables," *Phys. Rev. Lett.* **73**, 951–954 (1994).
- ⁴⁷redm: An r package for empirical dynamic modeling and convergent cross mapping," <https://mran.revolutionanalytics.com/snapshot/2018-07-06/web/packages/rEDM>, accessed: 2020-08-20.
- ⁴⁸S. L. Bressler and A. K. Seth, "Wiener–granger causality: a well established methodology," *Neuroimage* **58**, 323–329 (2011).
- ⁴⁹S. Ponczek, "To understand the wild u.s. stock rally, just forget about 2020," (2020).
- ⁵⁰M. Paluř, V. Albrecht, and I. Dvoř ak, "Information theoretic test for nonlinearity in time series," *Physics Letters A* **175**, 203–209 (1993).
- ⁵¹J. Hlinka, D. Hartman, M. Vejmelka, J. Runge, N. Marwan, J. Kurths, and M. Paluř, "Reliability of inference of directed climate networks using conditional mutual information," *Entropy* **15**, 2023–2045 (2013).
- ⁵²M. B. Kennel and M. Buhl, "Estimating good discrete partitions from observed data: Symbolic false nearest neighbors," *Physical Review Letters* **91**, 084102 (2003).
- ⁵³D. A. Hsieh, "Nonlinear dynamics in financial markets: evidence and implications," *Financial Analysts Journal* **51**, 55–62 (1995).
- ⁵⁴R. N. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B-Condensed Matter and Complex Systems* **11**, 193–197 (1999).
- ⁵⁵R. Cont, "Empirical properties of asset returns: stylized facts and statistical issues," *Quantitative finance* **1**, 223 (2001).
- ⁵⁶D. P. Doane and L. E. Seward, "Measuring skewness: a forgotten statistic?" *Journal of statistics education* **19** (2011).
- ⁵⁷J. C. Hull, *Options futures and other derivatives* (Pearson Education India, 2003).

Bibliography

- [1] S. H. Strogatz, *Nonlinear Dynamics and Chaos With Applications to Physics, Biology, Chemistry and Engineering*. CRC Press, 2018. DOI: 10.1017/CB09780511998188.
- [2] E. Lorenz, “The butterfly effect”, *World Scientific Series on Nonlinear Science Series A*, vol. 39, pp. 91–94, 2000.
- [3] A. Wolf, J. B. Swift, H. L. Swinney, & J. A. Vastano, “Determining lyapunov exponents from a time series”, *Physica D: Nonlinear Phenomena*, vol. 16, no. 3, pp. 285–317, 1984. DOI: [https://doi.org/10.1016/0167-2789\(85\)90011-9](https://doi.org/10.1016/0167-2789(85)90011-9).
- [4] V. Anishchenko, T. Vadivasova, D. Postnov, & M. Safonova, “Synchronization of chaos”, *International Journal of Bifurcation and Chaos*, vol. 2, no. 03, pp. 633–644, 1992.
- [5] I. Newton, *Philosophiae naturalis principia mathematica*. G. Brookman, 1833, vol. 1.
- [6] J. Barrow-Green, *Poincaré and the three body problem*. American Mathematical Soc., 1997.
- [7] E. N. Lorenz, “Deterministic nonperiodic flow”, *Journal of atmospheric sciences*, vol. 20, no. 2, pp. 130–141, 1963.
- [8] B. Mandelbrot, “How long is the coast of britain? statistical self-similarity and fractional dimension”, *science*, vol. 156, no. 3775, pp. 636–638, 1967.
- [9] D. Prosperino, “Estimating parameters of governing equations of non-linear systems from data using synchronisation and machine learning”, Master’s Thesis, Ludwig-Maximilians-Universität, 2022.
- [10] J. Runge, “Causal network reconstruction from time series: From theoretical assumptions to practical estimation”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075 310, 2018.
- [11] H. R. Brown & D. Lehmkuhl, “Einstein, the reality of space and the action–reaction principle”, in *Einstein, Tagore and the nature of reality*, Routledge, 2016, pp. 27–54.
- [12] J. S. Bell, “Free variables and local causality”, in *Quantum mechanics, high energy physics and accelerators. Selected papers of John S. Bell (with commentary)*, 1995.
- [13] T. Schreiber, “Measuring information transfer”, *Physical review letters*, vol. 85, no. 2, p. 461, 2000.
- [14] C. W. Granger, *Essays in econometrics: collected papers of Clive WJ Granger*. Cambridge University Press, 2001, vol. 32.
- [15] G. Sugihara, R. May, H. Ye, et al., “Detecting causality in complex ecosystems”, *science*, vol. 338, no. 6106, pp. 496–500, 2012.
- [16] K. Ried, M. Agnew, L. Vermeyden, D. Janzing, R. W. Spekkens, & K. J. Resch, “A quantum advantage for inferring causal structure”, *Nature Physics*, vol. 11, no. 5, pp. 414–420, 2015.
- [17] C. R ath & R. Monetti, “Surrogates with random fourier phases”, in *Topics on Chaotic Systems: Selected Papers from Chaos 2008 International Conference*, World Scientific, 2009, pp. 274–285.

- [18] E. Hairer, S. P. Nørsett, & G. Wanner, *Solving Ordinary Differential Equations I*. Springer, 1993. DOI: 10.1007/978-3-540-78862-1.
- [19] L. Perko, *Differential equations and dynamical systems*. Springer Science & Business Media, 2013, vol. 7.
- [20] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, & D. Sejdinovic, “Detecting and quantifying causal associations in large nonlinear time series datasets”, *Science Advances*, vol. 5, no. 11, Nov. 2019. DOI: 10.1126/sciadv.aau4996. [Online]. Available: <https://doi.org/10.1126%5C%2Fsciadv.aau4996>.
- [21] B. C. Daniels & I. Nemenman, “Automated adaptive inference of phenomenological dynamical models”, *Nature communications*, vol. 6, no. 1, pp. 1–8, 2015.
- [22] J. Pathak, A. Wikner, R. Fussell, et al., “Hybrid forecasting of chaotic processes: Using machine learning in conjunction with a knowledge-based model”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 4, 2018.
- [23] A. R. S. Parmezan, V. M. Souza, & G. E. Batista, “Evaluation of statistical and machine learning models for time series prediction: Identifying the state-of-the-art and the best conditions for the use of each model”, *Information sciences*, vol. 484, pp. 302–337, 2019.
- [24] M. I. Jordan & T. M. Mitchell, “Machine learning: Trends, perspectives, and prospects”, *Science*, vol. 349, no. 6245, pp. 255–260, 2015.
- [25] F. A. Gers, J. Schmidhuber, & F. Cummins, “Learning to forget: Continual prediction with lstm”, *Neural computation*, vol. 12, no. 10, pp. 2451–2471, 2000.
- [26] M. Schuster & K. K. Paliwal, “Bidirectional recurrent neural networks”, *IEEE transactions on Signal Processing*, vol. 45, no. 11, pp. 2673–2681, 1997.
- [27] A. S. Weigend, *Time series prediction: forecasting the future and understanding the past*. Routledge, 2018.
- [28] G. E. Karniadakis, I. G. Kevrekidis, L. Lu, P. Perdikaris, S. Wang, & L. Yang, “Physics-informed machine learning”, *Nature Reviews Physics*, vol. 3, no. 6, pp. 422–440, 2021.
- [29] R. Rai & C. K. Sahu, “Driven by data or derived through physics? a review of hybrid physics guided machine learning techniques with cyber-physical system (cps) focus”, *IEEE Access*, vol. 8, pp. 71 050–71 073, 2020.
- [30] M. Raissi, P. Perdikaris, & G. E. Karniadakis, “Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations”, *Journal of Computational physics*, vol. 378, pp. 686–707, 2019.
- [31] K. Law, A. Stuart, & K. Zygalakis, “Data assimilation”, *Cham, Switzerland: Springer*, vol. 214, p. 52, 2015.
- [32] H. Jaeger, “The “echo state” approach to analysing and training recurrent neural networks with an erratum note”, *Bonn, Germany: German National Research Center for Information Technology GMD Technical Report*, vol. 148, no. 34, p. 13, 2001.
- [33] W. Maass, T. Natschläger, & H. Markram, “Real-time computing without stable states: A new framework for neural computation based on perturbations”, *Neural computation*, vol. 14, no. 11, pp. 2531–2560, 2002.
- [34] L. Gonon & J.-P. Ortega, “Reservoir computing universality with stochastic inputs”, *IEEE transactions on neural networks and learning systems*, vol. 31, no. 1, pp. 100–112, 2019.
- [35] A. Haluszczynski & C. R ath, “Good and bad predictions: Assessing and improving the replication of chaotic attractors by means of reservoir computing”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 10, p. 103 143, 2019.
- [36] G. Tanaka, T. Yamane, J. B. H eroux, et al., “Recent advances in physical reservoir computing: A review”, *Neural Networks*, vol. 115, pp. 100–123, 2019.

- [37] G. Van der Sande, D. Brunner, & M. C. Soriano, “Advances in photonic reservoir computing”, *Nanophotonics*, vol. 6, no. 3, pp. 561–576, 2017.
- [38] H. O. Sillin, R. Aguilera, H.-H. Shieh, et al., “A theoretical and experimental study of neuromorphic atomic switch networks for reservoir computing”, *Nanotechnology*, vol. 24, no. 38, p. 384004, 2013.
- [39] J. Cao, X. Zhang, H. Cheng, et al., “Emerging dynamic memristors for neuromorphic reservoir computing”, *Nanoscale*, vol. 14, no. 2, pp. 289–298, 2022.
- [40] D. Duncan, “Interpreting reservoir computing with pca and exploring physics-informed rc”, Master’s Thesis, Feb. 2023.
- [41] K. Pearson, “VII. note on regression and inheritance in the case of two parents”, *Proceedings of the royal society of London*, vol. 58, no. 347-352, pp. 240–242, 1895.
- [42] A. Haluszczynski, I. Laut, H. Modest, & C. R ath, “Linear and nonlinear market correlations: Characterizing financial crises and portfolio optimization”, *Physical Review E*, vol. 96, no. 6, p. 062315, 2017.
- [43] M. Paluř, V. Albrecht, & I. Dvoř ak, “Information theoretic test for nonlinearity in time series”, *Physics Letters A*, vol. 175, no. 3-4, pp. 203–209, 1993.
- [44] J. Hlinka, D. Hartman, M. Vejmelka, D. Novotn a, & M. Paluř, “Non-linear dependence and teleconnections in climate data: Sources, relevance, nonstationarity”, *Climate dynamics*, vol. 42, no. 7-8, pp. 1873–1886, 2014.
- [45] R. N. Mantegna & H. E. Stanley, *Introduction to econophysics: correlations and complexity in finance*. Cambridge university press, 1999.
- [46] F. Jovanovic, R. N. Mantegna, & C. Schinckus, “When financial economics influences physics: The role of econophysics”, *Available at SSRN 3294548*, 2018.
- [47] G.-J. Wang, C. Xie, S. Chen, J.-J. Yang, & M.-Y. Yang, “Random matrix theory analysis of cross-correlations in the us stock market: Evidence from pearson’s correlation coefficient and detrended cross-correlation coefficient”, *Physica A: statistical mechanics and its applications*, vol. 392, no. 17, pp. 3715–3730, 2013.
- [48] G.-J. Wang, C. Xie, & H. E. Stanley, “Correlation structure and evolution of world stock markets: Evidence from pearson and partial correlation-based networks”, *Computational Economics*, vol. 51, pp. 607–635, 2018.
- [49] R. N. Mantegna & H. E. Stanley, “Scaling behaviour in the dynamics of an economic index”, *Nature*, vol. 376, no. 6535, pp. 46–49, 1995.
- [50] S. Ghashghaie, W. Breymann, J. Peinke, P. Talkner, & Y. Dodge, “Turbulent cascades in foreign exchange markets”, *Nature*, vol. 381, no. 6585, pp. 767–770, 1996.
- [51] J. Benesty, J. Chen, Y. Huang, & I. Cohen, “Pearson correlation coefficient”, in *Noise reduction in speech processing*, Springer, 2009, pp. 1–4.
- [52] A. L. Lloyd, “The coupled logistic map: A simple model for the effects of spatial heterogeneity on population dynamics”, *Journal of Theoretical Biology*, vol. 173, no. 3, pp. 217–230, 1995.
- [53] K. Hlav ckov a-Schindler, M. Paluř, M. Vejmelka, & J. Bhattacharya, “Causality detection based on information-theoretic approaches in time series analysis”, *Physics Reports*, vol. 441, no. 1, pp. 1–46, 2007.
- [54] L. Barnett, A. B. Barrett, & A. K. Seth, “Granger causality and transfer entropy are equivalent for gaussian variables”, *Physical review letters*, vol. 103, no. 23, p. 238701, 2009.
- [55] H. Toda, *Vector autoregression and causality*. Yale University, 1991.

- [56] D. B. Duncan, “Multiple range and multiple f tests”, *biometrics*, vol. 11, no. 1, pp. 1–42, 1955.
- [57] J. Morgan & J. Tatar, “Calculation of the residual sum of squares for all possible regressions”, *Technometrics*, vol. 14, no. 2, pp. 317–325, 1972.
- [58] S. L. Bressler & A. K. Seth, “Wiener–granger causality: A well established methodology”, *Neuroimage*, vol. 58, no. 2, pp. 323–329, 2011.
- [59] A. Shojaie & E. B. Fox, “Granger causality: A review and recent advances”, *Annual Review of Statistics and Its Application*, vol. 9, pp. 289–319, 2022.
- [60] M. Mynter, “Evaluation and extension of the transfer entropy calculus for the measurement of information flows between futures time series during the covid-19 pandemic”, Master’s Thesis, Ludwig-Maximilians-Universität, 2022.
- [61] S. Baur & C. R ath, “Predicting high-dimensional heterogeneous time series employing generalized local states”, *Phys. Rev. Research*, vol. 3, p. 023 215, 2 Jun. 2021. DOI: 10.1103/PhysRevResearch.3.023215. [Online]. Available: <https://link.aps.org/doi/10.1103/PhysRevResearch.3.023215>.
- [62] M. Paluř & M. Vejmelka, “Directionality of coupling from bivariate time series: How to avoid false causalities and missed connections”, *Physical Review E*, vol. 75, no. 5, p. 056 211, 2007.
- [63] F. Takens, “Detecting strange attractors in turbulence”, in *Dynamical Systems and Turbulence, Warwick 1980: proceedings of a symposium held at the University of Warwick 1979/80*, Springer, 2006, pp. 366–381.
- [64] D. M onster, R. Fusaroli, K. Tyl en, A. Roepstorff, & J. F. Sherson, “Causal inference from noisy time-series data—testing the convergent cross-mapping algorithm in the presence of noise and external influence”, *Future Generation Computer Systems*, vol. 73, pp. 52–62, 2017.
- [65] S. Wallot & D. M onster, “Calculation of average mutual information (ami) and false-nearest neighbors (fnn) for the estimation of embedding parameters of multidimensional time series in matlab”, *Frontiers in psychology*, vol. 9, p. 1679, 2018.
- [66] M. B. Kennel, R. Brown, & H. D. Abarbanel, “Determining embedding dimension for phase-space reconstruction using a geometrical construction”, *Physical review A*, vol. 45, no. 6, p. 3403, 1992.
- [67] J. M. McCracken & R. S. Weigel, “Convergent cross-mapping and pairwise asymmetric inference”, *Physical Review E*, vol. 90, no. 6, p. 062 903, 2014.
- [68] P. Krishna & A. K. Tangirala, “Inferring direct causality from noisy data using convergent cross mapping”, in *2019 58th Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE)*, IEEE, 2019, pp. 1523–1528.
- [69] G. Feng, K. Yu, Y. Wang, Y. Yuan, & P. M. Djuri c, “Improving convergent cross mapping for causal discovery with gaussian processes”, in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2020, pp. 3692–3696.
- [70] A. T. Clark, H. Ye, F. Isbell, et al., “Spatial convergent cross mapping to detect causal relationships from short time series”, *Ecology*, vol. 96, no. 5, pp. 1174–1181, 2015.
- [71] *Redm: An r package for empirical dynamic modeling and convergent cross mapping*, <https://ha0ye.github.io/rEDM/articles/rEDM.html>, Accessed: 2023-11-08.
- [72] C. R ath, M. Gliozzi, I. Papadakis, & W. Brinkmann, “Revisiting algorithms for generating surrogate time series”, *Physical review letters*, vol. 109, no. 14, p. 144 101, 2012.
- [73] E. O. Brigham, *The fast Fourier transform and its applications*. Prentice-Hall, Inc., 1988.

- [74] D. Prichard & J. Theiler, “Generating surrogate data for time series with several simultaneously measured variables”, *Physical review letters*, vol. 73, no. 7, p. 951, 1994.
- [75] E. Kasuya, “On the use of r and r squared in correlation and regression”, Wiley Online Library, Tech. Rep., 2019.
- [76] M. Maziarz, “A review of the granger-causality fallacy”, *The journal of philosophical economics: Reflections on economic and social issues*, vol. 8, no. 2, pp. 86–105, 2015.
- [77] S. J. Brown, W. Goetzmann, R. G. Ibbotson, & S. A. Ross, “Survivorship bias in performance studies”, *The Review of Financial Studies*, vol. 5, no. 4, pp. 553–580, 1992.
- [78] J. B. DeLong & K. Magin, *A short note on the size of the dot-com bubble*, 2006.
- [79] H. P. Krishnan, A. Bennington, H. P. Krishnan, & A. Bennington, “The vix “volmaggedon”, with exchange-traded notes destabilizing the market”, *Market Tremors: Quantifying Structural Risks in Modern Financial Markets*, pp. 83–119, 2021.
- [80] E. Zivot, J. Wang, E. Zivot, & J. Wang, “Rolling analysis of time series”, *Modeling financial time series with S-Plus®*, pp. 299–346, 2003.
- [81] G. Vidyamurthy, *Pairs Trading: quantitative methods and analysis*. John Wiley & Sons, 2004, vol. 217.
- [82] J. C. Hull, *Options, futures and other derivatives*, 2019.
- [83] H. M. Markowitz, “Foundations of portfolio theory”, *The journal of finance*, vol. 46, no. 2, pp. 469–477, 1991.
- [84] W. F. Sharpe, “The sharpe ratio”, *Streetwise—the Best of the Journal of Portfolio Management*, vol. 3, pp. 169–85, 1998.
- [85] D. Duffie & J. Pan, “An overview of value at risk”, *Journal of derivatives*, vol. 4, no. 3, pp. 7–49, 1997.
- [86] E. F. Fama & K. R. French, “The capm is wanted, dead or alive”, *The Journal of Finance*, vol. 51, no. 5, pp. 1947–1958, 1996.
- [87] D. F. Williamson, R. A. Parker, & J. S. Kendrick, “The box plot: A simple visual method to interpret data”, *Annals of internal medicine*, vol. 110, no. 11, pp. 916–921, 1989.
- [88] X. Wan, W. Wang, J. Liu, & T. Tong, “Estimating the sample mean and standard deviation from the sample size, median, range and/or interquartile range”, *BMC medical research methodology*, vol. 14, no. 1, pp. 1–13, 2014.
- [89] J. L. Breeden & A. Hübler, “Reconstructing equations of motion from experimental data with unobserved variables”, *Physical Review A*, vol. 42, no. 10, p. 5817, 1990.
- [90] T. Eisenhammer, A. Hübler, N. Packard, & J. S. Kelso, “Modeling experimental time series with ordinary differential equations”, *Biological cybernetics*, vol. 65, no. 2, pp. 107–112, 1991.
- [91] S. L. Brunton, J. L. Proctor, & J. N. Kutz, “Discovering governing equations from data by sparse identification of nonlinear dynamical systems”, *Proceedings of the national academy of sciences*, vol. 113, no. 15, pp. 3932–3937, 2016.
- [92] K. Champion, B. Lusch, J. N. Kutz, & S. L. Brunton, “Data-driven discovery of coordinates and governing equations”, *Proceedings of the National Academy of Sciences*, vol. 116, no. 45, pp. 22 445–22 451, 2019.
- [93] I. P. Mariño & J. Míguez, “An approximate gradient-descent method for joint parameter estimation and synchronisation of coupled chaotic systems”, *Phys. Lett. A*, vol. 351, no. 4-5, pp. 262–267, 2006. DOI: <https://doi.org/10.1016/j.physleta.2005.11.005>.
- [94] H. D. I. Abarbanel, D. R. Creveling, R. Farsian, & M. Kostuk, “Dynamical state and parameter estimation”, *SIAM J. Appl. Dyn. Syst.*, vol. 8, no. 4, 2009.

- [95] D. Eroglu, J. S. W. Lamb, & T. Pereira, “Synchronisation of chaos and its applications”, *Contemp. Phys.*, vol. 58, no. 2, pp. 207–243, 2017. DOI: <https://doi.org/10.1080/00107514.2017.1345844>.
- [96] D. P. Kingma & J. L. Ba, “Adam: A Method for Stochastic Optimization”, in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Y. Bengio & Y. LeCun, Eds., 2015.
- [97] S. J. Reddi, S. Kale, & S. Kumar, “On the Convergence of Adam and Beyond”, *CoRR*, 2019.
- [98] R. Thomas, “Deterministic chaos seen in terms of feedback circuits: Analysis, synthesis, “labyrinth chaos””, *IJBC*, vol. 9, no. 10, pp. 1889–1905, 1999.
- [99] J. C. Sprott, “A dynamical system with a strange attractor and invariant tori”, *Phys. Lett. A*, vol. 378, pp. 1361–1363, 20 2014.
- [100] S. Dadras & H. R. Momeni, “A novel three-dimensional autonomous chaotic system generating two, three and four-scroll attractors”, *Phys. Lett. A*, vol. 373, pp. 3637–3642, 40 2009.
- [101] O. E. Rössler, “An equation for continuous chaos”, *Phys. Lett. A*, vol. 57, pp. 397–398, 5 1976.
- [102] J. C. Sprott, *Elegant Chaos*. World Scientific Publishing, 2010.
- [103] E. N. Lorenz, “Atmospheric models as dynamical systems”, in *Perspectives in Nonlinear Dynamics: 28 - 30 May 1985 Naval Surface Weapons Center*, M. F. Shlesinger, R. Cawley, A. W. Saenz, & W. Zachary, Eds., World Scientific Publishing, 1986.
- [104] L. Pan, W. Zhou, J. Fang, & D. Li, “A new three-scroll unified chaotic system coined”, *Int. J. Nonlinear Sci.*, vol. 10, no. 4, pp. 462–474, 2010.
- [105] E. N. Lorenz, “Predictability — a problem partly solved”, in *Predictability of Weather and Climate*, T. Palmer & R. Hagedorn, Eds. Cambridge University Press, 2006.
- [106] A. Creswell, T. White, V. Dumoulin, K. Arulkumaran, B. Sengupta, & A. A. Bharath, “Generative adversarial networks: An overview”, *IEEE signal processing magazine*, vol. 35, no. 1, pp. 53–65, 2018.
- [107] J. Zhang, Y. Wang, P. Molino, L. Li, & D. S. Ebert, “Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models”, *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 364–373, 2018.
- [108] R. Roscher, B. Bohn, M. F. Duarte, & J. Garcke, “Explainable machine learning for scientific insights and discoveries”, *Ieee Access*, vol. 8, pp. 42 200–42 216, 2020.
- [109] D. Prokhorov, “Echo state networks: Appeal and challenges”, in *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, IEEE, vol. 3, 2005, pp. 1463–1466.
- [110] A. Haluszczynski, J. Aumeier, J. Herteux, & C. Räth, “Reducing network size and improving prediction stability of reservoir computing”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 6, p. 063 136, 2020.
- [111] G. Holzmam, “Reservoir computing: A powerful black-box framework for nonlinear audio processing”, in *International Conference on Digital Audio Effects (DAFx)*, Citeseer, 2009.
- [112] D. J. Watts & S. H. Strogatz, “Collective dynamics of ‘small-world’ networks”, *nature*, vol. 393, no. 6684, pp. 440–442, 1998.
- [113] R. Albert & A.-L. Barabási, “Statistical mechanics of complex networks”, *Reviews of modern physics*, vol. 74, no. 1, p. 47, 2002.
- [114] A. D. Broido & A. Clauset, “Scale-free networks are rare”, *Nature communications*, vol. 10, no. 1, p. 1017, 2019.

- [115] M. Gerlach & E. G. Altmann, “Testing statistical laws in complex systems”, *Physical Review Letters*, vol. 122, no. 16, p. 168301, 2019.
- [116] A. Griffith, A. Pomerance, & D. J. Gauthier, “Forecasting chaotic systems with very low connectivity reservoir computers”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 12, p. 123108, 2019.
- [117] T. L. Carroll & L. M. Pecora, “Network structure effects in reservoir computers”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 29, no. 8, p. 083130, 2019.
- [118] R. J. Glauber, “Time-dependent statistics of the ising model”, *Journal of mathematical physics*, vol. 4, no. 2, pp. 294–307, 1963.
- [119] E. Boltt, “On explaining the surprising success of reservoir computing forecaster of chaos? the universal machine learning dynamical system with contrast to var and dmd? a3b2 show [feature]?_i”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 31, no. 1, p. 013108, 2021.
- [120] D. J. Gauthier, E. Boltt, A. Griffith, & W. A. Barbosa, “Next generation reservoir computing”, *Nature communications*, vol. 12, no. 1, p. 5564, 2021.
- [121] C. L. Weise, “The asymmetric effects of monetary policy: A nonlinear vector autoregression approach”, *Journal of Money, Credit and Banking*, pp. 85–108, 1999.
- [122] A. G. Hart, J. L. Hook, & J. H. Dawes, “Echo state networks trained by tikhonov least squares are $l_2(\mu)$ approximators of ergodic dynamical systems”, *Physica D: Nonlinear Phenomena*, vol. 421, p. 132882, 2021.
- [123] J. Pathak, Z. Lu, B. R. Hunt, M. Girvan, & E. Ott, “Using machine learning to replicate chaotic attractors and calculate lyapunov exponents from data”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 27, no. 12, p. 121102, 2017.
- [124] Z. Lu, B. R. Hunt, & E. Ott, “Attractor reconstruction by machine learning”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 6, p. 061104, 2018.
- [125] P. Erdos, A. Rényi, et al., “On the evolution of random graphs”, *Publ. Math. Inst. Hung. Acad. Sci.*, vol. 5, no. 1, pp. 17–60, 1960.
- [126] J. Herteux & C. R ath, “Breaking symmetries of the reservoir equations in echo state networks”, *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 30, no. 12, p. 123142, 2020.
- [127] A. E. Hoerl & R. W. Kennard, “Ridge regression: Applications to nonorthogonal problems”, *Technometrics*, vol. 12, no. 1, pp. 69–82, 1970.
- [128] S. L. Brunton & J. N. Kutz, *Data-driven science and engineering: Machine learning, dynamical systems, and control*. Cambridge University Press, 2022.
- [129] K. Kaheman, J. N. Kutz, & S. L. Brunton, “Sindy-pi: A robust algorithm for parallel implicit sparse identification of nonlinear dynamics”, *Proceedings of the Royal Society A*, vol. 476, no. 2242, p. 20200279, 2020.
- [130] U. Fasel, J. N. Kutz, B. W. Brunton, & S. L. Brunton, “Ensemble-sindy: Robust sparse model discovery in the low-data, high-noise limit, with active learning and control”, *Proceedings of the Royal Society A*, vol. 478, no. 2260, p. 20210904, 2022.
- [131] M. T. Rosenstein, J. J. Collins, & C. J. De Luca, “A practical method for calculating largest lyapunov exponents from small data sets”, *Physica D: Nonlinear Phenomena*, vol. 65, no. 1-2, pp. 117–134, 1993.
- [132] P. Grassberger & I. Procaccia, “Dimensions and entropies of strange attractors from a fluctuating dynamics approach”, *Physica D: Nonlinear Phenomena*, vol. 13, no. 1-2, pp. 34–54, 1984.
- [133] B. P. Bezruchko & D. A. Smirnov, *Extracting knowledge from time series: an introduction to nonlinear empirical modeling*. Springer, 2010.

- [134] C. C. Paige, “Bidiagonalization of matrices and solution of linear equations”, *SIAM Journal on Numerical Analysis*, vol. 11, no. 1, pp. 197–209, 1974.
- [135] S. G. Kobourov, “Spring embedders and force directed graph drawing algorithms”, *arXiv preprint arXiv:1201.3011*, 2012.
- [136] D. Yu, H. Wang, P. Chen, & Z. Wei, “Mixed pooling for convolutional neural networks”, in *Rough Sets and Knowledge Technology: 9th International Conference, RSKT 2014, Shanghai, China, October 24-26, 2014, Proceedings 9*, Springer, 2014, pp. 364–375.
- [137] J. C. Sprott, “A dynamical system with a strange attractor and invariant tori”, *Physics Letters A*, vol. 378, pp. 1361–1363, 20 2014.
- [138] D. Prosperino, H. Ma, & C. R ath, “A modern and generalized gradient-descent method for estimating parameters of complex systems using synchronization”, *Submitted to Chaos: An Interdisciplinary Journal of Nonlinear Science*, 2023.
- [139] A. Haluszczynski, D. Koeglmayr, & C. R ath, “Controlling dynamical systems to complex target states using machine learning: Next-generation vs. classical reservoir computing”, in *2023 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2023, pp. 1–7.
- [140] J. C. Sprott, *Elegant chaos: algebraically simple chaotic flows*. World Scientific, 2010.
- [141] V. P. Thoai, M. S. Kahkeshi, V. V. Huynh, A. Ouannas, & V.-T. Pham, “A nonlinear five-term system: Symmetry, chaos, and prediction”, *Symmetry*, vol. 12, no. 5, p. 865, 2020.