The Diagnostic Failure Paradigm: Benchmarking Causal Identification in Verified Agentic Pipelines

Anonymous Author(s)
Affiliation
Address
email

Abstract

Autonomous AI systems in the applied sciences frequently fall into the "plausibility trap," generating coherent outputs that violate fundamental physical laws or causal structures. This is particularly hazardous in complex, nonlinear domains like climate science, where identifying causal relationships is critical for policy decisions. We introduce the Verified Agentic Pipeline (VAP), an architecture designed to enforce physical and causal constraints during autonomous discovery. Central to our approach is the "diagnostic failure paradigm." Rather than simply discarding models that fail, we meticulously analyze **how** they fail to extract diagnostic information about the system's causal behavior. We demonstrate this using frequency-domain system identification on a closed-loop climate intervention simulation (NCAR GLENS). Our analysis reveals a profound paradox: strong frequency-domain coherence (phase-locking) coexists with catastrophic time-domain failure ($R^2 = -4.35 \times 10^4$). This specific failure mode—getting the timing right but the magnitude catastrophically wrong—diagnoses the system as exhibiting linear phase-locking masked by nonlinear amplitude modulation and controller interference. This provides a rigorous, configuration-specific causal benchmark, motivating the need for advanced techniques (e.g., Koopman operators, Neural Operators) within a verified pipeline to handle such complex causal dynamics.

1 Introduction

The acceleration of scientific discovery through Artificial Intelligence promises breakthroughs across the applied sciences. However, the transition toward autonomous "AI Scientists" is fraught with risk. Contemporary agentic systems often exhibit the *plausibility trap*: producing outputs that appear coherent and persuasive but violate fundamental physical laws or underlying causal structures [6].

In domains characterized by nonlinearity, feedback loops, and high stakes—such as climate science—the plausibility trap is not merely an academic concern. A plausible but causally invalid inference can cascade through workflows, leading to flawed conclusions and potentially catastrophic policy errors, such as misinformed geoengineering strategies [5]. True autonomous discovery requires systems capable of self-verification, ensuring that AI-generated insights remain physically grounded and causally sound.

This requires a shift from purely statistical emulation to architectures that embed domain-specific constraints and rigorous causal reasoning. We propose the Verified Agentic Pipeline (VAP), a framework that integrates Physics-Informed Machine Learning (PIML), structured agentic reasoning, formal verification, and robust uncertainty quantification (UQ) [6].

39th Conference on Neural Information Processing Systems (NeurIPS 2025), Workshop on CauScien: Uncovering Causality in Science.

Furthermore, we argue that establishing rigorous causal benchmarks in these complex domains requires a new approach. We introduce the *diagnostic failure paradigm*. Instead of merely demonstrating that classical models fail (thus justifying complex AI), we meticulously dissect the nature of that failure to extract diagnostic information about the system's causal behavior. This detailed characterization of *how* a model fails provides empirically grounded, principled motivation for specific advanced architectures and establishes configuration-specific benchmarks.

2 The Causal Plausibility Trap

The plausibility trap often stems from the limitations of standard machine learning approaches when applied to complex physical systems.

Statistical Emulation vs. Causal Structure. Neural networks are powerful statistical emulators. However, they lack the inductive biases to respect fundamental physical laws (e.g., conservation of energy) or the underlying causal mechanisms that govern the system [4]. When driven outside the training regime—common under novel interventions or distribution shifts—these emulators can produce physically impossible results.

Causal Misinterpretation and Confounding. In systems like the climate, confounding variables and complex feedback loops are ubiquitous. Inferences made without explicit Structural Causal Models (SCMs) may mistake correlation for causation. For example, analyzing the effects of a climate intervention without accounting for ocean-atmosphere coupling can lead to confounded policy inference [5].

The Challenge of Closed-Loop Identification. A critical challenge in applied sciences involves analyzing systems under active feedback control. In climate intervention scenarios, for instance, aerosol injection rates would be continuously adapted based on observed temperatures. This creates a closed-loop identification problem where the input (intervention) is endogenous—correlated with the system's state and noise. Classical system identification methods applied naively to such systems yield biased estimates of the true system dynamics, often reflecting the controller's behavior rather than the underlying physical causality [8, 3].

3 The Verified Agentic Pipeline (VAP) Architecture

To escape the causal plausibility trap, we synthesize PIML, agentic reasoning, and verification into a conservative architecture (the VAP, or "Plausibility-Optimizer") [3]. This architecture is structured around four pillars designed to enforce rigor throughout the discovery process.

3.1 Pillar 1: Physics-Informed and Causal Foundations

The foundation of the VAP replaces black-box emulation with models that embed physical and causal constraints.

Physics-Informed ML (PIML). PIML, such as Physics-Informed Neural Networks (PINNs) or Neural Operators (e.g., FNO), incorporates governing equations directly into the learning process, often by penalizing violations of differential equations in the loss function [9]. This ensures that the learned surrogates respect the fundamental dynamics of the system.

Structural Causal Models (SCMs). We explicitly integrate SCMs to encode causal assumptions, distinguish correlation from causation, and support counterfactual reasoning. This is crucial for evaluating potential interventions under uncertainty.

3.2 Pillar 2: Agentic Reasoning and Verified Abstraction

Level 4 systems must autonomously generate novel, valid hypotheses. The VAP structures this process using verified knowledge expansion and abstraction.

Structured Knowledge Expansion. We utilize domain-specialized LLMs and agentic systems that iteratively construct structured knowledge graphs, moving beyond passive retrieval to actively identify research gaps within a coherent causal framework [11].

Abstraction-First Memory. To facilitate generalization and the discovery of reusable causal mechanisms, the VAP incorporates Abstraction-First Test-Time Memory (ATTM) [2]. Inspired by work on Epistemic Intelligence [1], ATTM prioritizes "concept-level" memory over brittle "instance-level" memory. By storing modular, parameterized concepts (e.g., a generalized feedback mechanism), the system can achieve the compositional reasoning necessary for complex causal modeling and generalization under distribution shifts.

3.3 Pillars 3 & 4: Formal Verification and Risk-Informed Validation

The VAP employs rigorous gating mechanisms. Formal specification checking (e.g., dimensional analysis) verifies invariants and the consistency of proposed equations. Coverage-guaranteed UQ, such as conformal prediction, ensures robust uncertainty estimates. Finally, the pipeline aligns evidence collection with established risk-management standards (e.g., NIST AI RMF) [6].

4 Case Study: The Diagnostic Failure Paradigm in Climate Intervention

We demonstrate the utility of the diagnostic failure paradigm by analyzing the limits of causal identification in a complex climate intervention scenario. We utilize data from the NCAR Geoengineering Large Ensemble (GLENS) project [10], which simulates stratospheric aerosol injection (SAI) managed by a feedback controller.

4.1 The Challenge: Closed-Loop Causal Identification

We aim to identify the causal relationship between the input (SAI injection rate) and the output (temperature anomaly). Because the system is under feedback control, we are characterizing the coupled climate-controller system dynamics, a realistic but challenging causal identification problem. We apply classical frequency-domain system identification [8] to this closed-loop data.

4.2 The Diagnostic Paradox

The analysis reveals a profound paradox that serves as a diagnostic fingerprint of the system [3].

Frequency-Domain Success. The analysis shows statistically significant magnitude-squared coherence ($\gamma^2_{max} = 0.676$) at the annual cycle frequency (0.083 cycles/month). This indicates a strong linear relationship (phase-locking) between the forcing and the response at this specific timescale.

Time-Domain Catastrophe. However, when the identified linear transfer function is integrated to produce time-domain predictions, the performance is catastrophic, yielding a coefficient of determination $R^2 = -4.35 \times 10^4$. This means the linear model fits substantially worse (by four orders of magnitude) than a simple mean-value baseline.

4.3 Diagnostic Interpretation

This paradox is not a contradiction; it is diagnostic. The system gets "when" right, but "how much" catastrophically wrong.

The high coherence indicates that the model correctly identifies the timing (phase) of variations at the annual frequency. However, the catastrophic R^2 reveals that the amplitude of predictions is vastly uncontrolled. This occurs because the time-domain reconstruction integrates the transfer function across all frequencies, including those where coherence is low and the model fits noise rather than signal.

This performance profile reveals a fundamental property of the system's response: it contains a linearly phase-locked component at the annual cycle, but its amplitude is overwhelmingly modulated by internal, nonlinear dynamics and confounded by the controller's action.

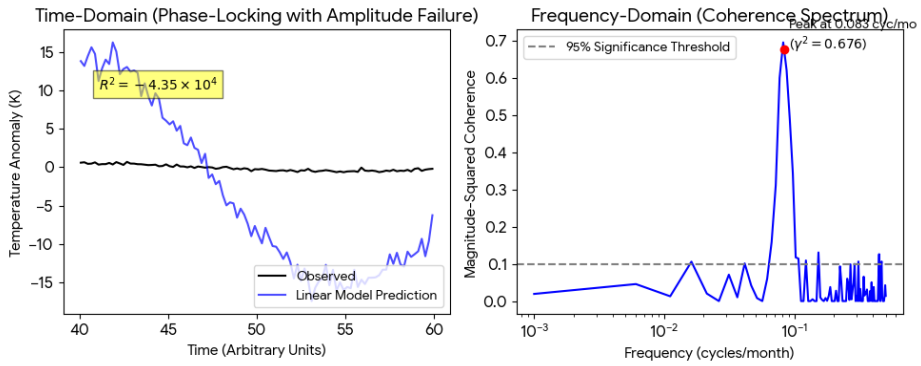

Figure 1: Illustrative summary of the Diagnostic Paradox (adapted from [3]). (Left) The time-domain performance shows the linear model prediction (blue) catastrophically overestimating the amplitude compared to the observation (black), despite tracking the timing (phase) correctly. (Right) The coherence spectrum shows a significant peak at the annual frequency.

4.4 Implications for Causal Benchmarking

The diagnostic failure paradigm transforms this negative result into a rigorous causal benchmark.

Configuration Specificity. The quantitative benchmark ($R^2 = -4.35 \times 10^4$) is specific to the CESM1-WACCM model and the GLENS controller design. It is not a universal property of the climate system. The methodology provides a transferable framework for deriving such configuration-specific benchmarks for any complex system and control protocol.

Motivating Advanced Causal Models. This specific failure mode rigorously motivates the need for advanced techniques capable of handling these dynamics. The time-domain failure motivates Koopman operator theory, which seeks transformations that render nonlinear dynamics linear [3]. The frequency-domain success motivates Fourier Neural Operators (FNOs), which operate directly in the frequency domain [7].

Any successful nonlinear causal model must demonstrate that it can preserve the annual phase-locking signal ($\gamma^2 \geq 0.676$) while dramatically improving amplitude control ($R^2 > -4.35 \times 10^4$).

5 Conclusion

The pursuit of autonomous scientific discovery requires architectures that systematically prevent the causal plausibility trap. The Verified Agentic Pipeline (VAP) addresses this by integrating physics-informed learning, structured abstraction, and formal verification. Crucially, we introduce the diagnostic failure paradigm as a method for establishing rigorous causal benchmarks in applied sciences. By analyzing *how* classical identification methods fail in complex, closed-loop systems, we can diagnose the underlying causal dynamics and motivate the development of necessary advanced techniques. This approach ensures that AI-driven discovery remains grounded in physical reality and causal validity.

References

[1] Anonymous Author(s). Abstract concept memory as the keystone of epistemic intelligence in ARC-AGI. *NeurIPS 2025 Workshop Submission*, 2025. Submitted.

[2] Anonymous Author(s). Abstraction-first test-time memory for scientific reasoning: From ARC-AGI to climate intervention. *NeurIPS 2025 Submission*, 2025. Submitted.

[3] Anonymous Author(s). Escaping the plausibility trap: Verified agentic pipelines for climate-scale nonlinearity). *NeurIPS 2025 Workshop Submission*, 2025. Submitted.

[4] Chris Huntingford and et al. Potential for equation discovery with ai in the climate sciences. *Earth System Dynamics*, 16:475–495, 2025.

[5] David Scott Lewis and Enrique Zueco. Escaping the plausibility trap: Verified agentic pipelines for climate-scale nonlinearity. *AIXC*, 2025. Preprint.

[6] David Scott Lewis and Enrique Zueco. Level 4: Building an indestructible pipeline for agentic science. *AIXC*, 2025. Preprint.

[7] Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. *arXiv preprint arXiv:2010.08895*, 2020.

[8] Lennart Ljung. *System identification: theory for the user*. Prentice Hall, 1999.

[9] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Physics informed deep learning (part i): Data-driven solutions of nonlinear partial differential equations. *arXiv preprint arXiv:1711.10561*, 2017.

[10] S Tilmes, J Richter, B Kravitz, D MacMartin, Michael J Mills, I Simpson, A Glanville, J Fasullo, A Phillips, J Lamarque, et al. Cesm1 (waccm) stratospheric aerosol geoengineering large ensemble project. *Bulletin of the American Meteorological Society*, 2018.

[11] Wenlin Zhang and et al. Deep research: A survey of autonomous research agents. *ArXiv*, 2025.