ORCD: ODE-residual causal decomposition for heterogeneous dynamical systems

Anonymous authors
Paper under double-blind review

Abstract

Causal discovery in multivariate time-series data is fundamentally limited by the entanglement of slow-scale invariant physical dynamics with fast-scale instantaneous causal forcing; a regime where existing approaches either require non-Gaussian noise, which is frequently violated in environmental and health systems, or lack a mechanism to decouple the two generative processes. We propose *ODE-residual causal decomposition* (ORCD), a two-step framework for multi-environment time-series data. A single global SINDy ODE is fit on pooled data; the resulting structured residuals are heteroscedastic across environments by construction, as local dynamics that the parsimonious model cannot represent, are forced into the residual. We prove formally that this heteroscedasticity suffices for unique DAG identifiability under Gaussian noise, bypassing the non-Gaussianity assumption that restricts LiNGAM-type methods, and that the global fit is the theoretically necessary mechanism for generating it. Causal discovery is then performed on the augmented space of original variables and whitened residuals, which serve as observed proxy covariates for environment-specific forcing. ORCD is algorithm-agnostic: across four discovery algorithms, it strictly reduces SHD in every synthetic benchmark setting (48/48), with a median reduction of 21% and up to 43% (Tabu) and 72% (LiNGAM) in the most challenging settings. On real-world data linking air pollution and cardiovascular mortality across six English regions (1981–2014), ORCD substantially improves structural accuracy under non-i.i.d. temporal hold-out (F1 = 0.973–0.991 vs. 0.895–0.963) while recovering richer causal graphs that generalise spatially: subtracting the residuals from parameter learning collapses CVD60 Spearman correlation from $r = 0.62$–0.66 to near zero ($r \approx 0.01$–0.04), confirming that the residuals encode genuine causal signals that transfer across unseen geographical environments.

1 Introduction

Identifying the directed acyclic graph (DAG) that characterises the causal structure of observational data is a pivotal task in modern causal machine learning (Pearl, 2009). In complex systems where deterministic physical dynamics are inextricably coupled with external forcing, this task is made substantially harder by the entanglement of two distinct generative regimes: slow-scale invariant laws governing the deterministic backbone of the system (e.g. seasonal atmospheric circulation driving background pollution levels), and fast-scale instantaneous causal forcing driven by unmodelled external inputs (e.g. from acute pollution episodes and their short-term cardiovascular effects, to measurement errors inherent in sensor arrays and spatial aggregation). Standard structural causal models (SCMs) cannot decouple these two regimes, assuming a static causal structure across all observations and leading to unreliable structural inferences (Peters et al., 2022; Runge et al., 2019).

The key difficulty is that current discovery methods are ill-equipped to unravel this entanglement. Non-Gaussianity-based approaches such as LiNGAM (Shimizu et al., 2006; 2011) and its variants (Suzuki & Yang, 2024) rely on distributional assumptions that are frequently violated: large-scale environmental and health data often exhibit approximate Gaussianity (Papanastasiou et al., 2023; 2025; Zhu et al., 2017), which precludes structural identifiability for methods dependent on non-normal noise. Score-based baselines such as NOTEARS (Zheng et al., 2018) and CGP-CDE (Dhir et al., 2024) capture complex functional forms

but have no mechanism to decouple invariant, slow-scale dynamics from instantaneous causal forcing; they consequently conflate physical laws with causal effects and yield entangled, often spurious structures.

We propose the *ODE-residual causal decomposition* (ORCD) framework to address both limitations simultaneously. ORCD integrates symbolic ODE discovery (SINDy; Brunton et al. 2016; Kaptanoglu et al. 2022) with Bayesian network structure learning in a principled two-step pipeline. A *single* global SINDy model is fit on data pooled across all environments simultaneously. This global fit captures the dominant shared dynamics and forces any local, environment-specific dynamics into the structured residuals $R_i = \dot{X}_i - \Theta(\mathbf{X})\hat{\boldsymbol{\Xi}}_i$, where $\dot{X}_i$ denotes the time derivative of variable i, $\Theta(\mathbf{X}) \in \mathbb{R}^{T\times D}$ is the library of D candidate basis functions evaluated on the data, and $\hat{\boldsymbol{\Xi}}_i \in \mathbb{R}^D$ is the sparse coefficient vector identified by SINDy (formal definitions in Section 3.1). Because the degree of local dynamic misfit varies across environments, these residuals are heteroscedastic across environments *by construction*, a property we prove is sufficient for unique DAG identifiability under Gaussian noise, invoking the multi-environment sufficient conditions of Xu et al. (2022) and Duong & Nguyen (2023). The residuals are then treated as explicit, observed proxy variables for confounding in an expanded SCM, and joint causal discovery is performed on the augmented $2N$-dimensional space.

Fitting separate per-environment ODEs, the seemingly natural alternative, would absorb local dynamics into per-environment coefficient estimates, driving the residual variance towards a common environment-independent level and eliminating the identifiability signal. The global fit is therefore not a computational shortcut but a theoretically necessary commitment, established formally in Corollary 3.7.

Our main contributions are:

1. A novel decompositional framework that separates slow-scale invariant dynamics from fast-scale instantaneous causal forcing using a single global ODE fit, whose structured residuals serve as observed confounding proxies in an expanded SCM (Section 4).
2. Formal identifiability theory establishing that *(i)* the global ODE fit mechanically generates residuals that are heteroscedastic across environments by construction under a single verifiable assumption (Proposition 3.6); *(ii)* this heteroscedasticity is sufficient for unique DAG identifiability under Gaussian noise, bypassing the non-Gaussianity assumptions that restrict LiNGAM-type methods (Theorem 3.10); and *(iii)* the global fit is not a computational convenience but a theoretical necessity; local per-environment fits provably eliminate the identifiability signal, as the SINDy candidate function library grows (Corollary 3.7).
3. A model-agnostic causal discovery pipeline that bridges domain knowledge and data-driven causal discovery by fusing sparse ODE-derived residuals with elementary physical priors, encoding expert knowledge as directional constraints between weather, pollution, and health, to recover a jointly consistent, physically interpretable causal network; demonstrated to yield structural and predictive gains uniformly across four algorithmically distinct discovery methods (Section 4).
4. Empirical validation on twelve synthetic settings and on large real-world pollution and cardiovascular mortality data across six English regions spanning 34 years, demonstrating consistent structural improvement and cross-environment causal generalisability (Section 5).

2 Background

2.1 Structural causal models and causal discovery

A *structural causal model* over variables $\{X_1, \ldots, X_N\}$ consists of structural equations $X_i = f_i(\mathrm{Pa}(X_i), \varepsilon_i)$, where $\mathrm{Pa}(X_i)$ denotes the causal parents in a DAG $\mathcal{G}$ and $\{\varepsilon_i\}$ are mutually independent exogenous noise variables (Pearl, 2009). Causal discovery aims to recover $\mathcal{G}$ from observational data, even in the absence of interventions, to represent the underlying data generating process (DGP). Without additional assumptions, only the *Markov equivalence class*, the set of DAGs encoding the same conditional independencies, is identifiable (Colombo & Maathuis, 2014).

Three main routes to unique identifiability beyond Markov equivalence have been established in the literature. Non-Gaussian noise in linear models (LiNGAM) identifies the unique DAG for the DGP (Shimizu et al., 2006). Linear Gaussian models with equal error variances are identifiable (Peters & Bühlmann, 2013). Most relevant here, *heteroscedastic noise* whose variance varies across environments suffices for unique identifiability even under Gaussian noise and without any linearity assumption (Xu et al., 2022; Duong & Nguyen,

2023). ORCD exploits this third route, generating the required heteroscedasticity mechanically from the global ODE fit rather than assuming it a priori.

Causal discovery algorithms fall into three classes. *Constraint-based* methods such as PC (Colombo & Maathuis, 2014) test conditional independencies but can only recover Markov equivalence classes. *Score-based* methods such as NOTEARS (Zheng et al., 2018) and DAGMA (Bello et al., 2022) optimise a continuous relaxation of the DAG acyclicity constraint but provide no mechanism to separate invariant dynamics from causal forcing. *Functional causal model* approaches such as Direct LiNGAM (Shimizu et al., 2011) achieve identifiability under distributional assumptions, often violated in practice (Papanastasiou et al., 2023; 2025).

2.2 ODE discovery and SINDy

The Sparse Identification of Nonlinear Dynamical Systems (SINDy) algorithm (Brunton et al., 2016) identifies governing equations by solving $\dot{\mathbf{X}}_i \approx \Theta(\mathbf{X})\boldsymbol{\Xi}_i$, where $\Theta(\mathbf{X}) \in \mathbb{R}^{T\times D}$ is a library of candidate basis functions and $\boldsymbol{\Xi}_i \in \mathbb{R}^D$ is a sparse coefficient vector. We use the SR3 optimiser (Champion et al., 2020) for numerical stability on large datasets. Zhang & Schaeffer (2019) prove that the sequential-thresholding algorithm converges to a local minimiser in at most D iterations (where D is the library size; (Definition 3.1)), and that every global minimiser is attainable as such a fixed point, guaranteeing that the sequential-thresholding algorithm always terminates at a fixed point, so the structured residuals of Definition 3.1 are well-defined for any output of the algorithm.

2.3 Related work

Heteroscedastic identifiability. Xu et al. (2022) establish that heteroscedastic noise across environments identifies causal direction without distributional assumptions on the noise. Duong & Nguyen (2023) extend this to multivariate settings. Both results are invoked in ORCD not by treating the ODE residuals as exogenous noise terms, which would require independence of $\delta_{i,k,t}$ (the local dynamic misfit, defined formally in Definition 3.5) from $\mathrm{Pa}(X_i)$, a condition that does not hold in general; but via the *multi-environment distribution shift* criterion: it suffices that the observed joint distribution of the augmented system $(X_i, \widetilde{R}_i)$, where $\widetilde{R}_i$ denotes the whitened residuals defined formally in Section 3.3, varies across environments in a structured way, a condition guaranteed mechanically by the global ODE fit (Propositions 3.6 and 3.8).

Score-based continuous optimisation. NOTEARS (Zheng et al., 2018), DAGMA (Bello et al., 2022), and CGP-CDE (Dhir et al., 2024) optimise continuous relaxations of the DAG acyclicity constraint. They capture complex functional forms but have no mechanism to decouple invariant physical dynamics from causal forcing; moreover, NOTEARS and CGP-CDE, as continuous optimisation methods, do not natively support hard structural constraints such as the domain-knowledge whitelists and blacklists used in ORCD.

Causal discovery in dynamical systems. Runge et al. (2019) survey temporal causal discovery with a focus on Granger-type formulations. Scutari et al. (2025) propose a spatio-temporal approach, but its additive-effects form limits flexibility. Lenatti et al. (2025) embed ODEs as local distributions in a DAG to enable counterfactual inference but without a discovery procedure. ORCD is, to our knowledge, the first framework to use a global ODE fit explicitly to generate identifiability-enabling residuals for multivariate causal discovery.

Nonstationarity- and heteroscedasticity-aware methods. FANTOM (Rahmani et al., 2025) handles heteroscedastic noise via regime segmentation, Regime-PCMCI (Saggioro et al., 2020) applies per-regime structure learning, and Löwe et al. (2022) learn graphs across samples with shared dynamics without identifiability guarantees. ORCD instead mechanically *generates* heteroscedasticity from a global ODE fit, yielding formal identifiability (Theorem 3.10) without regime labels, at the cost of requiring a shared dynamic law.

Decomposition-based causal discovery. Unlike methods separating time series into trend/seasonal/residual components as a stationarity preprocessing step (Ferdous & Gani, 2026), ORCD promotes residuals to explicit proxy nodes in the augmented SCM, making their cross-environment heteroscedasticity the direct source of identifiability rather than a means to stationarity.

3 Theoretical properties of ORCD

We now establish the formal theoretical foundations of ORCD. All proofs are collected in Appendix A.

3.1 Structural assumptions and residual decomposition

We begin by fixing notation. Let $\mathbf{X} \in \mathbb{R}^{N \times T \times K}$ denote the observed data tensor with N variables, T time points per environment, and K environments. Let $\Theta : \mathbb{R}^N \to \mathbb{R}^D$ be a fixed library of D candidate functions, and write $\Theta(\mathbf{X}_k) \in \mathbb{R}^{T \times D}$ for the library evaluated on environment k. Denote by $\dot{\mathbf{X}}_{i,k} \in \mathbb{R}^T$ the numerically estimated time derivative of variable i in environment k.

Definition 3.1 (Global SINDy coefficient and structured residual)**.** *The* global SINDy coefficient vector *for variable i is any element of*

$$\hat{\boldsymbol{\Xi}}_i \;\in\; \arg\min_{\boldsymbol{\xi} \in \mathbb{R}^D} \sum_{k=1}^{K} \left\| \dot{\mathbf{X}}_{i,k} - \Theta(\mathbf{X}_k)\boldsymbol{\xi} \right\|_2^2 + \lambda \left\| \boldsymbol{\xi} \right\|_0, \qquad \lambda > 0. \tag{1}$$

The structured residual *for variable i, environment k, time t is*

$$R_{i,k,t} \;:=\; \dot{X}_{i,k,t} \;-\; \left[\Theta(\mathbf{X}_k)\, \hat{\boldsymbol{\Xi}}_i \right]_t. \tag{2}$$

$\mathbf{R}_{i,k} := (R_{i,k,1}, \ldots, R_{i,k,T})^\top \in \mathbb{R}^T$ *denotes the residuals for variable i in environment k stacked over time.*

Remark 3.2. *Problem equation 1 is non-convex owing to the ℓ_0 penalty. Zhang & Schaeffer (2019) prove that the sequential-thresholding algorithm of Brunton et al. (2016) converges in at most D iterations to a local minimiser of equation 1 (Theorem 2.5 therein), and that every global minimiser is attainable as such a fixed point (Corollary 2.7 therein). We take $\hat{\boldsymbol{\Xi}}_i$ to be the output of this algorithm throughout.*

The following assumption is the only substantive structural condition required for all main results.

Assumption 3.3 (Dynamically heterogeneous environments)**.** *For every variable $i \in \{1, \ldots, N\}$, there exist environments $k, k' \in \{1, \ldots, K\}$ with $k \neq k'$ such that the true generating functions $f^*_{i,k}$ and $f^*_{i,k'}$ satisfy*

$$\left\| f^*_{i,k} - f^*_{i,k'} \right\|_{L^2(\mu)} \;>\; 0, \tag{3}$$

where μ is the marginal distribution of $\mathbf{X}_{k,t}$ and $\| \cdot \|_{L^2(\mu)}$ denotes the root-mean-square norm under μ, i.e. $\|g\|^2_{L^2(\mu)} = \int g(x)^2 \, \mathrm{d}\mu(x)$, and neither function is exactly representable by an s-sparse linear combination of Θ:

$$\inf_{\boldsymbol{\xi} \,:\, \|\boldsymbol{\xi}\|_0 \le s} \left\| f^*_{i,k} - \Theta\boldsymbol{\xi} \right\|_{L^2(\mu)} \;>\; 0 \qquad \text{for all } s < D. \tag{4}$$

Remark 3.4. *Condition equation 3 requires only that no two environments are dynamically identical for any variable. Condition equation 4 is the standard approximation-error condition in sparse regression; it holds generically whenever the true dynamics are not exactly polynomial with a degree bounded by the library order (Brunton et al., 2016). Neither condition involves the residuals, so Assumption 3.3 is verifiable from domain knowledge prior to fitting.*

The following definition formalises the three components of the structured residual and their statistical relationships.

Definition 3.5 (Residual components)**.** *Fix variable i, environment k, and time t. Assume the numerical differentiation scheme is symmetric (e.g. centred finite differences), so that $\mathbb{E}[v_{i,k,t}] = 0$. The structured residual equation 2 decomposes as:*

$$R_{i,k,t} \;=\; v_{i,k,t} \;+\; w_{i,k,t} \;+\; \delta_{i,k,t}, \tag{5}$$

where the three terms are:

(i) Derivative approximation error. *$v_{i,k,t} := \dot{X}^{\text{true}}_{i,k,t} - \hat{\dot{X}}_{i,k,t}$, the error from numerical differentiation. For a p-th order symmetric scheme with step size Δt, $|v_{i,k,t}| = O(\Delta t^p)$ pointwise, independently of k (Wu, 2021, Section II.B).*

(ii) Measurement noise component. *$w_{i,k,t}$: the contribution to the residual from additive observation noise $\delta\mathbf{X}_{k,t}$. Under $\delta\mathbf{X}_k \sim \mathcal{N}(0, \sigma_w^2 I)$ shared across environments, $\mathbb{E}[w_{i,k,t}^2] = O(\sigma_w^2)$ independently of k.*

(iii) Local dynamic misfit. *$\delta_{i,k,t}$: the gap between the true local dynamics and the global sparse approximation,*

$$\delta_{i,k,t} \;:=\; f_{i,k}^*(\mathbf{X}_{k,t}) \;-\; \big[\Theta(\mathbf{X}_k)\,\hat{\boldsymbol{\Xi}}_i\big]_t. \tag{6}$$

Under the zero-mean assumption on $v_{i,k,t}$ and independence of the measurement instrument from the model approximation, the three components are mutually uncorrelated:

$$\mathbb{E}[v_{i,k,t}\, w_{i,k,t}] = \mathbb{E}[v_{i,k,t}\, \delta_{i,k,t}] = \mathbb{E}[w_{i,k,t}\, \delta_{i,k,t}] = 0. \tag{7}$$

3.2 Heteroscedasticity of structured residuals

Proposition 3.6 (Heteroscedasticity of global-fit residuals)**.** *Under Assumption 3.3 and Definition 3.5, the per-environment residual variance satisfies:*

$$\mathrm{Var}(R_{i,k,t}) \;=\; \sigma_v^2 \;+\; \sigma_w^2 \;+\; \mathbb{E}\big[\delta_{i,k,t}^2\big], \tag{8}$$

where $\sigma_v^2 := \mathrm{Var}(v_{i,k,t})$ and $\sigma_w^2 := \mathrm{Var}(w_{i,k,t})$ are environment-independent constants. Moreover, for generic collections of dynamically heterogeneous environments satisfying Assumption 3.3, there exist $k \neq k'$ such that:

$$\mathrm{Var}(R_{i,k,t}) \neq \mathrm{Var}(R_{i,k',t}). \tag{9}$$

Here "generic" means that the set of environment pairs for which $\mathbb{E}[\delta_{i,k,t}^2] = \mathbb{E}[\delta_{i,k',t}^2]$ despite $\|f_{i,k}^* - f_{i,k'}^*\|_{L^2(\mu)} > 0$ has measure zero with respect to any non-degenerate prior over environment-specific dynamics; the result is therefore a generic property of pooled sparse approximation under heterogeneous dynamics rather than a deterministic consequence of Assumption 3.3 alone. Empirical confirmation via Bartlett's test (Table 19) verifies that this generic condition holds overwhelmingly in the pollution/CVD application.

Proof. See Appendix A.1. □

Empirical confirmation is provided by Bartlett's test of equal variances across the six English regions: all eleven variables yield $p = 0$ after Benjamini–Yekutieli (Benjamini & Yekutieli, 2001) multiplicity adjustment (Table 19).

Corollary 3.7 (Global fit is necessary for heteroscedasticity)**.** *Suppose separate local SINDy models are fit per environment, yielding coefficients $\hat{\boldsymbol{\Xi}}_i^{(k)}$ minimising the per-environment objective. As the library size $D \to \infty$ with sparsity s fixed, the local dynamic misfit satisfies $\mathbb{E}[(\delta_{i,k,t}^{(k)})^2] \to 0$ for each k, and consequently*

$$\mathrm{Var}(R_{i,k,t}^{(k)}) \;\to\; \sigma_v^2 + \sigma_w^2 \qquad \text{as } D \to \infty, \tag{10}$$

which is environment-independent. The heteroscedasticity equation 9 therefore does not arise from local fits, and the identifiability condition of Theorem 3.10 cannot be satisfied.

Proof. See Appendix A.2. □

3.3 Whitened residuals

Step 2 of ORCD operates on whitened residuals obtained via a two-stage procedure. The first stage addresses within-environment temporal autocorrelation: the derivative approximation error $v_{i,k,t}$ (Definition 3.5(i)) introduces autocorrelation into the raw residuals $\mathbf{R}_{i,k}$ along the time index, violating the independent-observation assumption of most causal discovery algorithms. When detected by the Durbin–Watson diagnostic (Durbin & Watson, 1950; 1951), an AR(p) pre-whitening filter is applied to $\mathbf{R}_{i,k}$ along t, with order p selected by BIC; $\mathbf{R}_{i,k}$ is then replaced by the AR residuals. The second stage addresses cross-environment correlation structure: the pre-whitened residuals are transformed as $\widetilde{R}_i := \Lambda^{-1/2} R_i$, where $R_i \in \mathbb{R}^{K \times T}$

collects the residuals with environments as rows and time points as columns, while $\Lambda \in \mathbb{R}^{K\times K}$ is the sample correlation matrix of R_i across environments, and $\Lambda^{-1/2}$ is its symmetric square-root inverse. This cross-environment transformation acts on the environment index k, not the time index, and therefore does not remove within-environment temporal autocorrelation; its sole role is to equalise cross-environment correlation structure while preserving the heteroscedasticity signal established in Proposition 3.6. Temporal independence after both stages is confirmed empirically by the Durbin–Watson test (Table 2).

Proposition 3.8 (Whitening preserves heteroscedasticity)**.** *Let* $\mathrm{Var}(R_{i,k,t}) = \sigma^2_{i,k}$ *be non-uniform across* k *as guaranteed by Proposition 3.6, and let* $\Lambda = \mathrm{Corr}(\mathbf{R}_i)$ *be symmetric positive definite. Then the whitened residuals satisfy*

$$\mathrm{Var}\Big(\widetilde{R}_{i,k,t}\Big) \;=\; \sum_{k'=1}^{K} \big[\Lambda^{-1/2}\big]^2_{kk'}\, \sigma^2_{i,k'}, \tag{11}$$

and there exist $k \neq k'$ *such that* $\mathrm{Var}(\widetilde{R}_{i,k,t}) \neq \mathrm{Var}(\widetilde{R}_{i,k',t})$*, provided* Λ *is not a scalar multiple of the identity. Furthermore, if* $\mathrm{Cov}(R_{i,k,t},\, R_{i,k,t+s}) = 0$ *for all* k *and* $s \neq 0$*, i.e., after AR pre-whitening has been applied along* t*, then* $\mathrm{Cov}(\widetilde{R}_{i,k,t},\, \widetilde{R}_{i,k,t+s}) = 0$ *for all* k *and* $s \neq 0$*.*

Proof. See Appendix A.3. □

Remark 3.9. *Proposition 3.8 establishes two properties of the cross-environment whitening step: it preserves the heteroscedasticity of* $\mathbf{R}_i$ *across environments, and it preserves zero temporal autocorrelation* if already present*. Cross-environment mixing along the environment index* k *cannot affect the within-environment time structure and therefore cannot remove temporal autocorrelation when present. This is why AR(p) pre-whitening along* t *is applied as a first stage, prior to cross-environment whitening, whenever the Durbin–Watson diagnostic detects autocorrelation. The temporal uncorrelatedness condition in the second part of Proposition 3.8 therefore refers to the post-AR-pre-whitening residuals* $\mathbf{R}_{i,k}$*, not to the raw residuals from the ODE fit. The full proof is given in Appendix A.3.*

3.4 Causal identifiability

Consider the augmented structural causal model on $2N$ nodes, comprising the original variables $X_1, \ldots, X_N$ and the whitened residuals $\widetilde{R}_1, \ldots, \widetilde{R}_N$, with structural equations

$$X_i \;=\; f_i\Big(\mathrm{Pa}_{\mathcal{G}}(X_i),\; \widetilde{R}_i\Big) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0,\, \tau_i^2), \quad i = 1, \ldots, N, \tag{12}$$

where $\mathcal{G}$ is the underlying DAG on the $\{X_i\}$ nodes, ε_i are jointly independent across i and independent of $\mathrm{Pa}_{\mathcal{G}}(X_i)$ and $\widetilde{R}_i$, and each $\widetilde{R}_i$ enters as an *observed proxy covariate* rather than as exogenous noise (see Remark 4.1). The *self-residual bottleneck constraint* restricts the augmented DAG: $\widetilde{R}_i$ may appear as a parent only of X_i, not of any X_j with $j \neq i$, and not of any $\widetilde{R}_j$; consequently $\mathrm{Pa}(\widetilde{R}_i) = \emptyset$ for all i.

Theorem 3.10 (Restricted identifiability of ORCD)**.** *Let the data be generated by the SCM equation 12 subject to the self-residual bottleneck, with* $K \geq 2$ *environments. Under Assumption 3.3 and the faithfulness assumption with respect to* G *(Meek, 1995), the DAG* G *is generically identifiable within the admissible DAG class from the multi-environment observational distribution of the augmented system* $(X_1, \ldots, X_N, \widetilde{R}_1, \ldots, \widetilde{R}_N)$*. Specifically, the observed joint distribution varies across environments in a structured way by Propositions 3.6 and 3.8, satisfying the multi-environment distribution shift conditions of Xu et al. (2022) and Duong & Nguyen (2023); and for every pair of distinct admissible DAGs* $G \neq G'$*, their induced multi-environment distributions are generically distinct.*

Proof. See Appendix A.4. □

Remark 3.11 (Relation to single-environment identifiability)**.** *In the single-environment Gaussian setting, a DAG is generally identifiable only up to its Markov equivalence class (Meek, 1995; Colombo & Maathuis, 2014). Faithfulness is invoked in the standard sense: it ensures that no conditional independence holds in the distribution beyond those entailed by d-separation in* G*, so that the environment-stratified likelihood functions induced by distinct admissible DAGs remain distinguishable. Under this assumption, Theorem 3.10 achieves*

Algorithm 1 ODE-Residual Causal Decomposition (ORCD)

Require: Data $\mathbf{X}$; sparsity λ, relaxation ν; whitelist $\mathcal{W}$, blacklist $\mathcal{B}$; bootstrap samples B, threshold τ
Ensure: Averaged DAG $\hat{\mathcal{G}}$ over $2N$ nodes

```
// Step 1: global ODE discovery
for each variable i = 1, ..., N do
    Solve SR3 on pooled data to obtain Ξ̂_i                                  (Equation 13)
    Compute R_{i,k,t} = Ẋ_{i,k,t} − [Θ(X_k)Ξ̂_i]_t                            (Equation 2)
    Diagnose temporal autocorrelation in R_i via Durbin–Watson test
    if autocorrelation detected then
        Apply AR(p) pre-whitening to R_{i,k} along t; BIC selects p
    end if
    Cross-environment whiten: R̃_i = Λ^{−1/2} R_i
end for
// Step 2: joint causal structure learning
Form augmented dataset D = {X_1, ..., X_N, R̃_1, ..., R̃_N}
Enforce W (R̃_i → X_i for all i; tier hierarchy) and B (no arcs into residual nodes; no cross-variable
residual edges; physically implausible arcs)
for b = 1, ..., B do
    Bootstrap resample D; run each of Tabu search, Direct LiNGAM, PC Stable, and DAGMA with
constraints W, B
end for
Ĝ ← averaged network at threshold τ for each algorithm independently
```

point identifiability through two simultaneous mechanisms: (i) *multi-environment heteroscedasticity of the augmented system* $(X_i, \widetilde{R}_i)$ *breaks Markov equivalence by the distribution shift criterion of Xu et al. (2022), invoked here in its multi-environment form, not as an additive-noise model exogeneity condition, since* $\widetilde{R}_i$ *enters the SCM as an observed proxy covariate rather than as exogenous noise (see Remark 4.1); and* (ii) *the self-residual bottleneck restricts the admissible DAG space so that no two distinct admissible structures remain Markov equivalent under the heteroscedasticity pattern induced by Propositions 3.6 and 3.8.*

4 The ORCD framework

4.1 Overview and notation

Let $\mathbf{X} = [x_{itk}]$ be an $N \times T \times K$ data tensor where i indexes variables, t time points, and k environments. ORCD is summarised in Algorithm 1. The backbone structural equation $X_i = f\Big(\mathrm{Pa}(X_i) \cup \widetilde{R}_i\Big) + \varepsilon_i$ defines a $2N$-node augmented SCM in which each whitened residual $\widetilde{R}_i$ enters as an *observed proxy covariate* for environment-specific forcing, not as an exogenous noise term.

4.2 Step 1: global ODE discovery and residual extraction

ORCD adopts an open discovery policy: all N variables are permitted on the right-hand side of each equation, enabling automated feature selection and ensuring that shared cross-environment physics is captured by the global model. The SR3 optimiser (Champion et al., 2020) solves

$$\hat{\boldsymbol{\Xi}}_i = \arg\min_{\boldsymbol{\Xi}_i} \tfrac{1}{2}\left\|\dot{\mathbf{X}}_i - \Theta(\mathbf{X})\boldsymbol{\Xi}_i\right\|_2^2 + \lambda\, R(u) + \tfrac{1}{2\nu}\|w-u\|_2^2, \tag{13}$$

with $\lambda = 0.3$, $\nu = 0.5$, over 100 iterations on the full pooled dataset across all K environments. Full hyperparameter specification and the open-discovery rationale are in Appendix B.

A restricted ODE that excludes some variables would leave "shadow physics" in the residuals: representable deterministic patterns that the causal discovery step would misattribute to the causal structure. By allowing all variables in the library, residuals are purged of representable dynamics, leaving only local dynamic misfit $\delta_{i,k,t}$ and unmodelled external forcing as genuine causal signal.

Remark 4.1 (Residuals as observed proxy covariates, not exogenous noise)**.** *The local dynamic misfit* $\delta_{i,k,t} = f^*_{i,k}(\mathbf{X}_{k,t}) - [\Theta(\mathbf{X}_k)\hat{\mathbf{\Xi}}_i]_t$ *is a function of* $\mathbf{X}_{k,t}$ *and is therefore not, in general, independent of* $\mathrm{Pa}(X_i)$*. ORCD does* not *treat* $\widetilde{R}_i$ *as an exogenous noise term in the sense required by additive noise model identifiability; invoking Xu et al. (2022) and Duong & Nguyen (2023) on that basis would be incorrect. Instead,* $\widetilde{R}_i$ *enters the augmented SCM as an* observed, pre-computed proxy covariate*: it is included as an explicit node in the augmented graph, conditioned on in the BIC regression (Equation 14), and its edge orientation is fixed by the self-residual bottleneck (Section 4.4). The identifiability result of Theorem 3.10 invokes the* multi-environment distribution shift *criterion of Xu et al. (2022) and Duong & Nguyen (2023), which requires only that the observed joint distribution of the augmented system* $(X_i, \widetilde{R}_i)$ *varies across environments in a structured way, a condition guaranteed by the cross-environment heteroscedasticity of* $\widetilde{R}_i$ *established in Propositions 3.6 and 3.8, regardless of whether* δ *is independent of* $\mathrm{Pa}(X_i)$*. The uncorrelatedness conditions of Equation (7) ensure that the three residual components do not conflate in the variance decomposition of Proposition 3.6; they are not invoked to establish exogeneity of* δ*.*

4.3 Step 2: joint causal structure learning

Causal discovery is performed on the augmented $2N$-dimensional space. By the self-residual bottleneck (Section 4.4), each $\widetilde{R}_i$ is a parentless source node in the augmented DAG, so $\mathrm{Pa}(\widetilde{R}_i) = \emptyset$ for all i and its marginal score contributes a constant to the BIC that does not affect structure learning over the $\{X_i\}$ nodes. The BIC score therefore decomposes as

$$\mathrm{BIC}(\mathcal{G}) = \sum_{i=1}^{N} \mathrm{BIC}\Big(X_i \mid \mathrm{Pa}(X_i),\, \widetilde{R}_i\Big) + \sum_{i=1}^{N} \mathrm{BIC}\Big(\widetilde{R}_i\Big)\,, \tag{14}$$

and the local regression model for each X_i is

$$X_i = \beta_i^\top \mathrm{Pa}(X_i) + \gamma_i\, \widetilde{R}_i + \varepsilon_i, \qquad \mathrm{Pa}(X_i) \subset \{X_j\}_{j \neq i}. \tag{15}$$

where $\varepsilon_i \sim \mathcal{N}(0, \tau_i^2)$ is the idiosyncratic error term as in Equation equation 12, independent across i and of $\mathrm{Pa}(X_i)$ and $\widetilde{R}_i$, with τ_i^2 estimated from the data. All discovery algorithms are wrapped in bootstrap aggregation ($B = 200$ resamples, threshold $\tau = 0.5$), producing a single averaged DAG per algorithm (Scutari & Nagarajan, 2013). Full implementation details are in Appendix B.

4.4 Structural constraints

Hierarchical physical prior. Whitelisted arc directions enforce the domain knowledge hierarchy weather (W) $\rightarrow$ air pollution (AP) $\rightarrow$ health (H); blacklists prevent physically implausible reverse arcs (e.g., CVD60 $\rightarrow$ WS). Edges within each tier are unconstrained. An additional directional constraint ($NO_2 \rightarrow O_3$) is imposed based on atmospheric chemistry.

Self-residual bottleneck. Each whitened residual $\widetilde{R}_i$ is permitted to causally affect only its own measured variable X_i:

$$\widetilde{R}_i \rightarrow X_i, \qquad \forall\, i \in \{1, \ldots, N\}. \tag{16}$$

No arcs enter any residual node, and no arcs connect residual nodes to one another, so $\mathrm{Pa}(\widetilde{R}_i) = \emptyset$ for all i, which is why the second sum in Equation equation 14 is unconditional. Together with Propositions 3.6 and 3.8, this constraint enables Theorem 3.10: it fixes the orientation of all edges incident to $\widetilde{R}_i$ by construction, eliminating the residual ambiguities that would otherwise persist within the heteroscedastic Markov equivalence class.

Within-environment temporal pre-whitening. The cross-environment whitening $\widetilde{\mathbf{R}}_i = \Lambda^{-1/2}\mathbf{R}_i$, where $\Lambda \in \mathbb{R}^{K \times K}$ is the sample correlation matrix of $\mathbf{R}_i$ *across environments*, mixes observations along the environment index k and does not, in general, remove temporal autocorrelation within each environment. Proposition 3.8 establishes only that cross-environment whitening *preserves* zero temporal autocorrelation when already present; it cannot create it. When the Durbin–Watson diagnostic (Algorithm 1, line 5) detects residual autocorrelation within an environment, an ARMA(p,q) pre-whitening filter is applied to $\mathbf{R}_{i,k}$

Table 1: Win rates (out of 12 settings) and representative SHD for $N = 20, n = 100, d = 0.5$. ORCD reduces SHD in every algorithm–setting combination.

Algorithm	Win rate (out of 12)		SHD for $N = 20, n = 100, d = 0.5$	
	SHD	SID	ORCD	Baseline
Tabu search	12/12	11/12	1.831	3.225
Direct LiNGAM	12/12	11/12	1.239	4.408
PC Stable	12/12	12/12	1.296	1.408
DAGMA	12/12	12/12	1.239	1.282

along the time index prior to cross-environment whitening, with orders $p, q \leq 5$ selected automatically via `auto.arima()`, following standard practice (Brockwell & Davis, 1991). The order p is selected by BIC on the AR model for each variable–cluster combination. The corrected two-stage whitening procedure is:

1. **Within-environment:** fit AR(p) to $\mathbf{R}_{i,k}$ along t; replace $\mathbf{R}_{i,k}$ with its AR residuals.
2. **Cross-environment:** apply $\widetilde{\mathbf{R}}_i = \Lambda^{-1/2}\mathbf{R}_i$ to equalise cross-environment correlation structure and preserve heteroscedasticity (Proposition 3.8).

Durbin–Watson statistics after both stages confirm the absence of residual autocorrelation (Table 2).

5 Experiments and results

All experiments use four causal discovery algorithms: Tabu search (Glover & Laguna, 1998), Direct LiNGAM (Shimizu et al., 2011), PC Stable (Colombo & Maathuis, 2014), and DAGMA (Bello et al., 2022), subject to the structural constraints of Section 4.4. Each algorithm is wrapped in bootstrap aggregation ($B = 200$ resamples, arc inclusion threshold $\tau = 0.5$, corresponding to a majority-vote criterion). All variables are min-max normalised to $[0, 1]$ prior to SINDy fitting.

5.1 Simulation study

Design. We evaluate ORCD across twelve settings defined by $N \in \{10, 20\}$ nodes, $n \in \{100, 200\}$ observations per environment, and edge density $d \in \{0.1, 0.5, 0.8\}$. Random DAGs are generated and data simulated from linear additive Gaussian SCMs with environment-varying dynamics satisfying Assumption 3.3. For each setting, we compare ORCD against the same algorithm run without residuals (the X_i-only baseline), reporting scaled SHD (divided by the number of true arcs) and scaled SID (divided by $(|\mathcal{A}|/2)^2$, where $\mathcal{A}$ denotes the arc set of the true DAG), against the known true DAG. Full per-setting results are in Appendix D.5.

Results. Table 1 summarises results. ORCD strictly reduces SHD in **48 out of 48** algorithm–setting combinations (100%). SID improvements hold in 46/48 combinations (100% for PC Stable and DAGMA, 91.7% for Tabu and LiNGAM). Gains are substantial: median SHD reduction of 36% across all settings, reaching 43% for Tabu and 72% for LiNGAM in the most challenging large-graph, low-sample-size regime ($N = 20, n = 100, d = 0.5$). These results are consistent with Theorem 3.10: the heteroscedastic residuals consistently improve structural accuracy regardless of which discovery algorithm is used.

5.2 Real-world analysis: England air pollution and cardiovascular mortality

5.2.1 Data and preprocessing

We apply ORCD to the data published by Vitolo et al. (2018) and made publicly available via Zenodo (Zenodo, 2017), covering England from 1981 to 2014. The dataset links cardiovascular mortality rates for adults aged 60 and above (CVD60), four air pollution variables (O_3, PM10, SO_2, NO_2; PM2.5 and CO removed due to collinearity), and six meteorological variables (WS, WD, T2M, TP, BLH, SSR). Raw data ($\approx$50 million observations) were spatially aggregated into six geographical clusters via k-means on station coordinates (Carvalho et al., 2016), approximately Greater London (Cl.1), South West (Cl.2), East (Cl.3), Central (Cl.4), North West (Cl.5), and North East England (Cl.6), and temporally aggregated to one daily

Table 2: Proportion of Durbin–Watson p-values below 0.01 after Benjamini–Yekutieli adjustment, by variable and region, computed on the raw ODE residuals prior to AR pre-whitening. Non-zero values motivate Stage 1; all reduce to zero after Stage 1 is applied. Cl.1–Cl.6: Greater London (Cl.1), South West (Cl.2), East (Cl.3), Central (Cl.4), North West (Cl.5), North East England (Cl.6).

Variable	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6
BLH	0.3	0.0	0.3	0.3	0.0	0.3
CVD60	0.2	0.0	0.2	0.2	0.0	0.1
NO2	0.5	0.0	0.5	0.7	0.0	0.1
O3	0.3	0.0	0.4	0.3	0.0	0.2
PM10	0.2	0.0	0.2	0.3	0.0	0.3
SO2	0.2	0.0	0.2	0.5	0.0	0.1
SSR	0.3	0.0	0.3	0.3	0.0	0.1
T2M	0.3	0.0	0.3	0.4	0.0	0.3
TP	0.1	0.1	0.2	0.2	0.0	0.1
WD	0.4	0.0	0.4	0.5	0.0	0.3
WS	0.3	0.0	0.3	0.3	0.0	0.3

value per cluster using a moving average, yielding 54,769 observations. Missing air pollution values are missing not at random (Little & Rubin, 1987) and were imputed using the LSTM-based model of Kushwah & Agrawal (2024), which outperformed MICE by 15–45% in mean squared error across pollutants. Full variable details are in Appendix C.

5.2.2 ODE fitting and residual diagnostics

A single SINDy model was fit on the full pooled dataset with $\lambda = 0.3$, $\nu = 0.5$, and 100 SR3 iterations, with all variables in the library (open discovery). Two diagnostics confirm the theoretical conditions of Section 3.

Within-environment temporal autocorrelation. Table 2 reports the proportion of Durbin–Watson p-values below 0.01 (Benjamini–Yekutieli adjusted) for the raw residuals $\mathbf{R}_{i,k}$ prior to whitening. Non-zero values confirm temporal autocorrelation in the raw residuals, motivating AR(p) pre-whitening as Stage 1 of the two-stage whitening procedure; all values reduce to zero after Stage 1.

Cross-environment heteroscedasticity. Bartlett's test yields $p \leq 6.07 \times 10^{-19}$ for all eleven variables after Benjamini–Yekutieli adjustment, with four variables (CVD60, NO2, SO2, BLH) returning $p = 0$ at machine precision (Table 19, Appendix D.6), providing overwhelming empirical confirmation of Proposition 3.6 and Assumption 3.3. Additional diagnostics: Pearson and Spearman correlation matrices, normal QQ plots, and ACF plots of $\widetilde{\mathbf{R}}_i$, are in Appendix D.6.

5.2.3 Experimental protocols and metrics

Experiment A (temporal generalisation). For each cluster, the BN is trained on a random 80% sample of the combined $\{X_i, \widetilde{R}_i\}$ space and evaluated on the held-out 20%, using pre-extracted residuals to avoid propagating the derivative estimation error into the structural evaluation. Random sampling breaks temporal autocorrelation, so the BN learns a time-stable causal backbone and the held-out 20% provides an unbiased estimate of predictive accuracy on unseen observations.

Experiment B (spatial generalisation). The BN is trained on one cluster's complete dataset and applied to predict outcomes in each of the five remaining clusters. Because residuals derive from a single pooled model, no additional fitting is required for test clusters.

Evaluation metrics: predictive Spearman correlation; scaled SHD and SID relative to the per-cluster consensus average DAG; F1 score. All algorithms produce DAGs with approximately 2 arcs per node.

Table 3: Predictive Spearman correlation for CVD60, Experiment A (temporal generalisation). Cl.1–Cl.6 denote geographical clusters as defined in Table 2.

Algorithm	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6
Tabu search	0.626	0.626	0.626	0.626	0.625	0.626
Direct LiNGAM	0.623	0.623	0.625	0.625	0.623	0.626
PC Stable	0.609	0.622	0.608	0.609	0.622	0.609
DAGMA	0.633	0.617	0.624	0.613	0.621	0.592

Table 4: Structural metrics from Experiment A against the per-cluster consensus DAG. All four algorithms are run within the ORCD framework on the augmented $\{X_i, \widetilde{R}_i\}$ space. SHD scaled by consensus arc count; SID scaled by $(|\mathcal{A}|/2)^2$, where $\mathcal{A}$ is the arc set of the consensus DAG. Cl.1–Cl.6 denote geographical clusters as defined in Table 2.

Algorithm	Metric	Cl.1	Cl.2	Cl.3	Cl.4	Cl.5	Cl.6
Tabu	Arcs	45	43	43	44	45	44
	F1	0.984	0.989	0.984	0.988	0.982	0.982
	Scaled SHD	0.267	0.209	0.326	0.227	0.311	0.295
	Scaled SID	0.162	0.160	0.190	0.153	0.204	0.192
LiNGAM	Arcs	41	41	41	38	37	42
	F1	0.991	0.991	0.989	0.988	0.987	0.991
	Scaled SHD	0.171	0.146	0.244	0.263	0.297	0.167
	Scaled SID	0.131	0.102	0.159	0.213	0.275	0.077
PC	Arcs	37	37	37	37	39	37
	F1	0.982	0.989	0.984	0.987	0.982	0.981
	Scaled SHD	0.432	0.270	0.405	0.324	0.410	0.432
	Scaled SID	0.412	0.263	0.307	0.333	0.376	0.389
DAGMA	Arcs	43	42	45	43	47	44
	F1	0.982	0.983	0.982	0.980	0.973	0.973
	Scaled SHD	0.395	0.333	0.378	0.419	0.468	0.500
	Scaled SID	0.208	0.177	0.140	0.188	0.152	0.213

5.2.4 Results: Experiment A

Predictive accuracy. Table 3 reports CVD60 Spearman correlations. All algorithms achieve consistent accuracy (0.592–0.633), with DAGMA achieving the highest point estimate in Cl.1 ($r = 0.633$) and Tabu and LiNGAM showing the most uniform cross-region performance ($r = 0.625$–0.626 and 0.623–0.626, respectively). The predictive Spearman correlation of each whitened residual node with its own variable is below 0.034 in absolute value across all algorithms and regions, confirming that the whitened residuals act as structural conduit variables rather than direct informative predictors.

Structural accuracy. Table 4 reports structural metrics against the per-cluster consensus DAG. ORCD recovers 37–47 arcs per cluster (F1 = 0.973–0.991). Direct LiNGAM achieves the lowest divergence (scaled SHD 0.15–0.30; scaled SID 0.08–0.28), followed by Tabu search. PC Stable misses more arcs; DAGMA tends to include spurious ones. The structural gains over the X_i-only baseline are substantial: the ablation in Section 5.2.6 shows the baseline recovers only 22–28 arcs with F1 falling to 0.895–0.963 versus 0.973–0.991 for ORCD.

5.2.5 Results: Experiment B

Table 5 reports CVD60 Spearman correlations across all 30 cross-cluster pairs. ORCD achieves $r = 0.643$–0.656, at or above the level of Experiment A, demonstrating that the learned causal structure transfers to

Table 5: CVD60 Spearman correlation, Experiment B (spatial generalisation). Mean and standard deviation over 30 cross-cluster pairs.

Algorithm	Mean	Std. Dev.
Tabu search	0.655	0.018
Direct LiNGAM	0.655	0.018
PC Stable	0.643	0.020
DAGMA	0.656	0.017

unseen geographical environments without degradation. Learned DAGs contain 38–50 arcs per cluster. The predictive Spearman correlation of each whitened residual node with its own variable remains below 0.11 in absolute value; the whitened residual of CVD60 carries a consistent mean correlation of approximately 0.13 with the predicted CVD60 values, reflecting the genuine causal signal encoded by the whitened residuals that the BN exploits during parameter learning. Full per-variable results are in Appendix D.2.

5.2.6 Ablation: role of the residuals

We perform a two-stage ablation under both protocols to verify that $\widetilde{R}_i$ carries genuine causal information. Table 6 summarises CVD60 Spearman correlations across all conditions; full tables are in Appendix D.3.

Stage 1 (only X_i). Causal discovery and parameter learning are performed on X_i alone. Under Experiment A (non-i.i.d. temporal hold-out), predictive accuracy is broadly comparable between ORCD and the X_i-only baseline for Tabu (0.626 vs. 0.626) and DAGMA (0.617 vs. 0.614), while ORCD leads clearly for LiNGAM (0.624 vs. 0.599); structural accuracy, however, is substantially worse for all algorithms: 22–28 arcs versus 37–47 for ORCD, and F1 falling to 0.895–0.963 versus 0.973–0.991. The non-i.i.d. split inflates predictive accuracy for both models by allowing temporal autocorrelation to leak across train and test, making this an unreliable basis for comparison. Experiment B provides the decisive test: under i.i.d. spatial transfer to entirely unseen geographical regions, ORCD *strictly dominates* the X_i-only baseline for every algorithm—Tabu (0.655 vs. 0.634), LiNGAM (0.655 vs. 0.618), PC (0.643 vs. 0.630), and DAGMA (0.656 vs. 0.647)—with gaps of 0.009–0.037 Spearman points. The LiNGAM gap of 0.037 substantially exceeds one standard deviation (0.018–0.036), constituting a meaningful difference. This uniform dominance confirms that the richer causal graphs recovered by ORCD (38–50 arcs per cluster versus 18–33 for the X_i-only baseline) encode genuine cross-regional causal dependencies rather than region-specific noise. The non-i.i.d. split in Experiment A masks this advantage by allowing temporal autocorrelation to inflate predictive accuracy for sparser models; Experiment B, being fully i.i.d. across environments, provides the clean comparison, and the outcome is unambiguous.

Stage 2 (without $\widetilde{R}_i$ in parameter learning). Causal discovery is performed on X_i alone, but parameter learning uses $X_i - \hat{R}_i$. If the residuals were noise, subtracting them would improve predictions; the opposite occurs, dramatically. Under Experiment A, CVD60 collapses to 0.013–0.037 across all algorithms—a decrease of approximately 0.58–0.61 Spearman points versus ORCD. Under Experiment B, the collapse is equally severe: CVD60 falls to 0.019–0.033, a decrease of approximately 0.62–0.64 points versus ORCD ($r = 0.643$–0.656), reducing predictions to near chance level. Precipitation (TP) undergoes a comparable collapse in both protocols (Exp A: $r \approx 0.02$–0.10 versus $r \approx 0.46$–0.49 for ORCD; Exp B mean $r \approx 0.06$–0.11 versus $r \approx 0.47$–0.49). The near-total collapse of CVD60 predictive accuracy upon residual subtraction confirms that the whitened residuals encode the dominant causal signal exploited by the Bayesian network, and that this signal generalises robustly across unseen geographical environments.

5.2.7 Temporal invariance of the residuals

Each cluster's whitened residuals are split into two chronological halves (9,127 observations each); a BN is learned from each half and evaluated on the other. The cross-predictive symmetry ratio $\rho^{(s,s')}/\rho^{(s',s)}$ has median 0.994 (Tabu), 0.995 (LiNGAM), and 0.995 (PC Stable), with 90% confidence intervals including 1 in all cases, confirming no systematic directional bias between the two halves. Scaled SHD between DAGs from the two halves has median 0.394, 0.410, and 0.455, respectively, with overlapping confidence intervals across

Table 6: CVD60 Spearman correlation across ablation conditions. Experiment A reports means across six regions; Experiment B reports means across 30 cross-cluster pairs (standard deviations in parentheses). ORCD: full model; onlyXi: structure and parameter learning on X_i alone; withoutRi: structure on X_i, parameter learning on $X_i - \hat{R}_i$.

Protocol	Condition	Tabu	LiNGAM	PC	DAGMA
Exp. A	ORCD	0.626	0.624	0.613	0.617
	onlyXi	0.626	0.599	0.616	0.614
	withoutRi	0.016	0.013	0.019	0.037
Exp. B	ORCD	0.655 (0.018)	0.655 (0.018)	0.643 (0.020)	0.656 (0.017)
	onlyXi	0.634 (0.024)	0.618 (0.036)	0.630 (0.021)	0.647 (0.018)
	withoutRi	0.021 (0.028)	0.019 (0.031)	0.025 (0.029)	0.033 (0.027)

all three algorithms, confirming no statistically significant structural differences. The residuals, therefore, capture a temporally stable causal structure. Full results are in Appendix D.4.

6 Discussion

The central finding of this work is that a single global ODE fit across multiple environments mechanically generates structured residuals whose cross-environment heteroscedasticity is both provable (Proposition 3.6) and empirically confirmed ($p \leq 6.07 \times 10^{-19}$ by Bartlett's test across all eleven variables, with three variables returning $p = 0$ at machine precision). This heteroscedasticity is sufficient for unique DAG identifiability under Gaussian noise, a regime where LiNGAM-type methods fail, and is eliminated by per-environment local fits, making the global fit a necessary design commitment rather than a simplification. The practical consequence, confirmed across both simulation and real-world evaluation, is that incorporating the whitened residuals as observed proxy covariates in a Bayesian network search consistently improves structural accuracy, with median SHD reductions of 36% in the simulation study, and a causal structure that generalises across both time and geography in the pollution/CVD analysis.

A subtlety worth highlighting for practitioners concerns the relationship between predictive and structural accuracy. Under non-i.i.d. temporal hold-out (Experiment A), omitting the residuals can match or marginally exceed ORCD's raw predictive accuracy for some algorithms while simultaneously *degrading* structural accuracy: the X_i-only baseline recovers only 22–28 arcs with F1 = 0.895–0.963, versus 37–47 arcs and F1 = 0.973–0.991 for ORCD. This pattern reflects non-i.i.d. inflation: temporal autocorrelation leaks across train and test splits, benefiting sparser models that rely more heavily on direct parametric associations. Under i.i.d. spatial transfer (Experiment B), this inflation is absent, and ORCD's structural advantage translates into a clear and consistent predictive advantage, with ORCD strictly dominating the X_i-only baseline for every algorithm by 0.009–0.037 Spearman points. The Stage 2 ablation confirms the underlying mechanism: subtracting the residuals from the outcome variable before parameter learning, collapses CVD60 Spearman correlations by approximately 0.60 points, from $r \approx 0.62$–0.66 to near zero ($r \approx 0.01$–0.04), across all algorithms and protocols, demonstrating that the whitened residuals encode the dominant causal signal. The predictive Spearman correlation of each whitened residual node with its own variable is below 0.034 in Experiment A, confirming their role as structural conduit variables; in Experiment B, the whitened residual of CVD60 carries a consistent mean correlation of ≈ 0.13 with the predicted CVD60 values, reflecting the genuine cross-environment forcing captured by the global ODE fit. Predictive accuracy under non-i.i.d. conditions is therefore not a reliable proxy for causal accuracy when evaluating decompositional causal discovery methods.

Limitations and future work. The SINDy library is a sparse combination of elementary functions; complex local dynamics not in this library appear in the residuals, potentially conflating causal signal with representational limitations. Methods discovering source terms via symbolic regression, such as D-CIPHER (Kacprzyk et al., 2023), are natural extensions. The global fit presupposes that a meaningful shared dynamic law exists across all K environments; developing a diagnostic for cases where this assumption fails is important future work. The hierarchical physical prior must be specified by the analyst, and ORCD

does not model latent confounders without environmental manifestation. As established in Remark 4.1, the local dynamic misfit $\delta_{i,k,t}$ is not in general independent of $\mathrm{Pa}(X_i)$; identifiability is invoked via the multi-environment distribution shift criterion rather than additive noise model exogeneity, and future work should characterise more precisely the conditions under which this distinction matters empirically. Finally, the two-stage whitening procedure addresses within-environment temporal autocorrelation via AR pre-whitening before cross-environment whitening; developing theoretical guarantees for the combined procedure under general autocorrelation structures is an open problem. As ORCD targets a distinct regime, i.e., heterogeneous environments sharing a global dynamic law rather than discrete temporal regime shifts, direct comparison with nonstationarity-aware baselines (Saggioro et al., 2020; Rahmani et al., 2025) requires carefully controlled shared benchmarks, which we leave for future work.

Broader Impact Statement

ORCD addresses a challenge that arises across a wide class of physical, biological, and social systems: the entanglement of slow-scale invariant dynamics with fast-scale instantaneous causal forcing driven by external inputs. This entanglement is the norm rather than the exception in multivariate time series collected across heterogeneous environments, e.g., from atmospheric and climate systems, to neurophysiology, to economics, and ORCD provides a principled path to separating these two regimes wherever a meaningful shared dynamic law exists across environments. Its most direct application here, linking environmental exposures to health outcomes, illustrates both the potential and the responsibility: erroneous causal conclusions in this domain carry real societal costs. Causal inference from observational data always carries risk: if Assumption 3.3 is violated or the library is misspecified, the residuals will not carry genuine causal signal and the recovered DAG may be unreliable. We recommend that practitioners verify Bartlett's test and the Durbin–Watson diagnostics (Section 5.2.2) before drawing policy conclusions from any ORCD analysis.

Reproducibility. Full implementation details are provided in Appendix B. The real-world dataset is publicly available via Zenodo (2017). Code to reproduce all experiments will be made publicly available upon acceptance.

Acknowledgments

References

K. Bello, B. Aragam, and P. Ravikumar. DAGMA: Learning DAGS via M-Matrices and a Log-Determinant Acyclicity Characterization. *Advances in Neural Information Processing Systems*, 35:8226–8239, 2022.

Y. Benjamini and D. Yekutieli. The Control of the False Discovery Rate in Multiple Testing Under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

P. J. Brockwell and R. A. Davis. *Time Series: Theory and Methods*. Springer, 2nd edition, 1991.

S. L. Brunton, J. L. Proctor, and J. N. Kutz. Discovering Governing Equations From Data by Sparse Identification of Nonlinear Dynamical Systems. *Proceedings of the National Academy of Sciences*, 113(15): 3932–3937, 2016.

M. Carvalho, P. Melo-Gonçalves, J. Teixeira, and A. Rocha. Regionalization of Europe Based on a K-Means Cluster Analysis of the Climate Change of Temperatures and Precipitation. *Physics and Chemistry of the Earth, Parts a/B/C*, 94:22–28, 2016.

K. Champion, P. Zheng, A. Y. Aravkin, S. L. Brunton, and J. N. Kutz. A Unified Sparse Optimization Framework to Learn Parsimonious Physics-Informed Models From Data. *IEEE Access*, 8:169259–169271, 2020.

D. Colombo and M. H. Maathuis. Order-Independent Constraint-Based Causal Structure Learning. *Journal of Machine Learning Research*, 15:3921–3962, 2014.

A. Dhir, R. Sedgwick, A. Kori, B. Glocker, and M. van der Wilk. Continuous bayesian model selection for multivariate causal discovery. arXiv preprint arXiv:2411.10154, 2024.

B. Duong and T. Nguyen. Heteroscedastic Causal Structure Learning. In *ECAI 23*, volume 372 of *Frontiers in Artificial Intelligence and Applications*, pp. 598–605, 2023.

J. Durbin and G. S. Watson. Testing for Serial Correlation in Least Squares Regression. I. *Biometrika*, 37 (3–4):409–428, 1950.

J. Durbin and G. S. Watson. Testing for Serial Correlation in Least Squares Regression. Ii. *Biometrika*, 38 (1–2):159–178, 1951.

M. H. Ferdous and M. O. Gani. DCD: Decomposition-Based Causal Discovery from Autocorrelated and Non-Stationary Temporal Data. *arXiv preprint arXiv:2602.01433*, 2026. Under review at TMLR.

F. Glover and M. Laguna. *Tabu Search*, pp. 2093–2229. Springer, 1998.

K. Kacprzyk, Z. Qian, and M. van der Schaar. D-CIPHER: Discovery of Closed-Form Partial Differential Equations. In *Advances in Neural Information Processing Systems*, volume 36, pp. 27609–27644, 2023.

A. Kaptanoglu, B. de Silva, U. Fasel, K. Kaheman, A. Goldschmidt, J. Callaham, C. Delahunt, Z. Nicolaou, K. Champion, J. Loiseau, J. Kutz, and S. Brunton. PySINDy: A Comprehensive Python Package for Robust Sparse System Identification. *Journal of Open Source Software*, 7(69):3994, 2022.

V. Kushwah and P. Agrawal. Hybrid Model for Air Quality Prediction Based on LSTM with Random Search and Bayesian Optimization Techniques. *Earth Science Informatics*, 18:32, 2024.

M. Lenatti, M. Zaffalon, A. Antonucci, P. F. De Paola, L. Multerer, M. Mongelli, A. Paglialonga, and L. Azzimonti. Counterfactual Inference Using Ordinary Differential Equations to Assess the Effect of Physical Activity on Type 2 Diabetes Onset. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Medicine*, pp. 212–221, 2025.

R. J. A. Little and D. B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 1987.

S. Löwe, D. Madras, R. Zemel, and M. Welling. Amortized Causal Discovery: Learning to Infer Causal Graphs from Time-Series Data. In *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 509–525. PMLR, 2022.

C. Meek. Causal Inference and Causal Explanation with Background Knowledge. In *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, pp. 403–410, 1995.

G. Papanastasiou, G. Yang, D. I. Fotiadis, and et al. Large-Scale Deep Learning Analysis to Identify Adult Patients at Risk for Combined and Common Variable Immunodeficiencies. *Communications Medicine*, 3 (1):189, 2023.

G. Papanastasiou, M. Scutari, R. Tachdjian, and et al. Large Scale Causal Modeling to Identify Adults at Risk for Combined and Common Variable Immunodeficiencies. *npj Digital Medicine*, 8(1):361, 2025.

J. Pearl. *Causality*. Cambridge University Press, 2009.

J. Peters and P. Bühlmann. Identifiability of Gaussian Structural Equation Models with Equal Error Variances. *Biometrika*, 101(1):219–228, 2013.

J. Peters, S. Bauer, and N. Pfister. Causal Models for Dynamical Systems. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 671–690. Association for Computing Machinery (ACM), 2022.

A. Rahmani et al. Flow-Based Non-Stationary Temporal Regime Causal Structure Learning. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2506.17065.

W. Rudin. *Principles of Mathematical Analysis*. International Series in Pure and Applied Mathematics. McGraw-Hill, New York, 3rd edition, 1976.

J. Runge, S. Bathiany, E. Bollt, G. Camps-Valls, D. Coumou, E. Deyle, C. Glymour, M. Kretschmer, M. D. Mahecha, J. Muñoz-Marí, et al. Inferring Causation From Time Series in Earth System Sciences. *Nature Communications*, 10(1):2553, 2019.

E. Saggioro, J. de Wiljes, M. Kretschmer, and J. Runge. Reconstructing Regime-Dependent Causal Relationships from Observational Time Series. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 30 (11):113115, 2020. doi: 10.1063/5.0020538.

G. Schwarz. Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464, 1978.

M. Scutari and R. Nagarajan. On Identifying Significant Edges in Graphical Models of Molecular Networks. *Artificial Intelligence in Medicine*, 57(3):207–217, 2013.

M. Scutari, D. Kerob, J. Krutmann, and S. Salah. Causal Networks of Infodemiological Data: Modelling Dermatitis. In *Proceedings of the 23rd International Conference on Artificial Intelligence in Medicine (AIME25)*, pp. 397–407, 2025.

S. Shimizu, P. O. Hoyer, A. Hyvärinena, and A. Kerminen. A Linear Non-Gaussian Acyclic Model for Causal Discovery. *Journal of Machine Learning Research*, 7(72):2003–2030, 2006.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washio, P. O. Hoyer, K. Bollen, and P. Hoyer. DirectLiNGAM: A Direct Method for Learning a Linear Non-Gaussian Structural Equation Model. *Journal of Machine Learning Research*, 12(Apr):1225–1248, 2011.

J. Suzuki and T.-L. Yang. Generalization of LiNGAM that allows confounding. In *2024 IEEE International Symposium on Information Theory (ISIT)*, pp. 3540–3545. IEEE, 2024.

C. Vitolo, M. Scutari, A. Tucker, and A. Russell. Modelling Air Pollution, Climate and Health Data Using Bayesian Networks: a Case Study of the English Regions. *Earth and Space Science*, 5(4):76–88, 2018.

Y. Wu. Error Processing of Sparse Identification of Nonlinear Dynamical Systems via L_∞ Approximation, 2021. URL `https://arxiv.org/abs/2107.06142`.

S. Xu, O. A. Mian, A. Marx, and J. Vreeken. Inferring Cause and Effect in the Presence of Heteroscedastic Noise. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 24615–24630, 2022.

Zenodo. MEHRA Data and Model for the English Regions, 2017. URL `https://zenodo.org/record/838571`.

L. Zhang and H. Schaeffer. On the Convergence of the SINDy Algorithm. *Multiscale Modeling & Simulation*, 17(3):948–972, 2019.

X. Zheng, B. Aragam, P. K. Ravikumar, and E. P. Xing. DAGs with No Tears: Continuous Optimization for Structure Learning. *Advances in Neural Information Processing Systems*, 31:0472–9483, 2018.

J. Y. Zhu, C. Zhang, H. Zhang, S. Zhi, V. O. Li, J. Han, and Y. Zheng. pg-Causality: Identifying Spatiotemporal Causal Pathways for Air Pollutants with Urban Big Data. *IEEE Transactions on Big Data*, 4(4):571–585, 2017.

A Proofs of theoretical results

A.1 Proof of Proposition 3.6

Proof. **Variance decomposition.** By Definition 3.5 and the mutual uncorrelatedness equation 7,

$$\mathrm{Var}(R_{i,k,t}) \;=\; \mathrm{Var}(v_{i,k,t}) + \mathrm{Var}(w_{i,k,t}) + \mathrm{Var}(\delta_{i,k,t}). \tag{17}$$

Environment-independence of $\mathrm{Var}(v_{i,k,t})$**.** For a p-th order symmetric scheme with uniform step size Δt, the leading error has magnitude $O(\Delta t^p)$ with a prefactor proportional to the p-th derivative of $X_{i,k,\cdot}$. Under the regularity conditions of Wu (2021, Section II.B), this prefactor is bounded uniformly over k by a constant depending only on p and the global Lipschitz constant of f_i^*. Hence $\mathrm{Var}(v_{i,k,t}) = \sigma_v^2$ for all k.

Environment-independence of $\mathrm{Var}(w_{i,k,t})$**.** Under $\delta\mathbf{X}_k \sim \mathcal{N}(0, \sigma_w^2 I)$ shared across environments, $w_{i,k,t}$ enters the residual as $w_{i,k,t} = -[\Theta(\delta\mathbf{X}_k)\hat{\boldsymbol{\Xi}}_i]_t$ to first order. Since both σ_w^2 and $\hat{\boldsymbol{\Xi}}_i$ are independent of k, $\mathrm{Var}(w_{i,k,t}) = \sigma_w^2$ for all k.

Environment-dependence of $\mathrm{Var}(\delta_{i,k,t})$**.** Let $\hat{S} \subseteq \{1, \ldots, D\}$ be the active support identified by the thresholding algorithm, and let $\Theta_{\hat{S}}(\mathbf{X}_k)$ be the corresponding submatrix. The vector $\hat{\boldsymbol{\Xi}}_i$ restricted to $\hat{S}$ satisfies the normal equations of the pooled support-restricted least-squares subproblem:

$$\sum_{k=1}^{K} \Theta_{\hat{S}}(\mathbf{X}_k)^\top \Big(\dot{\mathbf{X}}_{i,k} - \Theta_{\hat{S}}(\mathbf{X}_k)\hat{\boldsymbol{\Xi}}_{i,\hat{S}} \Big) = \mathbf{0}, \tag{18}$$

confirmed as a local minimiser by Zhang & Schaeffer (2019, Theorem 2.5). Thus $\hat{\boldsymbol{\Xi}}_i$ is the pooled least-squares projection of $(f_{i,1}^*, \ldots, f_{i,K}^*)$ onto the column space of $\Theta_{\hat{S}}$.

Under Assumption 3.3, condition equation 4 ensures that neither $f_{i,k}^*$ nor $f_{i,k'}^*$ lies in the column space of $\Theta_{\hat{S}}$. The expected squared misfit satisfies

$$\mathbb{E}\big[\delta_{i,k,t}^2\big] \;=\; \big\| f_{i,k}^* - \mathrm{Proj}_{\mathcal{S}(\Theta,s)}\, f_{i,k}^* \big\|_{L^2(\mu)}^2 + \mathrm{bias}_k(\hat{\boldsymbol{\Xi}}_i), \tag{19}$$

where $\mathrm{bias}_k(\hat{\boldsymbol{\Xi}}_i) \geq 0$ is the additional bias of the pooled estimator relative to the per-environment oracle. The pooled minimiser allocates approximation capacity in proportion to the magnitude of each environment's gap. Condition equation 3 guarantees $\|f_{i,k}^* - f_{i,k'}^*\|_{L^2(\mu)} > 0$, but two functions that differ in $L^2(\mu)$ can in principle be equidistant from a finite-dimensional subspace, so $\mathrm{bias}_k(\hat{\Xi}_i) \neq \mathrm{bias}_{k'}(\hat{\Xi}_i)$ does not follow deterministically. However, the set of environment pairs for which equal projected distances arise despite unequal functions has measure zero with respect to any non-degenerate prior over environment-specific dynamics (a standard genericity argument in approximation theory). Therefore, for generic environments satisfying Assumption 3.3, $\mathbb{E}[\delta_{i,k,t}^2] \neq \mathbb{E}[\delta_{i,k',t}^2]$.

Conclusion. Substituting into equation 17 gives equation 8. Since $\mathbb{E}[\delta_{i,k,t}^2] \neq \mathbb{E}[\delta_{i,k',t}^2]$ for the k, k' identified in Assumption 3.3, inequality equation 9 follows. □

A.2 Proof of Corollary 3.7

Proof. Let $\hat{\boldsymbol{\Xi}}_i^{(k)}$ be any local minimiser of the per-environment objective, with local dynamic misfit $\delta_{i,k,t}^{(k)} := f_{i,k}^*(\mathbf{X}_{k,t}) - [\Theta(\mathbf{X}_k)\hat{\boldsymbol{\Xi}}_i^{(k)}]_t$. As $D \to \infty$ with s fixed, by the Stone–Weierstrass theorem (Rudin, 1976, Theorem 7.26) the set $\mathcal{S}(\Theta, s)$ of s-sparse combinations of a polynomial or trigonometric library becomes dense in $L^2(\mu)$. Under the standard regularity assumption that each $f_{i,k}^*$ is continuous on the compact support of μ, Stone–Weierstrass guarantees uniform approximation of $f_{i,k}^*$ specifically, so $\inf_{\boldsymbol{\xi}:\|\boldsymbol{\xi}\|_0 \leq s} \|f_{i,k}^* - \Theta\boldsymbol{\xi}\|_{L^2(\mu)}^2 \to 0$ as $D \to \infty$ with s fixed.

Consequently $\inf_{\boldsymbol{\xi}:\|\boldsymbol{\xi}\|_0 \leq s} \|f_{i,k}^* - \Theta\boldsymbol{\xi}\|_{L^2(\mu)}^2 \to 0$, so $\mathbb{E}[(\delta_{i,k,t}^{(k)})^2] \to 0$ for each k. Applying equation 17 to local fits gives $\mathrm{Var}(R_{i,k,t}^{(k)}) \to \sigma_v^2 + \sigma_w^2$ as $D \to \infty$, which is environment-independent. Inequality equation 9 therefore fails in the limit, and the hypothesis of Theorem 3.10 is not satisfied. □

A.3 Proof of Proposition 3.8

Proof. **Variance formula.** Write $\mathbf{r}_{i,t} := (R_{i,1,t}, \ldots, R_{i,K,t})^\top \in \mathbb{R}^K$, so $\widetilde{\mathbf{r}}_{i,t} = \Lambda^{-1/2}\mathbf{r}_{i,t}$. With $\mathbf{D} := \mathrm{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,K}^2)$,

$$\mathrm{Var}(\widetilde{R}_{i,k,t}) = e_k^\top \Lambda^{-1/2}\mathbf{D}\,\Lambda^{-1/2} e_k = \sum_{k'=1}^{K} \big[\Lambda^{-1/2}\big]_{kk'}^2 \sigma_{i,k'}^2, \tag{20}$$

establishing equation 11.

Preservation of heteroscedasticity. Since Λ is symmetric positive definite, so is $\Lambda^{-1/2}$. If $\Lambda \neq cI$ for any scalar $c > 0$, at least two diagonal entries of $\Lambda^{-1/2}$ differ. The weight vectors $\mathbf{a}_k := ([\Lambda^{-1/2}]^2_{kk'})_{k'}$ then differ for distinct k, and since $\boldsymbol{\sigma}^2 := (\sigma^2_{i,1}, \ldots, \sigma^2_{i,K})$ is non-constant by Proposition 3.6, the inner products $\mathbf{a}_{k_1}^\top \boldsymbol{\sigma}^2 \neq \mathbf{a}_{k_2}^\top \boldsymbol{\sigma}^2$ for some $k_1 \neq k_2$, confirming heteroscedasticity.

Temporal uncorrelatedness. Since $\Lambda^{-1/2}$ acts on the environment index rather than the time index,

$$\mathrm{Cov}(\widetilde{R}_{i,k,t},\, \widetilde{R}_{i,k,t+s}) = \sum_{k',k''} \left[\Lambda^{-1/2}\right]_{kk'} \left[\Lambda^{-1/2}\right]_{kk''} \mathrm{Cov}(R_{i,k',t},\, R_{i,k'',t+s}). \tag{21}$$

Under the stated assumption, every term vanishes. □

A.4 Proof of Theorem 3.10

Proof. We invoke the multi-environment distribution shift criterion of Xu et al. (2022) and Duong & Nguyen (2023) directly on the observed joint distribution of the augmented system, rather than on an effective noise term, for the reason explained in Remark 4.1: the local dynamic misfit $\delta_{i,k,t}$ is a function of $\mathbf{X}_{k,t}$ and is not in general independent of $\mathrm{Pa}(X_i)$, so the additive noise model exogeneity condition does not apply. Instead, the identifiability argument proceeds in two steps.

Step 1: environment-varying distribution of the augmented system. By Propositions 3.6 and 3.8, the marginal variance $\mathrm{Var}(\widetilde{R}_{i,k,t})$ is non-constant across environments k for every variable i. Since $\widetilde{R}_i$ is an explicit observed node in the augmented DAG and enters the structural equation for X_i via

$$X_i = f_i\Big(\mathrm{Pa}_{\mathcal{G}}(X_i),\, \widetilde{R}_i\Big) + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \tau_i^2), \tag{22}$$

the marginal distribution of X_i in environment k depends on $\mathrm{Var}(\widetilde{R}_{i,k,t})$ and therefore varies across environments. This constitutes the structured cross-environment distribution shift required by Xu et al. (2022), Theorem 1 and Duong & Nguyen (2023), Theorem 3, applied to the augmented observational system $(X_1, \ldots, X_N, \widetilde{R}_1, \ldots, \widetilde{R}_N)$.

Step 2: the bottleneck eliminates residual Markov equivalences. Consider any two distinct admissible DAGs $\mathcal{G}$ and $\mathcal{G}'$ differing in the orientation of exactly one edge, $X_j \to X_i$ in $\mathcal{G}$ versus $X_i \to X_j$ in $\mathcal{G}'$. Under $\mathcal{G}$, the environment-varying distribution of $\widetilde{R}_j$ propagates to X_i via $X_j \to X_i$, producing an environment-variance profile for X_i determined by $\mathrm{Var}(\widetilde{R}_{j,k})$. Under $\mathcal{G}'$ the profile is determined by $\mathrm{Var}(\widetilde{R}_{i,k})$. Since Assumption 3.3 holds for every variable, applying Proposition 3.6 independently to variables i and j gives $\mathrm{Var}(\widetilde{R}_{i,k}) \neq \mathrm{Var}(\widetilde{R}_{j,k})$ for some k, because the two residuals arise from distinct ODE fits with distinct local dynamic misfits. The two orientations, therefore, induce distinct environment-stratified likelihood functions over the augmented system, and the BIC search in Step 2 selects the unique maximiser by Duong & Nguyen (2023), Theorem 3.

Conclusion. Since Steps 1 and 2 hold for every pair of distinct admissible DAGs under the faithfulness assumption, the multi-environment distribution of the augmented system generically identifies G within the admissible DAG class. □

B Implementation details

B.1 Step 1: global SINDy fit and two-stage whitening

The SR3 optimiser solves Equation equation 13 with $\lambda = 0.3$, $\nu = 0.5$, over 100 iterations on the full pooled dataset.

Temporal autocorrelation diagnostic. Residuals are tested for within-environment temporal autocorrelation using the Durbin–Watson statistic at lags $l = 1, \ldots, 10$:

$$\mathrm{DW}_l(\mathbf{R}_i) = \frac{\sum_{t=l+1}^{T}(R_{i,t} - R_{i,t-l})^2}{\sum_{t=1}^{T} R_{i,t}^2}. \tag{23}$$

p-values are adjusted via Benjamini–Yekutieli (Benjamini & Yekutieli, 2001).

Stage 1: within-environment AR pre-whitening. When the Durbin–Watson diagnostic detects statistically significant autocorrelation in $\mathbf{R}_{i,k}$, an AR(p) model is fit to $\mathbf{R}_{i,k}$ along the time index t, with order p selected by BIC (Schwarz, 1978). The raw residuals $\mathbf{R}_{i,k}$ are replaced by the residuals of the AR(p) fit, eliminating within-environment temporal autocorrelation prior to cross-environment mixing. This stage is necessary because the cross-environment whitening in Stage 2 operates on the environment index k and cannot, in general, remove autocorrelation along t; Proposition 3.8 only guarantees preservation of zero autocorrelation when already present, not its removal.

Stage 2: cross-environment whitening. Following AR pre-whitening, the cross-environment transformation $\widetilde{\mathbf{R}}_i = \Lambda^{-1/2}\mathbf{R}_i$ is applied, where Λ is the sample correlation matrix of $\mathbf{R}_i$ across environments. This equalises the cross-environment correlation structure while preserving the heteroscedasticity signal required for identifiability (Proposition 3.8).

Verification. Durbin–Watson statistics are recomputed after both stages to confirm the absence of residual autocorrelation. The identifiability argument of Corollary 3.7 provides a formal rationale for the global ODE fit beyond computational convenience: as the library grows richer, local per-environment fits drive $\delta_{i,k,t}^{(k)}$ to zero, eliminating the cross-environment variance differences that Theorem 3.10 requires. The global fit deliberately accepts reduced per-environment accuracy in order to preserve the structured residual signal on which causal identifiability depends. In addition, the global pooled regression enjoys a numerical stability advantage: the effective sample size is $K \cdot T$ rather than T, so when environments are not collinear, $\kappa(\mathbf{X}_{\text{pooled}}) \leq \kappa(\mathbf{X}_k)$ because additional observations stabilise the regression (Brockwell & Davis, 1991).

B.2 Step 2: joint causal structure learning

Tabu search and Direct LiNGAM maximise BIC over the augmented $2N$-node graph. BIC penalty multipliers: $2\log(n)$ for Tabu search and $3\log(n)$ for Direct LiNGAM, where n is the size of the training sample. The threshold p-value of the conditional independence tests in the PC Stable algorithm is 0.01 in Experiment A and 0.05 in Experiment B; these thresholds are chosen to achieve a graph density between $2 \times 2N$ and $3 \times 2N$ arcs per DAG, which is typical for sparse models in the Bayesian network literature. DAGMA is run with a linear loss (ℓ_2), arc-weight threshold $w_{\text{threshold}} = 0.02$, and $T = 30$ outer iterations (Bello et al., 2022), across all experiments. All variables are standardised to mean zero and unit variance prior to structure learning. Bootstrap aggregation is applied uniformly across all four algorithms: 200 resamples, arc inclusion threshold $\tau = 0.5$, parallel execution on 30 cores.

The full blacklist encodes: *(i)* tier hierarchy for data nodes ($\mathcal{W} \to \mathcal{AP} \to \mathcal{H}$); *(ii)* mirrored tier hierarchy for residual nodes; *(iii)* no cross-variable residual-to-data edges except whitelisted $\widetilde{R}_i \to X_i$; *(iv)* CVD60 cannot cause any residual node; *(v)* no arcs among residual nodes; *(vi)* $NO_2 \to O_3$ orientation excluded.

C Data details

Six clusters were derived by k-means on monitoring station coordinates (Carvalho et al., 2016): Greater London (Cl.1), South West/Exeter (Cl.2), East/Norfolk (Cl.3), Central/Nottingham (Cl.4), North West/Stoke-on-Trent (Cl.5), and North East/Darlington (Cl.6). All variables were aggregated to one daily value per cluster via moving average (54,769 total observations). Missing air pollution values (missing not at random, Little & Rubin 1987) were imputed with the LSTM-based model of Kushwah & Agrawal (2024), outperforming MICE by 15–45% in mean squared error.

Table 7: Variables in the Vitolo et al. (2018) dataset used in ORCD. PM2.5 and CO were removed due to collinearity.

Category	Name	Abbrev.	Unit	Coverage
Health	CVD mortality (60+)	CVD60	%	35 yr
Pollution	Ozone	O_3	$\mu g/m^3$	14 yr
	Particulate (<10μm)	PM10	$\mu g/m^3$	11 yr
	Sulfur dioxide	SO_2	$\mu g/m^3$	12 yr
	Nitrogen dioxide	NO_2	$\mu g/m^3$	12 yr
Weather	Wind speed	WS	m/s	35 yr
	Wind direction	WD	degrees	35 yr
	2-m temperature	T2M	K	35 yr
	Total precipitation	TP	mm	35 yr
	Boundary layer height	BLH	m	35 yr
	Surface solar radiation	SSR	W/m^2s	35 yr

D Full empirical results

D.1 Experiment A: full per-variable results

Table 8 reports predictive Spearman correlations for all eleven variables and six regions under Experiment A (temporal generalisation). Standard error ≈ 0.015 throughout. The predictive Spearman correlation of each whitened residual node with its own variable is below 0.034 in absolute value and is omitted for brevity.

Table 8: Full Experiment A Spearman correlations for all variables and geographical clusters. Each row is one cluster (Cl.1–Cl.6). Cl.1–Cl.6 denote geographical clusters as defined in Table 2.

Algo.	Cl	WS	WD	T2M	TP	BLH	SSR	O3	PM10	SO2	NO2	CVD60
Tabu	1	.775	.301	.721	.467	.761	.830	.552	.555	.581	.536	.626
	2	.775	.312	.721	.459	.761	.830	.563	.555	.581	.537	.626
	3	.774	.159	.533	.463	.769	.811	.550	.554	.582	.536	.626
	4	.775	.315	.725	.459	.762	.831	.558	.552	.581	.539	.626
	5	.773	.293	.719	.455	.758	.831	.551	.553	.566	.542	.625
	6	.775	.315	.722	.470	.762	.830	.557	.558	.579	.535	.626
LiNGAM	1	.772	.310	.725	.464	.759	.834	.569	.554	.586	.535	.623
	2	.772	.312	.728	.464	.760	.833	.569	.553	.586	.530	.623
	3	.775	.312	.723	.470	.762	.831	.567	.556	.583	.530	.625
	4	.567	.308	.716	.321	.371	.770	.568	.557	.579	.529	.625
	5	.545	.306	.713	.320	.352	.774	.567	.555	.588	.533	.623
	6	.773	.310	.723	.465	.761	.831	.571	.556	.585	.533	.626
PC	1	.408	.282	.734	.487	.483	.730	.557	.555	.570	.515	.609
	2	.475	.300	.717	.444	.495	.735	.540	.555	.582	.508	.622
	3	.658	.271	.470	.393	.682	.383	.539	.557	.570	.513	.608
	4	.460	.283	.712	.437	.487	.744	.554	.556	.568	.517	.609
	5	.474	.294	.711	.438	.487	.736	.537	.554	.572	.507	.622
	6	.408	.284	.729	.488	.492	.722	.542	.557	.575	.518	.609
DAGMA	1	.799	.317	.741	.487	.794	.834	.589	.584	.520	.491	.633
	2	.784	.360	.734	.466	.790	.830	.553	.551	.598	.536	.617
	3	.800	.343	.740	.462	.802	.831	.588	.584	.528	.494	.624
	4	.773	.350	.738	.458	.779	.826	.579	.517	.570	.547	.613
	5	.803	.311	.745	.491	.795	.828	.530	.540	.563	.553	.621
	6	.802	.334	.750	.484	.781	.846	.584	.543	.598	.526	.592

D.2 Experiment B: full per-variable results

Table 9 reports mean and standard deviation of Spearman correlations across all 30 cross-cluster pairs under Experiment B (spatial generalisation). The predictive Spearman correlation of each whitened residual node with its own variable remains below 0.11 in absolute value; the whitened residual of CVD60 carries a consistent mean correlation of approximately 0.13 with the predicted CVD60 values, reflecting the genuine causal signal that the BN exploits during parameter learning.

Table 9: Full Experiment B results: mean ± standard deviation of cross-cluster Spearman correlations.

Variable	Tabu	LiNGAM	PC	DAGMA
WS	0.801±0.027	0.802±0.027	0.694±0.138	0.802±0.025
WD	0.332±0.062	0.327±0.063	0.242±0.088	0.311±0.068
T2M	0.745±0.015	0.745±0.014	0.729±0.011	0.743±0.011
TP	0.480±0.051	0.486±0.058	0.457±0.048	0.473±0.058
BLH	0.800±0.036	0.799±0.036	0.660±0.143	0.797±0.037
SSR	0.840±0.016	0.840±0.016	0.783±0.042	0.835±0.020
O_3	0.550±0.085	0.560±0.091	0.530±0.094	0.552±0.086
PM10	0.541±0.070	0.542±0.068	0.536±0.071	0.543±0.071
SO_2	0.560±0.089	0.561±0.090	0.538±0.088	0.562±0.089
NO_2	0.532±0.112	0.536±0.112	0.507±0.115	0.527±0.113
CVD60	0.655±0.018	0.655±0.018	0.643±0.020	0.656±0.017

D.3 Ablation: full per-variable results

Tables 10 and 11 report full per-variable Spearman correlations for the two ablation stages under Experiment A (temporal generalisation). Tables 12 and 13 report the corresponding results under Experiment B (spatial generalisation). The CVD60 collapse under the withoutRi condition ($r \approx 0.01$–0.04 under both protocols versus $r \approx 0.62$–0.66 for ORCD) is the most dramatic signal that the residuals encode genuine dynamic causal information. TP also collapses sharply ($r \approx 0.02$–0.11 versus $r \approx 0.46$–0.49 for ORCD). The irregular values in the DAGMA columns for some regions reflect that algorithm's higher sensitivity to bootstrap aggregation thresholds under the X_i-only graph, not data errors.

Table 10: Experiment A ablation Stage 1 (onlyXi): full per-variable Spearman correlations. Each row is one cluster (Cl.1–Cl.6). Cl.1–Cl.6 denote geographical clusters as defined in Table 2.

Algo.	Cl	WS	WD	T2M	TP	BLH	SSR	O3	PM10	SO2	NO2	CVD60
Tabu	1	.774	.315	.725	.492	.759	.831	.569	.554	.578	.518	.626
	2	.775	.303	.722	.471	.757	.830	.569	.555	.580	.519	.626
	3	.567	.277	.498	.431	.327	.593	.566	.553	.579	.520	.626
	4	.771	.330	.441	.411	.760	.795	.556	.555	.577	.546	.626
	5	.774	.305	.719	.490	.756	.831	.568	.555	.579	.522	.626
	6	.770	.308	.726	.388	.759	.832	.556	.553	.577	.528	.626
LiNGAM	1	.763	.288	.721	.471	.760	.833	.577	.543	.519	.494	.608
	2	.763	.291	.733	.471	.761	.830	.578	.541	.514	.490	.602
	3	.764	.291	.732	.472	.759	.831	.566	.542	.515	.490	.602
	4	.764	.290	.732	.472	.761	.830	.577	.534	.490	.491	.574
	5	.763	.288	.731	.472	.760	.829	.576	.542	.518	.492	.603
	6	.763	.288	.732	.470	.757	.831	.569	.541	.516	.490	.602
PC	1	.777	.267	.740	.437	.767	.822	.555	.549	.538	.514	.623
	2	.777	.269	.735	.426	.771	.820	.538	.551	.534	.510	.623
	3	.774	.271	.734	.432	.769	.819	.537	.554	.530	.513	.608
	4	.774	.266	.734	.424	.771	.820	.539	.553	.529	.514	.609
	5	.777	.273	.732	.426	.770	.820	.535	.555	.527	.535	.622
	6	.776	.258	.734	.428	.767	.814	.539	.555	.536	.516	.609
DAGMA	1	.786	.308	.716	.434	.785	.829	.580	.552	.514	.503	.619
	2	.472	.201	.729	.405	.312	.781	.571	.545	.507	.488	.605
	3	.777	.181	.727	.218	.783	.824	.592	.429	.252	.369	.610
	4	.377	.193	.727	.223	.353	.769	.591	.552	.520	.482	.608
	5	.800	.172	.734	.158	.785	.823	.575	.358	.503	.512	.628
	6	.777	.195	.734	.198	.781	.819	.556	.479	.372	.456	.613

Table 11: Experiment A ablation Stage 2 (withoutRi): full per-variable Spearman correlations. Parameter learning uses $X_i - \hat{R}_i$. Note the near-total collapse of CVD60 ($r \approx 0.01$–0.04) and the sharp reduction in TP ($r \approx 0.02$–0.10) across all algorithms, confirming that the residuals encode genuine dynamic causal information. Cl.1–Cl.6 denote geographical clusters as defined in Table 2.

Algo.	Cl	WS	WD	T2M	TP	BLH	SSR	O3	PM10	SO2	NO2	CVD60
Tabu	1	.704	.308	.600	.088	.698	.793	.545	.528	.555	.522	.015
	2	.703	.299	.601	.097	.699	.792	.551	.529	.557	.512	.013
	3	.490	.278	.146	.028	.293	.566	.548	.528	.556	.512	.028
	4	.703	.329	.109	.098	.704	.735	.542	.532	.556	.544	.014
	5	.700	.302	.599	.102	.695	.793	.550	.528	.554	.515	.016
	6	.701	.303	.601	.066	.697	.792	.540	.528	.554	.523	.010
LiNGAM	1	.698	.285	.595	.089	.708	.792	.572	.532	.489	.490	.012
	2	.697	.285	.599	.087	.709	.792	.573	.532	.488	.484	.011
	3	.696	.288	.598	.090	.700	.792	.548	.531	.490	.484	.022
	4	.699	.288	.598	.089	.709	.791	.572	.532	.487	.485	.002
	5	.697	.285	.596	.085	.708	.791	.572	.532	.488	.486	.015
	6	.695	.284	.598	.090	.699	.792	.550	.532	.487	.485	.013
PC	1	.708	.274	.626	.103	.722	.783	.541	.525	.503	.511	.023
	2	.709	.278	.632	.102	.726	.783	.537	.533	.489	.532	.028
	3	.707	.276	.624	.092	.724	.782	.535	.527	.503	.512	.012
	4	.704	.137	.626	.107	.704	.779	.538	.526	.503	.515	.021
	5	.707	.277	.629	.103	.721	.783	.534	.532	.492	.532	.022
	6	.709	.260	.625	.102	.722	.777	.538	.531	.509	.514	.006
DAGMA	1	.717	.322	.631	.085	.733	.791	.560	.530	.488	.490	.041
	2	.388	.197	.631	.092	.353	.753	.561	.537	.483	.494	.039
	3	.717	.153	.632	.022	.722	.789	.565	.321	.033	.376	.025
	4	.353	.187	.635	.026	.371	.749	.560	.537	.502	.488	.042
	5	.716	.181	.630	.038	.720	.787	.560	.311	.463	.497	.044
	6	.711	.172	.634	.024	.718	.789	.563	.456	.241	.481	.029

Table 12: Experiment B ablation Stage 1 (onlyXi): mean ± standard deviation of cross-cluster Spearman correlations across 30 pairs. Structure and parameter learning on X_i alone.

Variable	Tabu	LiNGAM	PC	DAGMA
WS	0.768±0.075	0.797±0.027	0.787±0.035	0.753±0.116
WD	0.309±0.048	0.320±0.055	0.309±0.052	0.202±0.116
T2M	0.704±0.094	0.736±0.014	0.733±0.010	0.731±0.013
TP	0.455±0.069	0.460±0.068	0.446±0.055	0.345±0.160
BLH	0.743±0.130	0.793±0.038	0.769±0.053	0.717±0.177
SSR	0.796±0.099	0.836±0.019	0.817±0.028	0.821±0.034
O_3	0.548±0.087	0.552±0.090	0.525±0.093	0.553±0.088
PM10	0.523±0.078	0.525±0.074	0.535±0.069	0.487±0.100
SO_2	0.554±0.091	0.548±0.090	0.553±0.094	0.431±0.168
NO_2	0.533±0.114	0.535±0.120	0.511±0.114	0.485±0.127
CVD60	0.634±0.024	0.618±0.036	0.630±0.021	0.647±0.018

D.4 Temporal invariance: full results

Each cluster's whitened residuals are split into two chronological halves of 9,127 observations each; a BN is learned from each half and evaluated on the other. Table 14 reports medians and 5th/95th percentiles across all variable–cluster combinations.

Table 13: Experiment B ablation Stage 2 (withoutRi): mean ± standard deviation of cross-cluster Spearman correlations across 30 pairs. Structure on X_i alone; parameter learning on $X_i - \hat{R}_i$. The near-total collapse of CVD60 ($r \approx 0.02$–0.03) and sharp reduction in TP ($r \approx 0.06$–0.11) confirm that residuals encode the dominant dynamic causal signal.

Variable	Tabu	LiNGAM	PC	DAGMA
WS	0.676±0.105	0.723±0.027	0.706±0.039	0.673±0.127
WD	0.302±0.047	0.311±0.048	0.303±0.050	0.201±0.111
T2M	0.551±0.198	0.630±0.021	0.626±0.019	0.635±0.018
TP	0.071±0.065	0.089±0.051	0.111±0.027	0.058±0.048
BLH	0.683±0.144	0.739±0.039	0.698±0.057	0.672±0.153
SSR	0.755±0.094	0.792±0.022	0.763±0.036	0.782±0.034
O_3	0.536±0.084	0.546±0.085	0.520±0.091	0.548±0.083
PM10	0.510±0.080	0.513±0.078	0.515±0.072	0.453±0.118
SO_2	0.545±0.088	0.539±0.087	0.533±0.086	0.378±0.196
NO_2	0.526±0.111	0.531±0.119	0.504±0.113	0.470±0.132
CVD60	0.021±0.028	0.019±0.031	0.025±0.029	0.033±0.027

Table 14: Temporal invariance of whitened residuals. Medians with 5th/95th percentiles across all variable–cluster combinations. $\rho^{(s,s')}/\rho^{(s',s)}$: cross-predictive symmetry; $\rho^{(s,s')}/\rho^{(s,s)}$: out-of-sample loss. Scaled SHD and SID are between DAGs from the two chronological halves.

Algo.	$\rho^{(s,s')}/\rho^{(s',s)}$	$\rho^{(s,s')}/\rho^{(s,s)}$	Sc. SHD	Sc. SID
Tabu	$0.994\,[0.704, 1.075]$	$0.948\,[0.267, 1.473]$	$0.394\,[0.244, 0.600]$	$0.064\,[0.021, 0.099]$
LiNGAM	$0.995\,[0.884, 1.067]$	$0.894\,[0.269, 1.068]$	$0.410\,[0.268, 0.611]$	$0.033\,[0.014, 0.054]$
PC	$0.995\,[0.873, 1.676]$	$0.934\,[0.198, 1.292]$	$0.455\,[0.258, 0.641]$	$0.071\,[0.040, 0.119]$

D.5 Simulation study: full per-setting results

Dataset names encode $[N][n][d]$ (nodes, observations per environment, edge density). "w" = with ORCD residuals; "w/o" = X_i-only baseline.

Table 15: Simulation results: Tabu search.

Dataset	SHD w	SHD w/o	SID w	SID w/o
[10][100][0.1]	0.816	1.026	0.391	0.416
[10][100][0.5]	0.719	1.125	0.273	0.320
[10][100][0.8]	0.895	1.132	0.355	0.366
[10][200][0.1]	1.133	1.333	0.431	0.462
[10][200][0.5]	0.829	1.143	0.362	0.451
[10][200][0.8]	0.969	1.281	0.387	0.461
[20][100][0.1]	1.783	2.638	0.357	0.432
[20][100][0.5]	1.831	3.225	0.417	0.503
[20][100][0.8]	1.710	2.348	0.420	0.468
[20][200][0.1]	1.556	1.833	0.387	0.359
[20][200][0.5]	1.500	1.926	0.438	0.485
[20][200][0.8]	1.243	1.541	0.343	0.375

Table 16: Simulation results: Direct LiNGAM.

Dataset	SHD w	SHD w/o	SID w	SID w/o
[10][100][0.1]	0.684	0.868	0.341	0.374
[10][100][0.5]	0.813	1.188	0.418	0.480
[10][100][0.8]	0.816	1.158	0.396	0.404
[10][200][0.1]	1.067	1.267	0.427	0.356
[10][200][0.5]	0.943	1.200	0.320	0.408
[10][200][0.8]	1.031	1.344	0.352	0.477
[20][100][0.1]	1.348	3.536	0.368	0.399
[20][100][0.5]	1.239	4.408	0.402	0.538
[20][100][0.8]	1.638	3.333	0.441	0.489
[20][200][0.1]	1.708	2.292	0.299	0.393
[20][200][0.5]	1.471	2.221	0.296	0.363
[20][200][0.8]	1.446	1.824	0.374	0.397

Table 17: Simulation results: PC Stable.

Dataset	SHD w	SHD w/o	SID w	SID w/o
[10][100][0.1]	0.737	0.947	0.357	0.374
[10][100][0.5]	0.844	1.063	0.359	0.418
[10][100][0.8]	0.868	1.105	0.399	0.421
[10][200][0.1]	1.067	1.233	0.444	0.471
[10][200][0.5]	0.914	1.057	0.379	0.447
[10][200][0.8]	1.031	1.250	0.527	0.566
[20][100][0.1]	1.275	1.391	0.344	0.358
[20][100][0.5]	1.296	1.408	0.360	0.371
[20][100][0.8]	1.159	1.333	0.392	0.409
[20][200][0.1]	1.264	1.458	0.289	0.331
[20][200][0.5]	1.118	1.382	0.388	0.423
[20][200][0.8]	1.054	1.216	0.364	0.369

Table 18: Simulation results: DAGMA.

Dataset	SHD w	SHD w/o	SID w	SID w/o
[10][100][0.1]	0.868	1.079	0.338	0.366
[10][100][0.5]	0.813	1.125	0.441	0.449
[10][100][0.8]	0.868	1.000	0.324	0.355
[10][200][0.1]	0.967	1.100	0.467	0.511
[10][200][0.5]	0.943	1.171	0.389	0.467
[10][200][0.8]	1.031	1.094	0.406	0.457
[20][100][0.1]	1.203	1.362	0.316	0.318
[20][100][0.5]	1.239	1.282	0.375	0.417
[20][100][0.8]	0.957	1.072	0.353	0.370
[20][200][0.1]	1.000	1.264	0.284	0.321
[20][200][0.5]	1.044	1.324	0.329	0.346
[20][200][0.8]	0.932	1.135	0.318	0.335

D.6 Residual diagnostic plots

Table 19 reports Bartlett test p-values for cross-environment heteroscedasticity of $\hat{\mathbf{R}}_i$ across the six English regions after Benjamini–Yekutieli adjustment, assessed on the AR pre-whitened residuals prior to cross-environment whitening. Figures 1–3 show the Pearson and Spearman correlation matrices, normal QQ plots, and ACF plots of the fully whitened residuals $\widetilde{\mathbf{R}}_i$ across all six regions. Low off-diagonal Pearson

correlations validate the heteroscedasticity preservation argument of Proposition 3.8; ACF plots confirm temporal independence after both whitening stages. Figures 4 and 5 show structural divergences between each algorithm's learned DAG and the per-cluster consensus DAG under ORCD and the X_i-only baseline, respectively.

Table 19: Bartlett test p-values for cross-environment heteroscedasticity of $\hat{\mathbf{R}}_i$ after Benjamini–Yekutieli adjustment, assessed on AR pre-whitened residuals prior to cross-environment whitening. All variables yield $p \leq 6.07 \times 10^{-19}$, with four variables returning $p = 0$ at machine precision, empirically confirming Assumption 3.3 and Proposition 3.6.

Variable	p-value
BLH	2.35×10^{-22}
CVD60	0
NO2	0
O3	6.07×10^{-19}
PM10	6.50×10^{-169}
SO2	0
SSR	7.80×10^{-9}
T2M	5.92×10^{-62}
TP	1.08×10^{-106}
WD	7.59×10^{-11}
WS	2.10×10^{-36}

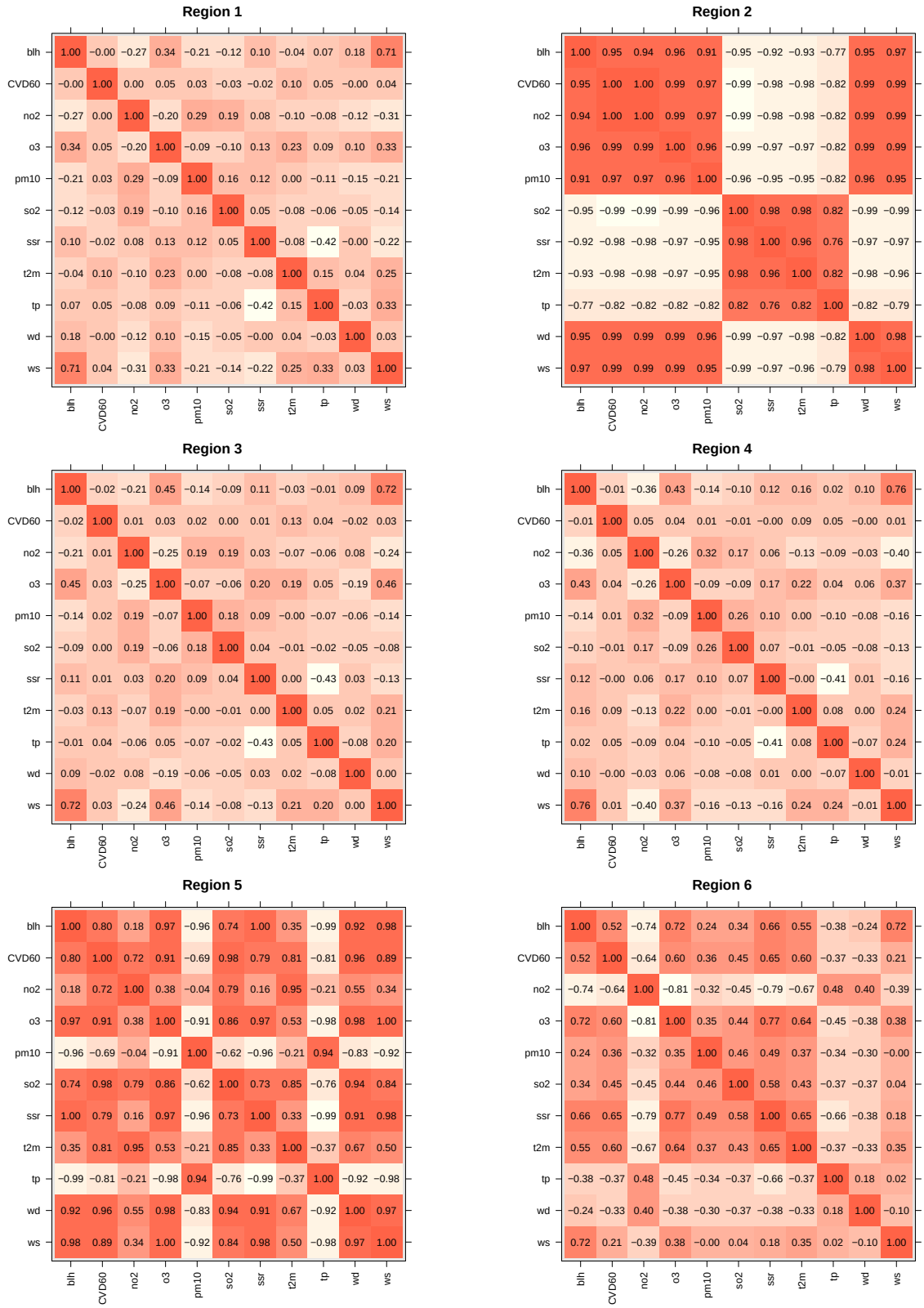

Figure 1: Marginal Pearson correlations between $\hat{R}_i$ for all variables in each geographical region. Low off-diagonal correlations confirm that the whitened residuals are approximately uncorrelated across variables within each region, supporting the independence assumptions of the augmented SCM.

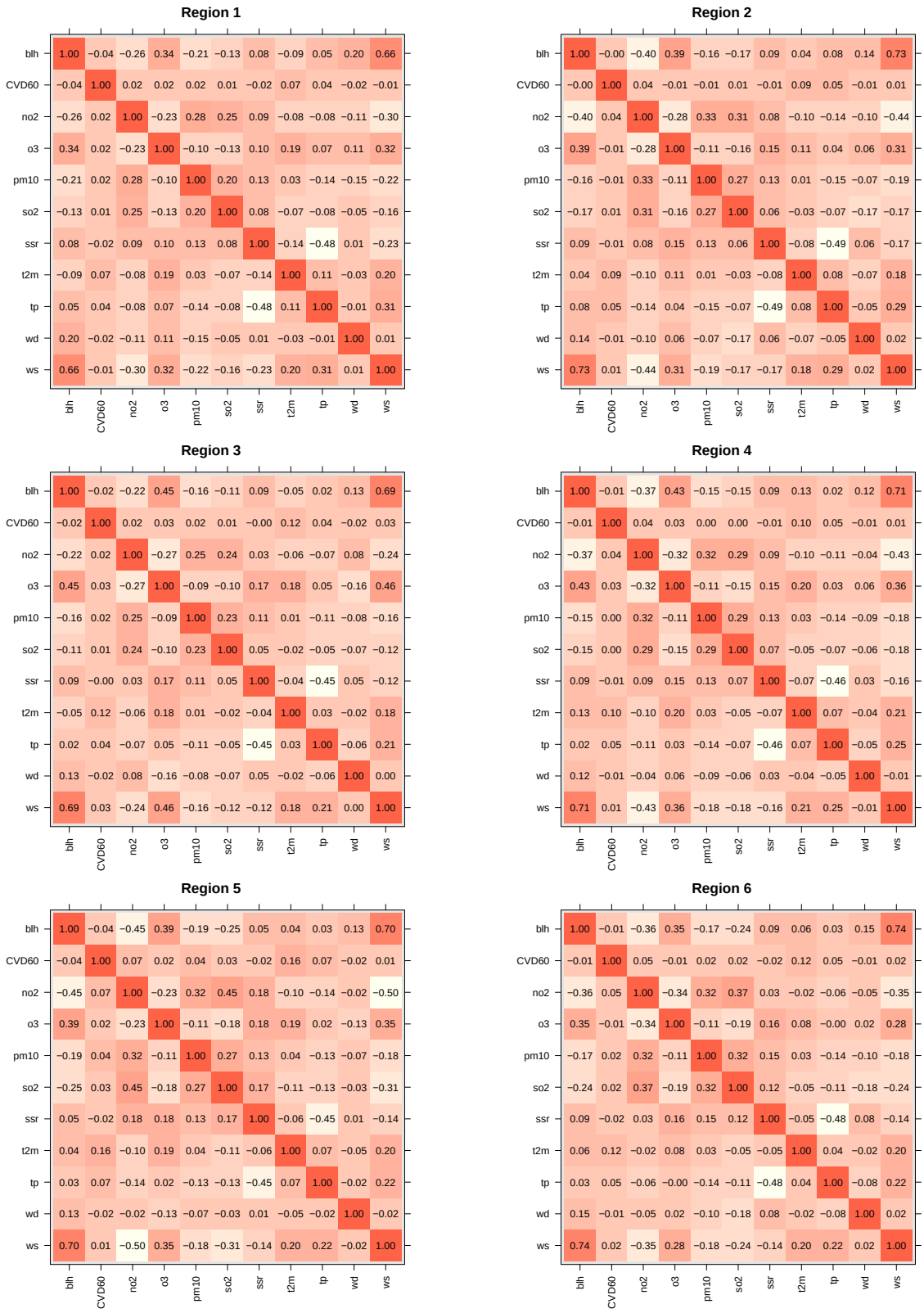

Figure 2: Marginal Spearman correlations between $\hat{R}_i$. Rank correlations are more consistent across regions than Pearson correlations, reflecting robustness to heavy-tailed outliers.

Figure 3: Normal QQ plots of $\hat{R}_i$ for all variables and regions. Distributions are symmetric and bell-shaped; air pollution variables show heavier tails but remain symmetric.

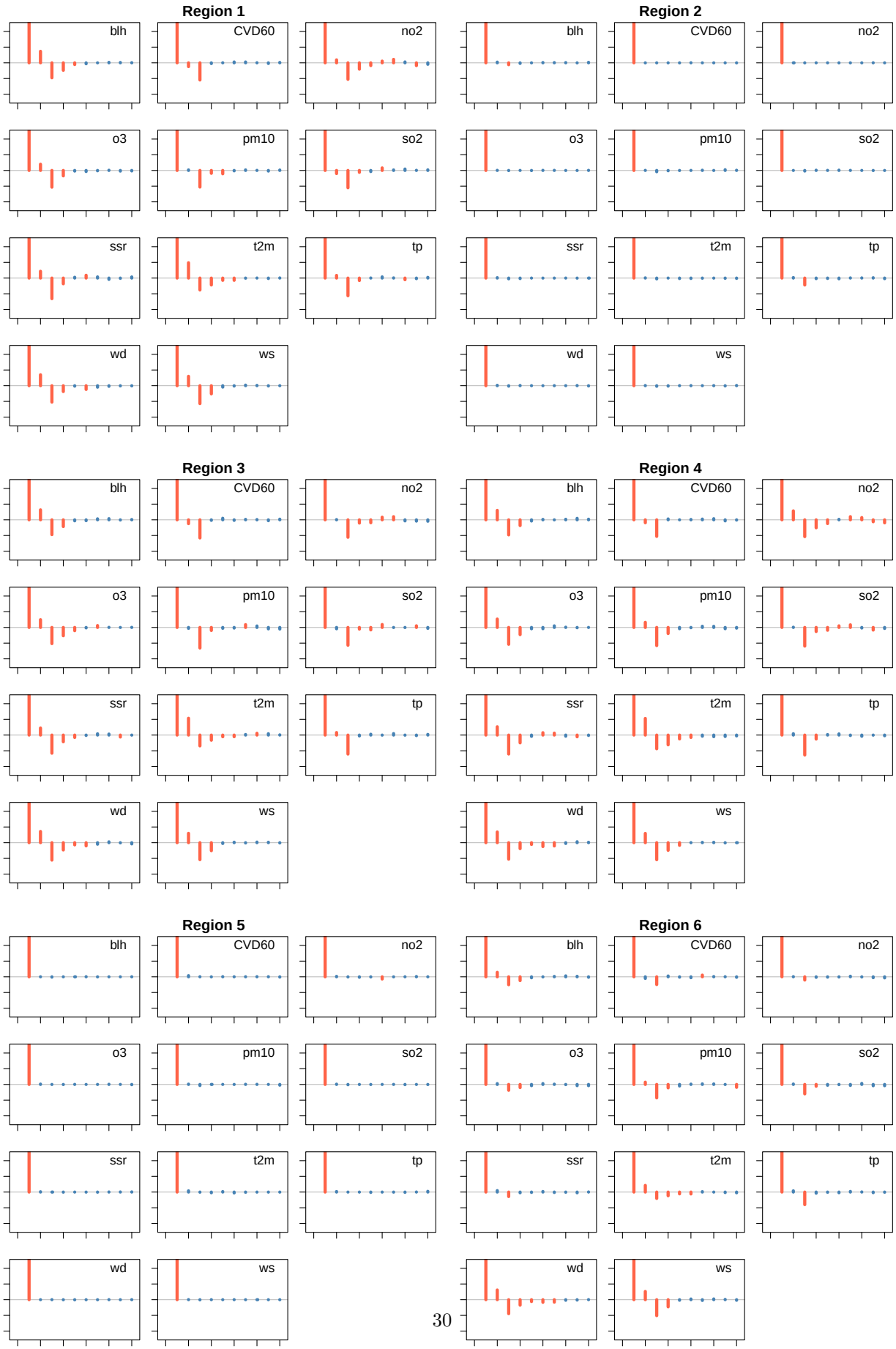
Region 1
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws
Region 2
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws
Region 3
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws
Region 4
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws
Region 5
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws
Region 6
blh
CVD60
no2
o3
pm10
so2
ssr
t2m
tp
wd
ws

Figure 3: ACF plots of $\widetilde{\mathbf{R}}_i$ for all variables and regions after both whitening stages. The absence of statistically significant autocorrelation confirms that the two-stage whitening procedure successfully removes within-environment temporal dependence prior to causal discovery.

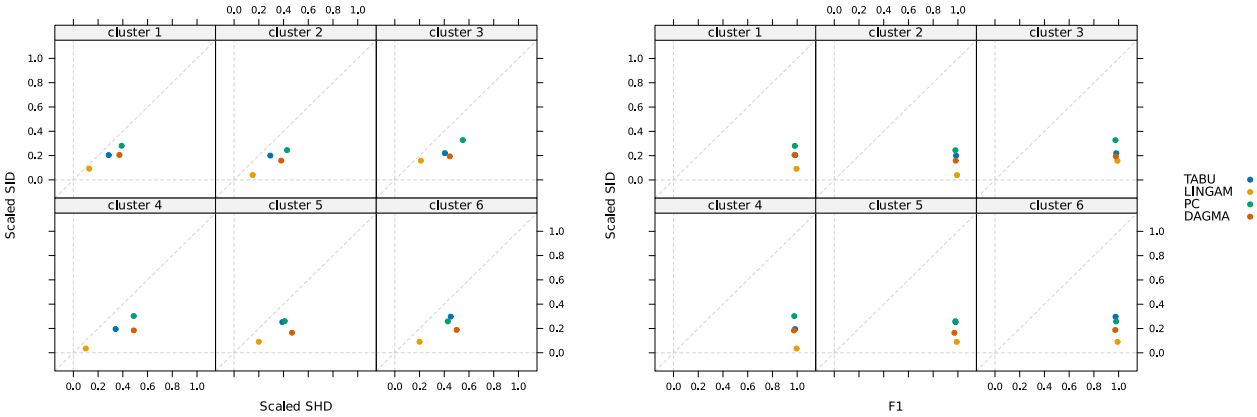

Figure 4: Structural divergences (scaled SHD vs scaled SID, left; F1 vs scaled SID, right) between each algorithm's learned DAG and the per-cluster consensus DAG under ORCD, across all six regions. Points closer to the origin indicate greater structural agreement with the consensus; LiNGAM consistently achieves the lowest divergence.

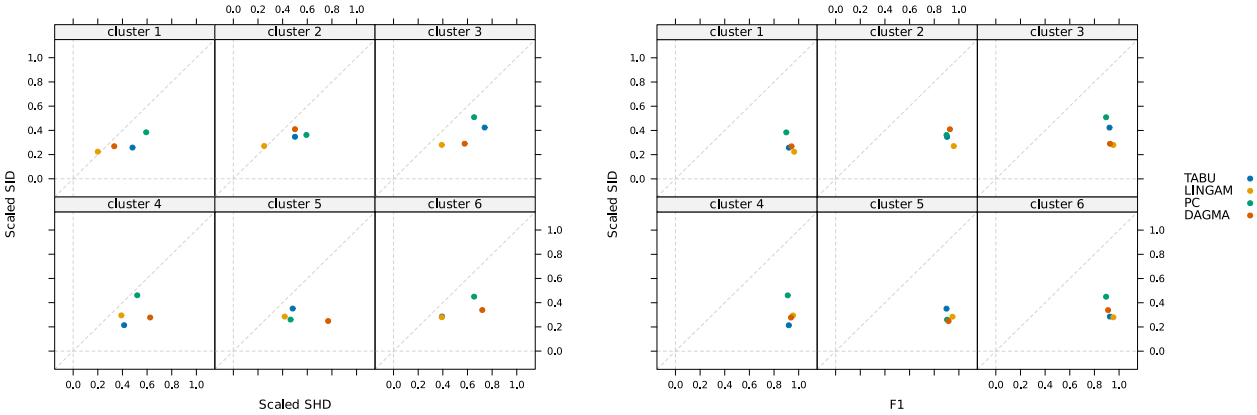

Figure 5: Structural divergences for the X_i-only baseline (same axes as Figure 4). Points are systematically further from the origin than in Figure 4, confirming that omitting the residuals degrades structural accuracy consistently across all algorithms and regions.