Out-of-distribution robustness for multivariate analysis via causal regularisation

Homer Durand **Gherardo Varando** **Nathan Mankovich** **Gustau Camps-Valls**
Image Processing Laboratory (IPL), Universitat de València, València, Spain

Abstract

We propose a regularisation strategy of classical machine learning algorithms rooted in causality that ensures robustness against distribution shifts. Building upon the anchor regression framework, we demonstrate how incorporating a straightforward regularisation term into the loss function of classical multivariate analysis algorithms, such as (orthonormalized) partial least squares, reduced-rank regression, and multiple linear regression, enables out-of-distribution generalisation. Our framework allows users to efficiently verify the compatibility of a loss function with the regularisation strategy. Estimators for selected algorithms are provided, showcasing consistency and efficacy in synthetic and real-world climate science problems. The empirical validation highlights the versatility of anchor regularisation, emphasizing its compatibility with multivariate analysis approaches and its role in enhancing replicability while guarding against distribution shifts. The extended anchor framework advances causal inference methodologies, addressing the need for reliable out-of-distribution generalisation.

1 INTRODUCTION

Data sources in contemporary machine learning applications are often heterogeneous, leading to potential distribution shifts (Sugiyama and Kawanabe, 2012; Liu et al., 2023). This is a particularly relevant problem in computer vision (Csurka, 2017), healthcare (Zhang et al., 2021), Earth and climate sciences (Tuia et al., 2016; Kellenberger et al., 2021), and social sciences (Jin et al., 2023), as variations in data patterns can significantly impact model performance and generalisation in the out-of-distribution (OOD) setting, also referred to as domain generalisation (Liu et al., 2023; Zhou et al., 2023).

In this work, we address the problem of predicting a target $Y \in \mathcal{Y}$ from covariates $X \in \mathcal{X}$ where the training data consists of samples from a subset $\mathcal{Q}$ of a broader class of distributions $\mathbb{Q}$. Our objective is to achieve strong performance across the entire class $\mathbb{Q}$. This challenge can be framed as the following minimax problem:

$$\text{argmin}_{\mathbf{\Theta}} \sup_{\mathcal{Q} \in \mathbb{Q}} \mathbb{E}_{(X,Y)\sim\mathcal{Q}}[\mathcal{L}(X, Y; \mathbf{\Theta})], \tag{1}$$

where $\mathcal{L}(X, Y; \mathbf{\Theta})$ is a loss function over X and Y with parameters $\mathbf{\Theta}$. The choice of the class $\mathbb{Q}$ is particularly important as different classes of distributions lead to various types of robustness. The class $\mathbb{Q}$, often referred to as the uncertainty set, can be constrained using distributional metrics such as f-divergence (see, e.g., Namkoong and Duchi, 2016) or the Wasserstein distance (see, e.g., Esfahani and Kuhn, 2015). The challenge of OOD generalisation is also closely related to the field of transfer learning (see, e.g., Zhuang et al., 2021). For an in-depth review of OOD generalisation, we recommend consulting (Liu et al., 2023; Zhou et al., 2023).

Of particular interest for this work, causal inference can be understood as a distributionally robust estimation where the class $\mathbb{Q}$ is the class of all distributions arising from arbitrarily large interventions on the variables of an SCM (see Meinshausen, 2018; Liu et al., 2023). Notably, the Instrumental Variable (IV) regression exhibits robustness to arbitrarily strong interventions on the instruments under some structural assumptions (Bowden and Turkington, 1990). However, it might be overconservative to pursue algorithms robust to arbitrarily strong interventions, especially when assuming access to knowledge regarding the intervention strength. Rothenhäusler et al. (2018) address this challenge by constraining the intervention to exogenous (so-called

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258.

anchor) variables and by bounding the intervention strength. This approach, known as Anchor Regression (AR), results in a straightforward regularisation of the Ordinary Least Squares (OLS) algorithm, enabling robust linear regression.

More concretely, in its original form, AR can be written, in the population case, as

$$b^{\gamma} = \text{argmin}_{\gamma} \mathbb{E}\left[((I - P_A)(Y - X^T b))^2\right] + \gamma \mathbb{E}\left[(P_A(Y - X^T b))^2\right],$$

with $Y \in \mathbb{R}$. The second term acts as a regulariser that, for $\gamma > 1$, *enforces the minimization of the residuals' projection onto the linear space defined by* A. This term is akin to the two-stage least squares used in IV regression, as it promotes decorrelation between the anchor and the residuals, thereby making the model more robust to distribution shifts arising from interventions on the anchor. This regularisation approach has proven useful in various applications, notably in environmental fields where Oberst et al. (2021) applied it to air-quality prediction, and Sippel et al. (2021) and Székely et al. (2022) utilized it for the Detection and Attribution (D&A) of climate change.

Several AR extensions have been proposed, demonstrating the theoretical versatility of this framework. Notable examples include Kernel Anchor Regression (Shi and Xu, 2023), which extends AR to nonlinear SCMs by working on reproducing kernel Hilbert spaces; targeted anchor regression (Oberst et al., 2021), which exploits prior knowledge of the direction of intervention on the anchor variables; and proxy anchor regression (Oberst et al., 2021), which addresses cases where the anchor is unobserved but proxy variables are available. In Kook et al. (2022), authors introduce a distributional extension of AR, expanding beyond the ℓ_2 loss to accommodate censored or discrete targets.

Previous literature has focused on regression problems with one-dimensional target variables and ordinary least squares as the loss function. Building upon the original approach of AR, we introduce a regularisation strategy applicable to a broader class of algorithms, including some classical MultiVariate Analysis (MVA) techniques (Bilodeau and Brenner, 1999; Arenas-Garcia et al., 2013; Borchani et al., 2015), also known as multi-output algorithms. This contributes on three fronts. First, we extend the applicability of AR algorithms to various loss functions, thereby ensuring compatibility and robustness against distribution shifts induced by constrained interventions on anchor variables. Second, we redefine conventional MVA algorithms within the anchor framework, encompassing techniques such as Partial Least Squares (PLS), Orthonormalized PLS (OPLS), Reduced Rank Regression (RRR), and Multilinear Regression (MLR). This adaptation ensures that their respective loss functions are *anchor-compatible*, and the application of straightforward regularisation provides robustness during testing. This, allows us to leverage the power and versatility of a diverse array of MVA algorithms while providing robustness guarantees against distributional shifts. Third, by introducing a new class of *anchor-compatible* loss functions, we offer a simple and straightforward method to determine whether an algorithm could benefit from anchor regularization and its generalization properties.

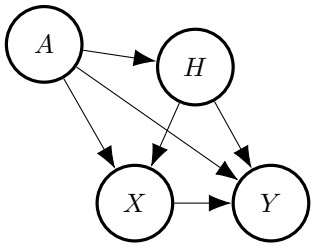

Figure 1: Directed Acyclic Graph (DAG) analyzed in this work, as induced by the Structural Causal Model (SCM) described in (2). The directions of the arrows between X, Y, and H are flexible, provided that the graph remains acyclic. All possible configurations of the DAG can be seen in Fig. 4 in Appendix.

The code used in this work can be found on GitHub.

2 ANCHOR FRAMEWORK

For notation, bold capital letters $\mathbf{U}$ represent matrices, bold lowercase letters $\mathbf{u}$ denote vectors, regular lowercase letters u signify scalars, and uppercase letters U represent random variables (potentially multivariate) which we assume are centered. Given two random vector U and V we define UV as the row-wise stacking operation $UV = (U^T, V^T)^T$. The symbol C_{UV} denotes Kronecker products involving random variables, i.e. $C_{UV} = UV \otimes UV$ and to simplify the notation we write $C_U = U \otimes U$. Since we assume centred random variables $\Sigma_{UV} = \mathbb{E}[C_{UV}]$ and $\Sigma_U = \mathbb{E}[C_U]$. By a slight abuse of notation, we denote $C_{UV|A} = \mathbb{E}[UV|A] \otimes \mathbb{E}[UV|A]$ and $\Sigma_{UV|A}$ its expectation.[1] Again for simplification, we write $\sigma_U = \mathbb{E}[C_U]$ and $\sigma_{UV} = \mathbb{E}[U \otimes V]$. We denote the identity matrix with $\mathbf{I}$, and its dimension is implied by the context when not explicitly stated. Finally, $\|.\|_F$ represents the Froebinius norm.

2.1 The Class of Anchor SCM

In the following, we assume the Directed Acyclic Graph (DAG) in Fig. 1, where $X \in \mathbb{R}^d$ denotes the observed covariates, $Y \in \mathbb{R}^p$ represents the response (or target) variables, $H \in \mathbb{R}^r$ are unobserved variables potentially confounding the causal relation between X and Y and $A \in \mathbb{R}^q$ the exogenous (so-called *anchor*) variables. We assume that the distribution of (X, Y, H) is entailed by the following linear SCM $\mathcal{C}$:

$$\begin{pmatrix} X \\ Y \\ H \end{pmatrix} = \mathbf{B} \begin{pmatrix} X \\ Y \\ H \end{pmatrix} + \varepsilon + \mathbf{M}A, \tag{2}$$

where $\mathbf{B} \in \mathbb{R}^{(d+p+r)\times(d+p+r)}$ and $\mathbf{M} \in \mathbb{R}^{(d+p+r)\times q}$ are unknown constant matrices and $\varepsilon \in \mathbb{R}^{d+p+r}$ is a vector of random noise. We assume that A and ε are independent and have finite variances, ε components are independent, and X and Y have zero mean. The model in Eq. 2 generalises the Instrumental Variables setting, offering a flexible framework that encompasses any linear model—whether or not hidden confounders are present—where an exogenous variable is observed, without requiring an exclusion restriction. Assuming that $(\mathbf{I} - \mathbf{B})$ is invertible, which is satisfied if the linear SCM is acyclic, we can easily express

$$\begin{pmatrix} X \\ Y \end{pmatrix} = \mathbf{D}(\varepsilon + \mathbf{M}A), \tag{3}$$

where $\mathbf{D} \in \mathbb{R}^{(d+p)\times(d+p+r)}$ existence is implied by the existence of $(\mathbf{I} - \mathbf{B})^{-1}$.

2.2 Bound on the Perturbed Covariance

Let us now build upon Eq. (3) to illustrate how distributional robustness can be formulated in a causal context and give some intuition of the proposed regularisation strategy. We show how anchor regularisation can be formulated in terms of covariances Σ_{XY} and $\Sigma_{XY|A}$, instead of squared residuals as originally proposed in Rothenhäusler et al. (2018), the former being more general. This relies on the simple idea that the squared residuals loss is a linear form on Σ_{XY} and anchor regularisation simply adds a constraint on the perturbed covariance matrix Σ_A (see Eq. (5)).

Note that, from Eq. (3), we can easily express Σ_{XY} as

$$\Sigma_{XY} = \mathbf{D}\Sigma_\varepsilon \mathbf{D}^T + \mathbf{D}\mathbf{M}\Sigma_A \mathbf{M}^T \mathbf{D}^T, \tag{4}$$

which combines the variance from exogenous observed (anchors) and unobserved (noise) variables.

[1]Not to be confused with the conditional covariance matrix $\mathrm{Cov}(UV|A)$. Our notation satisfies the total variance-covariance formula $\Sigma_{UV} = \mathbb{E}[\mathrm{Cov}(UV|A)] + \Sigma_{UV|A}$.

Consider the triplet $(X, Y, H) \sim \mathbb{P}_\nu$, where $\mathbb{P}_\nu = \mathbb{P}_{\mathcal{C}}^{do(A\sim\nu)}$ represents the perturbed distribution resulting from intervention (see Peters et al., 2017, Sec. 3.2) on the anchor variable A, where the distribution of ν is assumed to be mean-centered. From Eq. (4), the perturbed variance-covariance matrix $\Sigma_{XY}^{do(A\sim\nu)}$ can be expressed as

$$\Sigma_{XY}^{do(A\sim\nu)} = \mathbf{D}\Sigma_\varepsilon \mathbf{D}^T + \mathbf{D}\mathbf{M}\Sigma_\nu \mathbf{M}^T \mathbf{D}^T,$$

with $\Sigma_\nu = \mathbb{E}[\nu\nu^\top]$. By imposing the constraint $\Sigma_\nu \preceq \gamma\Sigma_A$[2] , we can bound the interventional variance-covariance matrix $\Sigma_{XY}^{do(A\sim\nu)}$ using the training covariances Σ_ε and Σ_A. An intuition for this constraint is that during testing, the covariance matrix of the anchor variable can be expected to be scaled by a factor of at most γ in all directions, thus constraining the strength of the intervention. The anchor regulariser parameter $\gamma \in \mathbb{R}^+$ thus controls the amount of causal regularisation. This yields the specific distribution class:

$$C^\gamma = \{\mathbb{P}_\nu : \Sigma_\nu \preceq \gamma\Sigma_A\}. \tag{5}$$

As demonstrated in Eq. (20) and Eq. (19) in Appendix, both Σ_ε and Σ_A can be expressed in terms of $\Sigma_{XY|A}$ and Σ_{XY} of the training distribution, which leads to the following inequality:

$$\Sigma_{XY}^{do(A\sim\nu)} \preceq \Sigma_{XY} + (\gamma - 1)\Sigma_{XY|A}. \tag{6}$$

Consequently, the interventional covariance matrix $\Sigma_{XY}^{do(A\sim\nu)}$ can be bounded using only the training (or unperturbed) distribution according to Eq. (6). As we will develop in §3.1, algorithms exploiting Σ_{XY} can thus be anchor-regularised, leading to distributionally robust estimators. An example is the ℓ_2 loss, defined for a one-dimensional response variable as $\mathbb{E}[\mathcal{L}(X, Y; \mathbf{b})] = \mathbb{E}[(Y - \mathbf{b}^T X)^2] = \sigma_Y - 2\mathbf{b}^T\sigma_{XY} + \mathbf{b}^T\Sigma_X\mathbf{b}$, which is evidently linear on Σ_{XY}. Similarly, the PLS algorithm (Abdi, 2010) aims to find a pair of matrices $\mathbf{W}_x \in \mathbb{R}^{d\times u}$ and $\mathbf{W}_y \in \mathbb{R}^{p\times v}$ that maximise the covariance between the transformed data $X\mathbf{W}_x$ and $Y\mathbf{W}_y$, i.e. assuming centered X and Y, we aim to maximise $\mathrm{tr}(\mathbf{W}_x^T\Sigma_{XY}\mathbf{W}_y)$ while constraining the columns of $\mathbf{W}_x$ and $\mathbf{W}_y$ to each be orthogonal. In this case, as the loss is also linear on matrix Σ_{XY}, i.e. PLS can be anchor-regularised, cf. §3.1. These observations lead us to define a class of loss functions and demonstrate their robustness over the perturbed distributions C^γ defined in Eq. (5).

[2]Here, $U \preceq V$ means that $V - U$ is positive semi-definite.

3 ROBUSTNESS THROUGH ANCHOR REGULARISATION

Building on the aforementioned idea, we demonstrate that anchor regularisation can be efficiently applied to a broad class of linear algorithms, including several classical MVA algorithms. This approach ensures distributional robustness for algorithms within a class $\mathbb{Q}$, which encompasses distributions arising from bounded interventions on A.

3.1 *Anchor-Compatible* Loss

The distributional robustness properties stemming from intervention on the anchor variables arise from the ability to bound $\Sigma_{XY}^{do(A\sim\nu)}$ using the *observational* covariance matrices Σ_{XY} and $\Sigma_{XY|A}$. Consequently, any loss function expressed as a linear form on Σ_{XY} can attain distributional robustness similar to AR using anchor regularisation.

Definition 3.1 (*Anchor-compatible* loss)**.** We say that a loss function $\mathcal{L}(X, Y; \mathbf{\Theta})$ is *anchor-compatible* if it can be written as $\mathcal{L}(X, Y; \mathbf{\Theta}) = f_{\mathbf{\Theta}}(C_{XY})$, where $f_{\mathbf{\Theta}} : \mathbb{R}^{d\times p} \to \mathbb{R}$ is a linear form[3] and $\mathbf{\Theta}$ are the parameters to be learned.

The following result extends Th. 1 in Rothenhäusler et al. (2018) to the previously defined class of loss functions.

Theorem 3.2. *Let the distribution of* (X, Y, H) *be entailed by the SCM* (2) *and* $\mathcal{L}(X, Y; \mathbf{\Theta})$ *be an* anchor-compatible *loss function. Then for any set of parameters* $\mathbf{\Theta}$*, and any causal regulariser* $\gamma \in \mathbb{R}^+$*, we have*

$$\sup_{\nu\in C^\gamma} \mathbb{E}_\nu[\mathcal{L}(X, Y; \mathbf{\Theta})] = f_{\mathbf{\Theta}}(\Sigma_{XY}) + (\gamma - 1) f_{\mathbf{\Theta}}(\Sigma_{XY|A}), \tag{7}$$

where $C^\gamma = \{\mathbb{P}_\nu : \Sigma_\nu \preceq \gamma\Sigma_A\}$*. Proof can be found in Appendix (3.2).*

The term $f_{\mathbf{\Theta}}(\Sigma_{XY|A})$ could be understood as a causal regulariser enforcing invariance of the loss w.r.t the anchor variable. More intuitively, the anchor regularisation can be seen as adding a penalty term to how much X and Y covary when projected in the span of the anchor variable. The theorem suggests that in the case of distributional shifts observed during testing due to constrained intervention on A, the anchor-regularised estimator $\mathbf{\Theta}^\gamma = \text{argmin}_{\mathbf{\Theta}}\, f_{\mathbf{\Theta}}(\Sigma_{XY}) + (\gamma-1) f_{\mathbf{\Theta}}(\Sigma_{XY|A})$ acts as an optimal estimator, as per the definition provided in Eq. (1). We will proceed to demonstrate the practical application of this general result to standard MVA algorithms.

[3] A linear form is a linear map from a vector space to its field of scalars. The trace function is, for example, a linear form.

3.2 Common Multivariate Analysis Algorithms

By virtue of Th. 3.2, any *anchor-compatible* loss function has theoretically grounded robustness properties to distribution shifts. This is practically relevant, as users can easily verify if a loss function is compatible using Def. 3.1. In that section, we show how this can be applied to a set of commonly used multivariate algorithms in both compatible and incompatible cases. As demonstrated in the experimental section, while non-compatible loss functions might also benefit from anchor regularisation, they exhibit suboptimal robustness properties, and this regularisation should thus be used with care.

A common and straightforward approach to extend Least Squares (LS) regression to multivariate settings is treating responses independently in multioutput regression (Borchani et al., 2015). This approach aims to identify the regression coefficients $\mathbf{W}$ such that $\hat{Y} = \mathbf{W}^T X$, e.g. by solving the LS problem $\widehat{\mathbf{W}} = \text{argmin}_{\mathbf{W}} \|Y - \mathbf{W}^T X\|_F^2$. When the columns of Y exhibit correlation, RRR addresses a similar optimisation problem assuming that the rank of $\mathbf{W}$ is lower than $\min(d, p)$, see Izenman (1975). Alternatively, in OPLS (Arenas-García and Gómez-Verdejo, 2015), we do not assume that $\mathbf{W}$ is low rank. Instead, $\mathbf{W}$ is determined by imposing the constraints $\mathbf{W} = \mathbf{V}^T\mathbf{U}$, where both $\mathbf{U}$ and $\mathbf{V}$ have a rank $\rho \leq \min(d, p)$. Additionally, OPLS is commonly solved through matrix deflation, eigenvalue, or generalised eigenvalue decomposition (Arenas-García and Gómez-Verdejo, 2015). An advantageous property of OPLS is that the columns of the solution $\mathbf{V}$ are ordered based on their predictive performance on Y. This is valuable when predicting Y is not the sole objective, and there is interest in learning a relevant information-retaining subspace of X (here $\mathbf{V}^T X$). These are natural extensions of LS regression and are *anchor-compatible*. See Appendix A for further details and proofs. PLS and CCA are typically preferred when prioritizing learning a pertinent subspace of X over predictive performance. These algoritms aim to maximise the similarity between the latent representations of predictor and target variables. Specifically, PLS assumes the following latent representation $X = \mathbf{P}^T T + N_x$ and $Y = \mathbf{Q}^T U + N_y$ and seeks to maximise the covariance between $\mathbf{W}_x^T X$ and $\mathbf{W}_y^T Y$, while CCA aims to maximise the correlation between the estimated latent spaces. Correlation as a similarity measure ensures equal importance for each dimension in the learned latent space, independent of data variance. PLS, being *anchor-compatible* (see Prop. A.2), benefits from causal regularisation for distributional robustness. In contrast, CCA lacks anchor compatibility due to nonlinearity in its loss function concerning

Table 1: Characterisation of common MVAs with anchor regularisation (loss, constraints, and anchor-compatibility).

	MLR	OPLS	RRR	PLS	CCA
Loss	$\|Y - \mathbf{W}^T X\|_F^2$	$\|Y - \mathbf{U}\mathbf{V}^T X\|_F^2$	$\|Y - \mathbf{W}X\|_F^2$	$-\text{tr}\left(\mathbf{W_x}^T X^T Y \mathbf{W_y}\right)$	$-\text{tr}\left(\mathbf{W_x}^T X^T Y \mathbf{W_y}\right)$
Const.	-	$\mathbf{U}^T\mathbf{U} = \mathbf{I}$	$\text{rank}(\mathbf{W}) = \rho$	$\mathbf{W_x}^T\mathbf{W_x} = \mathbf{I}$, $\mathbf{W_y}^T\mathbf{W_y} = \mathbf{I}$	$\mathbf{W_x}^T C_X \mathbf{W_x} = \mathbf{I}$, $\mathbf{W_y}^T C_Y \mathbf{W_y} = \mathbf{I}$
Comp.	✓	✓	✓	✓	✗

variance-covariance (see §A.3). As illustrated in Fig. 2, CCA shows reduced robustness under interventions on the anchor variable's variance. Recognizing the established equivalence between CCA and OPLS (Sun et al., 2009), it may appear puzzling that one algorithm is *anchor-compatible* while the other is not. We delve deeper into this matter in the Appendix (see §A).

4 TOWARDS INVARIANCE AND CAUSALITY

A direct implication of Th. 3.2 is the connection between anchor regularisation and two well-studied estimators: Instrumental Variable (IV) and Partialling-Out (PA), which we define as follow

$$\boldsymbol{\Theta}^{IV} = \text{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\Sigma_{XY|A}) = \lim_{\gamma \to +\infty} \boldsymbol{\Theta}^{\gamma} \tag{8}$$

$$\boldsymbol{\Theta}^{PA} = \text{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\Sigma^{\perp}_{XY|A}) = \boldsymbol{\Theta}^{0}. \tag{9}$$

Here $\Sigma^{\perp}_{XY|A} = \Sigma_{XY} - \Sigma_{XY|A}$. Note that for the MLR loss function $\boldsymbol{\Theta}^{IV}$ is equivalent to the two-stage least square estimation. A direct implication of Th. 3.2 is that the minimisers of the *anchor-regularised* loss function have the following properties: unregularised parameters $\boldsymbol{\Theta}^1$ are optimal for all $\nu \in C^1$, PA parameters $\boldsymbol{\Theta}^{PA}$ are optimal for all $\nu \in C^0$ and IV parameters $\boldsymbol{\Theta}^{IV}$ are optimal for all $\nu \in C^\infty$. However, although PA and IV estimations are generally used for causal inference, this shows that they provide robustness properties when considering specific interventions on the anchor variables, even if they do not necessarily retrieve the causal parameters. While it is known that for specific sets of interventions, causal parameters have robustness properties (Haavelmo, 1943), it has recently been recognized (Christiansen et al., 2022) that this is not always the case. The following proposition shows the connection between the optimal structural parameters in terms of error and the causal parameters.

Proposition 4.1. *Let us denote* $R(\tilde{\mathbf{B}}) = [X, Y, H]^T - \tilde{\mathbf{B}}[X, Y, H]^T$ *as the error term when reconstructing* (X, Y, H) *with* $\tilde{\mathbf{B}}$. *Given any matrix* $\mathbf{B}^\diamond$ *such that*

$$\sup_{\nu \in C^\gamma} \mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B}^\diamond) \otimes R(\mathbf{B}^\diamond))] \leq \sup_{\nu \in C^\gamma} \mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B}) \otimes R(\mathbf{B}))],$$

it holds that if the loss $f_{\boldsymbol{\Theta}}$ *is* anchor-compatible, *then*

$$\sup_{\nu \in C^\gamma} f_{\boldsymbol{\Theta}}((\mathbf{B} - \mathbf{B}^\diamond)\Sigma^{do(A\sim\nu)}_{XYH}(\mathbf{B} - \mathbf{B}^\diamond)^T) \leq 4 f_{\boldsymbol{\Theta}}(\Sigma_\epsilon) + 4\gamma f_{\boldsymbol{\Theta}}(\mathbf{M}\Sigma_A\mathbf{M}^T). \tag{10}$$

The worst case risk between the causal parameter $\mathbf{B}$ and the optimal one $\mathbf{B}^\diamond$ is bounded but increase linearly with γ. Thus the potential improvement enabled by using *anchor-regularisation* instead of seeking to retrieve causal parameters becomes clearer as intervention strength on the anchor increases.

5 ESTIMATORS

In the previous sections, we derived properties of a regularised version of various MVA methods in the population case. We consider now the sample setting where we are given n i.i.d observations of (X, Y, A). Observations are collectively organised row-wise, in the following matrices, $\mathbf{X} \in \mathbb{R}^{n\times d}$, $\mathbf{Y} \in \mathbb{R}^{n\times p}$ and $\mathbf{A} \in \mathbb{R}^{n\times q}$. For a given *anchor-compatible* loss $\mathcal{L}(X, Y; \boldsymbol{\Theta}) = f_{\boldsymbol{\Theta}}(C_{XY})$, we propose a simple estimator of the population parameters:

$$\widehat{\boldsymbol{\Theta}}^{\gamma} = \text{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\mathbf{S}_{XY}) + (\gamma - 1) f_{\boldsymbol{\Theta}}(\mathbf{S}_{XY|A}), \tag{11}$$

where $\mathbf{S}_{XY}$ and $\mathbf{S}_{XY|A}$ are respectively the empirical estimators of Σ_{XY} and of $\Sigma_{XY|A}$.

Parameter Selection When knowledge about the graph and anchor variable distributional shift is available, one can estimate the optimal γ for robustness against worst-case scenarios. For instance, if the expected anchor variable perturbation strength is up to 1.5, Rothenhäusler et al. (2018) recommends $\gamma = 1.5$. Without prior knowledge, various strategies can be considered: setting γ to a low value (e.g. $\gamma = 2$) might be a suitable default choice, while Sippel et al. (2021) and Székely et al. (2022) suggest selecting γ as a trade-off between prediction error (MSE or R^2 score) and the

correlation between residuals and the anchor variable, or the value of projected residuals in the anchor variable's span. Similar strategies have been proposed in deemed related algorithms (Cortés-Andrés et al., 2022; Li et al., 2022).

High Dimensional Estimators Dealing with high dimensional data challenges estimators and often requires introducing regularisation (Bühlmann and Van De Geer, 2011; Candes and Tao, 2007). Regularisation has been generally introduced to control models' capacity and avoid overfitting (Girosi et al., 1995; Tibshirani, 1996; Hastie et al., 2009), but also to introduce prior (domain) knowledge in the algorithms, as in AR. Adding a regularisation term $\Omega(\|\mathbf{\Theta}\|)$ to the empirical loss in Eq. (11) leads to estimate $\mathbf{\Theta}$ as $\operatorname{argmin}_{\mathbf{\Theta}} f_{\mathbf{\Theta}}(\mathbf{S}_{XY}) + (\gamma - 1) f_{\mathbf{\Theta}}(\mathbf{S}_{XY|A}) + \Omega(\|\mathbf{\Theta}\|)$.

In CCA, OPLS, and PLS regression (or the rank in RRR), the number of components can be viewed as regularisation hyperparameters. Consequently, the optimisation task for anchor-regularised MVA algorithms may involve many hyperparameters. For instance, in Anchor-regularised Reduced Rank Ridge Regression (RRRR) (Mukherjee and Zhu, 2011), three hyperparameters require tuning: ℓ_2 regularisation α, rank ρ, and anchor regularisation γ. Since each hyperparameter optimization addresses different objectives, it may be impractical to optimise all of them simultaneously.

Limitations Since anchor regularisation generally incurs only a minor additive computational cost (see Appendix §C), we believe it could be used in a variety of fields to leverage robustness guarantees. However, a few limitations should be kept in mind when considering its use. First, it should be clear to the reader that *anchor-regularisation* is primarily designed for distributional robustness, not for recovering causal parameters. Also, the data distribution must be entailed in SCM Eq. (3), which also entails linearity assumptions; without this, robustness guarantees are invalid. Lastly, the choice of the regularisation parameter γ is crucial for optimal performance, but finding a sensible value can be challenging in some contexts.

6 EXPERIMENTS

Our theoretical findings are substantiated through an extensive series of experiments that highlight the robustness properties of *anchor-regularised* MVA algorithms. Furthermore, we provide a high-dimensional example to elucidate the process of hyperparameter selection, and demonstrate its practical relevance by showcasing how anchor regularisation enhances climate predictions in a real-world climate science application.

6.1 Simulation Experiments

To demonstrate how anchor regularisation can be applied with different multivariate analysis algorithms, we adapt the experiments from Rothenhäusler et al. (2018) to a multioutput setting. In particular, we assume that the training data (A, X, Y) follows a distribution entailed by the following linear SCM

$$\begin{aligned} \varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y &\sim \mathcal{N}(0,1) \\ A &\leftarrow \varepsilon_A \\ H &\leftarrow \varepsilon_H \\ X &\leftarrow A\mathbf{1}_p^\top + H\mathbf{1}_p^\top + \varepsilon_X \\ Y &\leftarrow \mathbf{W}^T X + H\mathbf{1}_d^\top + \varepsilon_Y, \end{aligned} \tag{12}$$

where X and Y are of dimension $d = p = 10$ ($d = p = 300$ with $n = 200$ in the high dimensional setting), $\mathbf{1} \in \mathbb{R}^d$ is a ones vector, and $\mathbf{W}$ is low rank ρ (see Appendix §D.1 for more details on the experiments settings). We generate test data by intervening on the anchor variable's distribution, setting $A \sim \mathcal{N}(0, t)$, where t is the *perturbation strength*. Each MVA algorithm assumes oracle knowledge of the rank ρ of $\mathbf{C}$, aligning RRR's rank, PLS regression's component count, and CCA accordingly. We consider an IV setting (A only affects Y through X) and we show that even in this scenario, anchor regularised algorithms (with $\gamma = 5$) outperform the PA, IV, and unregularised algorithms. Shown in Fig. 2, all *anchor-compatible* algorithms exhibit robustness to distribution shifts. Anchor-regularised models (with $\gamma = 5$) excel with bounded-strength interventions; PA regularisation is optimal for weak interventions; and IV regularisation for unlimited perturbation strength. Overall, anchor-regularised algorithms maintain stable performance across various perturbation strengths. We also conducted a set of high-dimensional experiments ($p \gg n$ and $d \gg n$) using RRR and MLR both regularised with ℓ_2 norm (ridge regularisation). In both cases, the anchor-regularised algorithms exhibit robustness to increasing perturbation strength (see Fig. 5 in §D.1). This is also the case when A is a confounder (affecting both X and Y, potentially through H), in which case anchor-regularised algorithms present are equally optimal for a wide range of perturbation strength (see Fig. 8 in §D.1). Though not as pronounced as in *anchor-compatible* algorithms, anchor-regularised CCA also displays robustness to perturbations in A, prompting further investigation into its behaviour.

We showcase how hyperparameter selection can be performed through a simulation experiment in a high-dimensional setting using an anchor-regularised RRRR, an ℓ_2-regularized version of RRR with α the ridge hyperparameter and ρ the reduced rank. In this setting, we are given three hyperparameters: γ enforces robust-

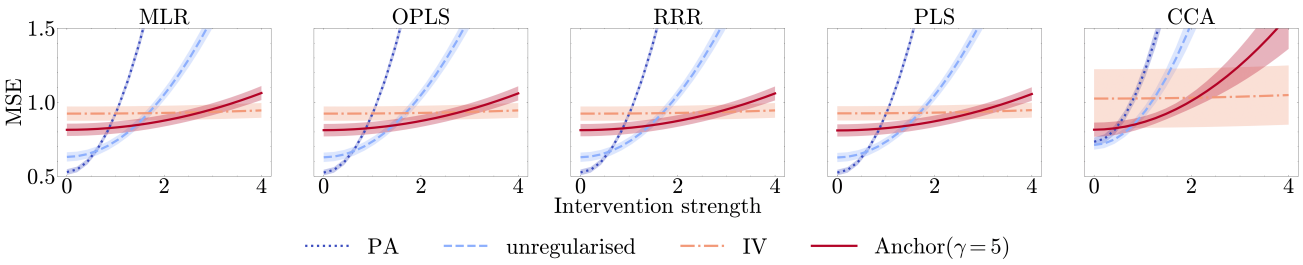

Figure 2: Robustness to increasing perturbation strength for PA, unregularised, IV, anchor-regularised ($\gamma = 5$) algorithms (MLR, OPLS, RRR, PLS and CCA). Anchor versions show robustness in terms of R^2 for bounded intervention strength. Shaded areas represent the range of two standard errors of the mean for running $B = 20$ times the experiment.

ness to intervention on the anchor, and α and ρ aim to maximise prediction performance. Following (Sippel et al., 2021) we select hyperparameters as a trade-off between predictive performance (measured via MSE) and correlation between anchor and residuals. We give equal weights to both objectives but this can be adapted regarding the knowledge available and the application. Thus hyperparameters are selected at testing such that they minimise the combination of the two objectives. Fig. D.1 shows how augmenting regularisation for distributional robustness to anchor intervention often reduces predictive performance in the training sample (Fig. D.1.A), yet the trade-off translates into improved predictive performance in testing samples. More details can be found in §D.2.

6.2 Robust Climate Prediction

We showcase the efficacy of our approach in a real-world application within the Detection and Attribution of Climate Change (D&A) domain. We extend the methodology of Sippel et al. (2021), who utilised AR for robust detection of forced warming against increased climate variability, by applying it to predict multidimensional local climate responses. Given the increased variability observed in recent climate models (Parsons et al., 2020) and observations (DelSole, 2006; Kociuba and Power, 2015; Cheung et al., 2017), our approach aims to ensure robustness against potential underestimation of decadal and multidecadal internal climate variability (Parsons et al., 2020; Mcgregor et al., 2018).

Objective We use a p-dimensional temperature field X_{mod} to predict a temperature response $Y^{\text{forced}}_{\text{mod}} \in \mathbb{R}^p$ to external forcings (such as greenhouse gas emissions, aerosols, solar radiation, or volcanic activity), a crucial step in detecting warming using the fingerprint method (see Hegerl et al., 1996), while ensuring robustness to climate's Decadal Internal Variability (DIV). This leads to the following linear model $Y^{\text{forced}}_{\text{mod}} = X_{\text{mod}}\beta + \epsilon$. We follow Sippel et al. (2021), leveraging multiple climate models from the Climate Model Intercomparison Project (CMIP), both Phase 5 (Taylor et al., 2011) and 6 (Eyring et al., 2016) (CMIP5 and CMIP6). Specifically, we employ four models (CCSM4, NorCPM1, CESM2, HadCM3) characterised by lower-scale DIV to train our MVA algorithm and validate its robustness using anchor regularisation against models exhibiting higher-magnitude DIV (CNRM-CM6-1, CNRM-ESM2-1, IPSL-CM6A-LR).

Estimators We employ RRRR as MVA algorithm. This choice is motivated by the correlation structures present in both the predictors (temperature fields) and the target (temperature response to external forcing), which arise due to spatial autocorrelation. We use DIV as anchor to protect against shifts in long term internal variability. As both α and ρ in RRRR serve to regularise the regression, we optimise them using cross-validation across the training models. We consider two levels of anchor regularisation: $\gamma = 5$ (low regularisation) and $\gamma = 100$ (high regularisation) to showcase how various regularisation strategies lead to different estimated forced response. We evaluate our results regarding two metrics: R^2 score (a standard metric when predicting spatial fields) and mean correlation between the anchor (DIV) and the regression residuals noted as r, similarly to what is done in (Sippel et al., 2021; Székely et al., 2022).

Data We selected 7 models from the CMIP5 and CMIP6 archives, each containing at least 8 members from historical simulations, to ensure accurate estimation of the climate response (refer to Tab. 4 for detailed model information). The data preprocessing procedure for each model involves re-gridding surface air temperature data to a regular $5° \times 5°$ grid and computing yearly anomalies by subtracting the mean surface air temperature for the reference period: years $1850-1900$. The forced response $Y^{\text{forced}}_{\text{mod}}$ is obtained using a standard approach (see Deser et al., 2020, and Appendix E), averaging over all available members in each model.

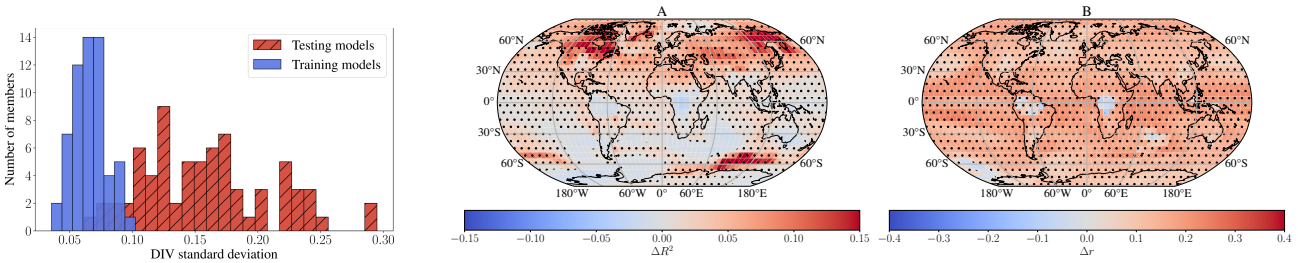

Figure 3: (Left) Standard deviation of DIV among training (red) and testing (blue) model members. (Right A) R^2 score differences between A-RRRR ($\gamma = 5$) and RRRR for test models. (Right B) Differences in residuals-DIV correlation (r) between A-RRRR ($\gamma = 5$) and RRRR for test models. Red and hatched areas indicate where A-RRRR performs better.

Furthermore, we use DIV as a proxy for multidecadal climate internal variability (see Parsons et al., 2020), achieved by removing the global forced response to global temperature and smoothing it using a 10-year running mean. This procedure is commonly employed for estimating the multidecadal variability of the climate system (see Sippel et al., 2021; Deser et al., 2020, for further details). Finally, all data are standardised.

Evaluation Procedure We categorise the models into two groups (low and high variability) based on the distribution of the standard deviation of their DIV among the members of each model. As depicted in Fig. 3 (Left), a significant proportion of members used for training display a notably smaller magnitude of DIV. Hyperparameters (α, ρ) are selected through a leave-half-models-out cross-validation procedure. Given $k = 4$ training models, we randomly select 2 models for training and 2 models for validation, and repeated $B = 10$ times. We use 500 samples from each model to ensure equal weights are given to each training model in the learning algorithm, despite variations in the number of members. For each of the B train/validation splits, we train an RRRR model and select the hyperparameters that yield the highest averaged R^2 score across the B sampling splits. R^2 provides a performance measure comparable across regions and is typically used when predicting spatial variables (see e.g. Sippel et al., 2019). Hyperparameter α is selected from 20 candidates in a logarithmically-spaced sequence ($\lambda \in [1; 10^6]$), and ρ is selected from 21 candidates in a linearly-spaced integer sequence ($\rho \in [300; 600]$).

Results For both anchor-regularised RRRR (A-RRRR) with $\gamma = 5$ and $\gamma = 100$, we observe an improvement in both metrics in the testing set (respectively 0.537 and 0.533 of R^2 for A-RRRR and 0.506 for unregularised RRRR), while experiencing a slight decrease of performance in the training samples (0.510 of R^2 for unregularised RRRR and respectively 0.500 and 0.487 for both A-RRRR). As the latter strongly protects against shifts in DIV, we notice that its mean correlation between residuals and DIV is lower, albeit with a slight decrease in the R^2 score (see Table 5 in Appendix §E). In Fig. 3, we observe that anchor-regularised RRRR performs better in the northern hemisphere, particularly in regions where the residuals are highly correlated with DIV and in the northern hemisphere where unregularised RRRR performs poorly (see Figs. 14 and 15 in Appendix §E). Conversely, we observe a decrease in the performance of A-RRRR in regions where unregularised RRRR already performs well or where residuals are anti-correlated with DIV. The latter challenge might be addressed by considering regional proxies of internal variability instead of a global one, which does not always accurately represent local internal variability.

These results suggest the potential of using anchor-regularised multi-output algorithms in D&A studies to detect and attribute local responses to external forcings, a fundamental problem in climate science.

6.3 Robust Air-Quality Prediction

We conduct an experiment to evaluate the effectiveness of anchor regularization in ensuring robust predictions under temporal distribution shifts in pollution variables. Our approach is compared against established robust prediction methods.

Objective We assess the performance of anchor regularization in handling temporal shifts by testing models across different seasons, where the relationship between meteorological variables and air quality indicators is known to vary.

Data We use the Air Quality dataset from Vito (2008), leveraging meteorological variables (Temperature, Humidity, and Relative Humidity) as predictors and 10 air quality indicators as outcomes (see Table 6 in Appendix). Seasons are treated as categorical variables, and models are tested on unseen seasons to

Table 2: Mean, Median, Max, and Min MSE values for different models, along with the number of cases where they performed better than LR. IRM shows a better worst-case risk control but A-PLS has a lower median MSE and has better performances than LR in 22 out of the 24 cases.

Model	Mean MSE	Median MSE	Max MSE	Min MSE	Better than LR
IRM	**0.6852**	0.7297	**0.8655**	**0.4580**	20
CVP	0.8171	0.7907	1.1872	0.5764	21
Ridge	0.7699	0.7518	0.9969	0.5525	17
LR	0.8213	0.7914	1.2146	0.5785	-
AR	0.7804	0.7406	1.2146	0.5058	16
A-Ridge	0.7809	0.7627	0.9969	0.5877	13
A-PLS	0.7231	**0.7251**	0.9931	0.4788	**22**

evaluate their robustness under temporal shifts.

Estimators Several linear model-based estimators are compared. Invariant Risk Minimization (IRM) follows the approach of Arjovsky et al. (2019). Conditional Variance Penalties (CVP), as proposed by Heinze-Deml and Meinshausen (2021), employing quantile binning for outcome discretization. We also consider Linear Regression (LR) and Ridge Regression (Ridge) as baseline models. Finally, we include Anchor Multioutput Regression (AR) and Anchor PLS Regression (A-PLS) to assess the impact of anchor regularization.

Evaluation Procedure Models are trained on two seasons, validated on a third for hyperparameter tuning, and tested on the fourth, cycling through all 24 possible combinations. The hyperparameters considered include anchor regularization, selected from a logarithmically-spaced sequence ($\gamma \in [10^{-10}; 10^{-10}]$); the number of PLS components, varying between one and three; IRM and CVP regularization parameters, selected from a logarithmic grid ($\lambda \in [10^{-3}; 1]$); and Ridge regularization, also selected from a logarithmic grid ($\lambda \in [10^{-10}; 10^{-10}]$). All hyperparameters are selected from 20 candidates. Mean Squared Error (MSE) is used as the primary evaluation metric. We report the mean, median, maximum, and minimum MSE across the 24 combinations, as well as the number of splits where a model outperforms LR in mean MSE.

Results All out-of-distribution methods—IRM, CVP, AR, and A-PLS—demonstrate superior performance compared to linear regression (LR). A-PLS shows competitive results alongside IRM, surpassing LR in 22 out of 24 combinations and achieving better median MSE. However, its slightly higher mean and maximum MSE suggest that IRM remains the preferred choice for minimising worst-case risk. Still, anchor-regularised PLS exhibits significantly better control over worst-case risk (Max MSE) compared to simple multi-output anchor regression, highlighting the potential of anchor regularisation to leverage advanced multivariate analysis techniques in practical, real-world scenarios.

7 CONCLUSION

In this study, we extend the causal framework of anchor regression proposed by Rothenhäusler et al. (2018), demonstrating the versatility of their regularisation approach across a wide range of MVA algorithms. Given the significant challenge of generalising models to OOD data in machine learning, we advocate for integrating anchor regularisation into a broader class of MVA algorithms, including RRR, OPLS, or PLS regression, particularly when domain knowledge suggests a bounded intervention strength on the anchor. Moreover, we highlight that anchor regularisation offers an interesting trade-off between prediction performance and invariance to distribution shifts, addressing concerns regarding over-conservativeness in some cases. Future theoretical advancements will entail extending the formulation nonlinear cases using kernel methods. Furthermore, an interesting avenue for exploration is understanding how statistical learning algorithms incompatible with anchor regularisation behave when this regularisation is applied. On the application front, we explore how anchor regularisation, when combined with various MVA algorithms, can be used to detect forced responses to more complex climate variables (e.g., precipitation, temperature, and their extremes) and attribute these responses to external forcing sources, such as greenhouse gas or aerosol emissions (anthropogenic factors), or natural phenomena like solar radiation and volcanic activity. Additionally, we demonstrate in an air quality prediction problem that, in the context of categorical anchors, anchor-regularised MVA yields results that are competitive with state-of-the-art approaches, such as IRM. We anticipate a broad development and application of anchor-regularised methods across various fields of science.

Acknowledgements

Authors acknowledge funding from the Horizon project AI4PEX (grant agreement 101137682), the Horizon project ELIAS (grant agreement 101120237), and the European Research Council (ERC) support under the ERC Synergy Grant USMILE (grant agreement 855187).

References

Masashi Sugiyama and Motoaki Kawanabe. *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. Towards out-of-distribution generalization: A survey. *arXiv preprint*, 2023. arXiv:2108.13624.

Gabriela Csurka. *Domain Adaptation in Computer Vision Applications*. Springer, 01 2017.

Haoran Zhang, Natalie Dullerud, Laleh Seyyed-Kalantari, Quaid Morris, Shalmali Joshi, and Marzyeh Ghassemi. An empirical framework for domain generalization in clinical settings. In *Proceedings of the Conference on Health, Inference, and Learning*, CHIL '21, page 279–290, New York, NY, USA, 2021. Association for Computing Machinery.

Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57, 2016.

Benjamin Kellenberger, Onur Tasar, Bharath Bhushan Damodaran, Nicolas Courty, and Devis Tuia. *Deep Domain Adaptation in Earth Observation*, chapter 7, pages 90–104. John Wiley & Sons, Ltd, 2021.

Ying Jin, Kevin Guo, and Dominik Rothenhäusler. Diagnosing the role of observable distribution shift in scientific replications. *arXiv preprint*, 2023. arXiv:2309.01056.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(4): 4396–4415, apr 2023.

Hongseok Namkoong and John C Duchi. Stochastic gradient methods for distributionally robust optimization with f-divergences. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.

Peyman Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171, 05 2015.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, jan 2021.

Nicolai Meinshausen. Causality from a distributional robustness point of view. In *2018 IEEE Data Science Workshop (DSW)*, pages 6–10, 2018.

Roger J Bowden and Darrell A Turkington. *Instrumental variables*. Cambridge university press, 1990.

Dominik Rothenhäusler, Peter Bühlmann, Nicolai Meinshausen, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 83, 01 2018.

Michael Oberst, Nikolaj Thams, Jonas Peters, and David Sontag. Regularizing towards causal invariance: Linear models with proxies. In *International Conference on Machine Learning*, pages 8260–8270. PMLR, 2021.

Sebastian Sippel, Nicolai Meinshausen, Enikő Székely, Erich Fischer, Angeline G. Pendergrass, Flavio Lehner, and Reto Knutti. Robust detection of forced warming in the presence of potentially large climate variability. *Science Advances*, 7(43):eabh4429, 2021.

Enikő Székely, Sebastian Sippel, Nicolai Meinshausen, Guillaume Obozinski, and Reto Knutti. Robust detection and attribution of climate change under interventions. *arXiv preprint*, 2022. arXiv:2212.04905.

Wenqi Shi and Wenkai Xu. Learning nonlinear causal effect via kernel anchor regression. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 1942–1952. PMLR, 31 Jul–04 Aug 2023.

Lucas Kook, Beate Sick, and Peter Bühlmann. Distributional anchor regression. *Statistics and Computing*, 2022.

Martin Bilodeau and David Brenner. *Theory of multivariate statistics*. Springer Science & Business Media, 1999.

Jeronimo Arenas-Garcia, Kaare Brandt Petersen, Gustavo Camps-Valls, and Lars Kai Hansen. Kernel multivariate analysis framework for supervised subspace learning: A tutorial on linear and kernel multivariate methods. *IEEE Signal Processing Magazine*, 30(4): 16–29, 2013.

Hanen Borchani, Gherardo Varando, Concha Bielza, and Pedro Larranaga. A survey on multi-output

regression. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 5, 07 2015.

Jonas Peters, Dominik Janzing, and Bernhard Schlkopf. *Elements of Causal Inference: Foundations and Learning Algorithms*. The MIT Press, 2017. ISBN 0262037319.

Hervé Abdi. Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:97–106, 01 2010.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5(2):248–264, 1975.

Jerónimo Arenas-García and Vanessa Gómez-Verdejo. Sparse and kernel OPLS feature extraction based on eigenvalue problem solving. *Pattern Recognition*, 48, 05 2015.

Liang Sun, Shuiwang Ji, Shipeng Yu, and Jieping Ye. On the equivalence between canonical correlation analysis and orthonormalized partial least squares. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence*, IJCAI'09, page 1230–1235, San Francisco, CA, USA, 2009. Morgan Kaufmann Publishers Inc.

Trygve Haavelmo. The statistical implications of a system of simultaneous equations. *Econometrica*, 11: 1, 1943.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):6614–6630, oct 2022. ISSN 0162-8828.

Jordi Cortés-Andrés, Gustau Camps-Valls, Sebastian Sippel, Enikő Székely, Dino Sejdinovic, Emiliano Diaz, Adrián Pérez-Suay, Zhu Li, Miguel Mahecha, and Markus Reichstein. Physics-aware nonparametric regression models for earth data analysis. *Environmental Research Letters*, 17(5):054034, 2022.

Zhu Li, Adrián Pérez-Suay, Gustau Camps-Valls, and Dino Sejdinovic. Kernel dependence regularizers and gaussian processes with applications to algorithmic fairness. *Pattern Recognition*, 132:108922, 2022.

Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313 – 2351, 2007.

Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

Trevor Hastie, Robert Tibshirani, Jerome H Friedman, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Ashin Mukherjee and Ji Zhu. Reduced rank ridge regression and its kernel extensions. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(6):612–622, 2011.

Luke A. Parsons, M. Kathleen Brennan, Robert C.J. Wills, and Cristian Proistosescu. Magnitudes and spatial patterns of interdecadal temperature variability in CMIP6. *Geophysical Research Letters*, 47(7), 2020.

Timothy DelSole. Low-frequency variations of surface temperature in observations and simulations. *Journal of Climate*, 19(18):4487–4507, 2006.

Greg Kociuba and Scott B. Power. Inability of CMIP5 models to simulate recent strengthening of the walker circulation: Implications for projections. *Journal of Climate*, 28(1):20–35, 2015.

Anson H. Cheung, Michael E. Mann, Byron A. Steinman, Leela M. Frankcombe, Matthew H. England, and Sonya K. Miller. Comparison of low-frequency internal climate variability in CMIP5 models and observations. *Journal of Climate*, 30(12):4763–4776, 2017.

Shayne Mcgregor, Malte Stuecker, Jules Kajtar, Matthew England, and M. Collins. Model tropical atlantic biases underpin diminished pacific decadal variability. *Nature Climate Change*, 8, 06 2018.

Gabriele C. Hegerl, Hans von Storch, Klaus Hasselmann, Benjamin D. Santer, Ulrich Cubasch, and Philip D. Jones. Detecting greenhouse-gas-induced climate change with an optimal fingerprint method. *Journal of Climate*, 9(10):2281–2306, 1996.

Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93 (4):485–498, 2011. Publisher: American Meteorological Society.

Veronika Eyring, Sandrine Bony, Gerald A. Meehl, Catherine A. Senior, Bjorn Stevens, Ronald J. Stouffer, and Karl E. Taylor. Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, 9(5):1937–1958, may 2016.

C. Deser, F. Lehner, K. Rodgers, Toby Ault, T. Delworth, P. DiNezio, A. Fiore, Claude Frankignoul, J. Fyfe, Daniel Horton, Jennifer Kay, Reto Knutti,

N. Lovenduski, J. Marotzke, K. McKinnon, S. Minobe, James Randerson, J. Screen, Isla Simpson, and Miao Ting. Insights from earth system model initial-condition large ensembles and future prospects. *Nature Climate Change*, 10:277–286, 04 2020.

Sebastian Sippel, Nicolai Meinshausen, Anna Merrifield, Flavio Lehner, Angeline G. Pendergrass, Erich Fischer, and Reto Knutti. Uncovering the forced climate response from a single ensemble member using statistical learning. *Journal of Climate*, 32(17): 5677–5699, 2019.

Saverio Vito. Air Quality. UCI Machine Learning Repository, 2008.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint*, 2019. arXiv:1907.02893.

Christina Heinze-Deml and Nicolai Meinshausen. Conditional variance penalties and domain shift robustness. *Machine Learning*, 110(2):303–348, 2021.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Michele Svanera, Mattia Savardi, Sergio Benini, Alberto Signoroni, Gal Raz, Talma Hendler, Lars Muckli, Rainer Goebel, and Giancarlo Valente. Transfer learning of deep neural network representations for fmri decoding. *Journal of Neuroscience Methods*, 328:108319, 2019. ISSN 0165-0270.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   The mathematical setting and assumptions are clearly stated in sections 2 and 3.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [No]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes] The code is available at this GitHub repo.

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes] The assumptions are stated in section 3.
   (b) Complete proofs of all theoretical results. [Yes] Proofs are available in sections A and B.
   (c) Clear explanations of any assumptions. [Yes] Intuition on the set of assumptions is stated in sections 2 and 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes] Experiments and simulations were run on a standard laptop with 16 GB of RAM and 12th Gen Intel(R) Core(TM) i7-12700H CPU. The presented algorithm do not need special computational infrastructure to be applied to standard data-sets as shown in this paper

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes] Algorithms used are stated in section D.1.
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

A ROBUSTNESS OF COMMON MULTIVARIATE ANALYSIS ALGORITHM

In this section, we aim to prove the compatibility and incompatibility of some standard multivariate analysis algorithms. We assume that the (X, Y, H, A) distribution be entailed in Eq. (2).

A natural extension of Ordinary Least Squares (ℓ_2) regression to the multioutput case is to solve an OLS problem for each output, which is equivalent to solving the optimisation problem:

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times p}} \mathbb{E}\|Y - \mathbf{W}^T X\|_F^2, \tag{13}$$

also known as multilinear (or multioutput) regression. Extra constraints on the regression coefficients can be added, leading, for example, to Reduced Rank Regression by constraining $\mathbf{W}$ to be of rank $\rho < \min(p, d)$. This constraint resembles constraining $\mathbf{W}$ to have the form $\mathbf{V}^T\mathbf{U}$, with $\mathbf{U} \in \mathbb{R}^{\rho\times d}$ and $\mathbf{V} \in \mathbb{R}^{\rho\times p}$.

This formulation is closely related to the formulation of Orthogonalised Partial Least Squares with an extra constraint on the ordering of the columns of $\mathbf{V}^T X$ based on their predictive performance on Y. Therefore, we derive the proof of anchor compatibility for the multilinear regression setting, which directly extends to anchor compatibility of OPLS and Reduced Rank Regression.

Since both OPLS and RRR aim to achieve the same objective as defined in Eq. (14) with different constraints on regression coefficients $\mathbf{W}$, we prove the general anchor compatibility of MLR, but this extends to OPLS and RRR.

Proposition A.1 (Multilinear, Reduced Rank and Orthonormalised Partial Least Square Regression are *anchor-compatible.*)**.** *The algorithm minimising the expectation of the following loss function:*

$$\mathcal{L}(X, Y; \mathbf{\Theta}) = \|Y - \mathbf{W}^T X\|_F^2 \tag{14}$$

is anchor-compatible*. Here* $\|.\|_F$ *defines the Froebinius norm.*

Proof. The proof of this proposition is straightforward, observing that the loss can be written as:

$$\mathcal{L}(X, Y; \mathbf{\Theta}) = \sigma_Y - \mathbf{W}^T\sigma_{XY} + \mathbf{W}^T\sigma_X\mathbf{W}, \tag{15}$$

which is linear over the variance-covariance Σ_{XY}.

□

Another type of Multivariate Analysis algorithm is not aimed at minimising prediction error. Instead, it focuses on maximising the similarity between latent representations of the predictors $X\mathbf{W}_x$ and the target $Y\mathbf{W}_y$. A standard measure of this similarity is covariance, leading directly to the Partial Least Squares (PLS) algorithm, for which we now present anchor compatibility.

Proposition A.2 (Partial Least Square regression is *anchor-compatible.*)**.** *The PLS Regression algorithm maximising the expectation of the following loss function is* anchor-compatible*:*

$$\mathcal{L}(X, Y; \mathbf{W}) = \frac{tr(\mathbf{W_x}^T C_{XY} \mathbf{W_y})}{\sqrt{tr(\mathbf{W_x}^T \mathbf{W_x})}\sqrt{tr(\mathbf{W_y}^T \mathbf{W_y})}}. \tag{16}$$

Proof. The loss function Eq. 16 is clearly linear with respect to the variance-covariance Σ_{XY}, which is sufficient to ensure the anchor compatibility of PLS regression. □

On the other hand, one could consider correlation as a measure of the similarity between the learned latent spaces, aiming to account for potential differences in the variance of each variable and thus assigning equal weight to each dimension of the latent space. This approach is generally known as Canonical Correlation Analysis (CCA) or PLS mode B. By considering the variance of the latent representation of the predictors and the target, we sacrifice linearity with respect to the variance-covariance of X and Y (as illustrated in Eq. 17), making CCA incompatible with anchor regularisation.

Example A.3 (Canonical Correlation Analysis is not *anchor-compatible.*). The Canonical Correlation Analysis solving the optimisation problem

$$\mathcal{L}(X;\mathbf{W}) = \frac{\operatorname{tr}({\mathbf{W_x}}^T \Sigma_{XY} \mathbf{W_y})}{\sqrt{\operatorname{tr}({\mathbf{W_x}}^T \Sigma_X \mathbf{W_x})}\sqrt{\operatorname{tr}({\mathbf{W_y}}^T \Sigma_Y \mathbf{W_y})}} \tag{17}$$

is not *anchor-compatible* as it is not linear over the variance-covariance matrix. This explains the different behaviour taken by anchor regularisation in Fig. 2 by the anchor-regularised CCA.

This is illustrated in Fig. 2, where we can observe that anchor-regularised CCA exhibits a distinct behavior compared to anchor-compatible MVA algorithms. It would be of interest to investigate to which extent the incompatibility of CCA with anchor regularisation impacts its distributional robustness properties and if its anchor regularisation could still be of interest.

OPLS Formulations and their Relation to CCA There are two versions of OPLS: the standard eigenvalue decomposition (EVD) and the generalized eigenvalue (GEV) formulations Arenas-García and Gómez-Verdejo (2015). In our work, we implement EVD-OPLS whose optimization problem is

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \|Y - \mathbf{U}\mathbf{V}^T X\|_F^2 \\ &\text{s.t.} \quad \mathbf{U}^T\mathbf{U} = \mathbf{I}. \end{aligned}$$

EVD-OPLS is linear in C_{XY} and is, therefore, an anchor-compatible loss. On the other hand, the GEV-OPLS loss is

$$\begin{aligned} &\min_{\mathbf{U},\mathbf{V}} \|Y - \mathbf{U}\mathbf{V}^T X\|_F^2 \\ &\text{s.t.} \quad \mathbf{V}^T C_X \mathbf{V} = \mathbf{I} \end{aligned}$$

Viewing the optimization for $\mathbf{V}$ as a generalized eigenvalue decomposition problem by absorbing the constraint into the loss gives us the equivalent optimization

$$\min_{\mathbf{v}} \frac{\mathbf{v}^T X^T Y Y^T X \mathbf{v}}{\mathbf{v}^T C_X \mathbf{v}}.$$

This is clearly *not* linear in C_{XY} because it is not linear in $C_X = X^T X$.

Furthermore, GEV-OPLS and CCA are shown to be equivalent up to an orthogonal rotation (see Sun et al., 2009, Theorem 2), and by similar reasoning, CCA is also not anchor-compatible.

B Proofs

B.1 Proof of Theorem 3.2

Proof. Let's first note that from the SCM Eq. 2 we have the following decomposition:

$$\begin{aligned} f(C_{XY}) &= f(\mathbf{D}C_\varepsilon\mathbf{D}^T + \mathbf{D}\mathbf{M}C_A\mathbf{M}^T\mathbf{D}^T) \\ &= f(\mathbf{D}C_\varepsilon\mathbf{D}^T) + f(\mathbf{D}\mathbf{M}C_A\mathbf{M}^T\mathbf{D}^T) \end{aligned}$$

by linearity of f. Thus when taking the supremum of the expectation of $f(C_{XY})$ over C^γ, we get

$$\sup_{\nu\in C^\gamma} \mathbb{E}_\nu[f(C_{XY})] = f(\mathbf{D}\Sigma_\varepsilon\mathbf{D}^T) + \sup_{\nu\in C^\gamma} f(\mathbf{D}\mathbf{M}\Sigma_\nu\mathbf{M}^T\mathbf{D}^T),$$

since $f(\mathbf{D}\Sigma_\varepsilon\mathbf{D}^T)$ is not affected by the intervention. Here $\mathbb{E}_\nu[.] = \mathbb{E}_{(X,Y)\sim\mathbb{P}_\nu}[.]$ the expectation for the intervention distribution $\mathbb{P}_\nu$. Using the definition of C^γ leads to

$$\sup_{\nu\in C^\gamma} \mathbb{E}_\nu[f(C_{XY})] = f(\mathbf{D}\Sigma_\varepsilon\mathbf{D}^T) + \gamma f(\mathbf{D}\mathbf{M}\Sigma_A\mathbf{M}^T\mathbf{D}^T). \tag{18}$$

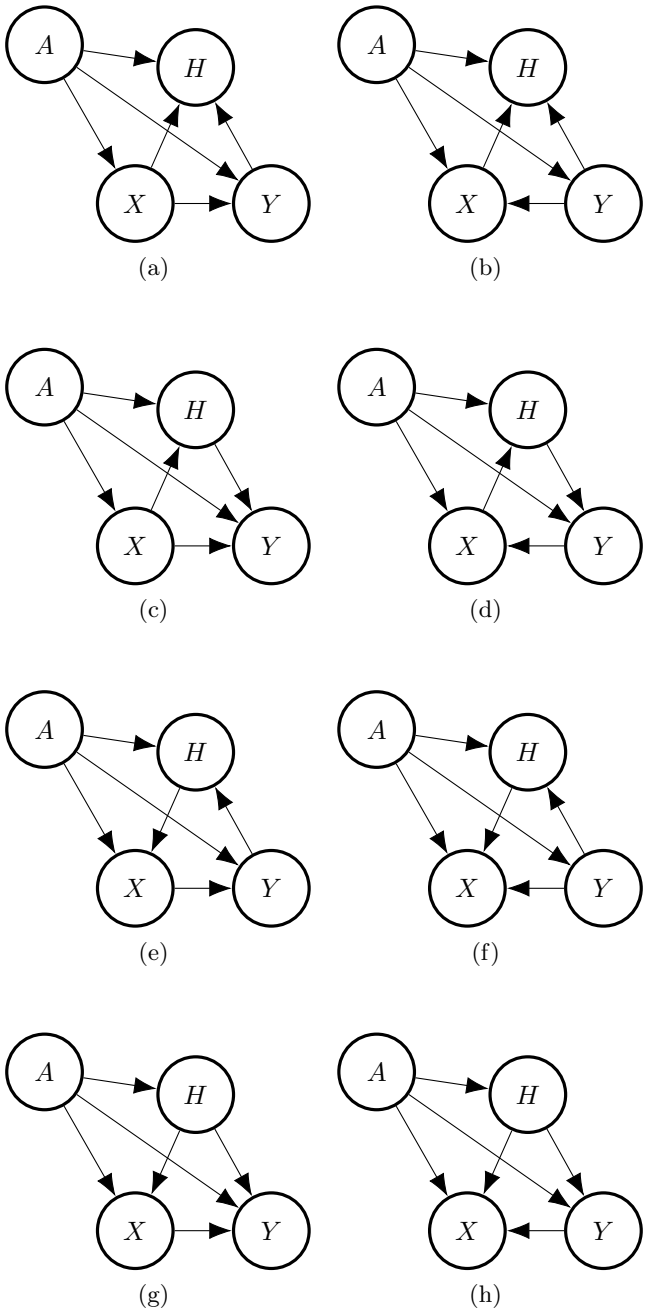

Figure 4: All possible DAGs compatible with the anchor framework.

Let's now note that for $(A, X, Y) \sim \mathbb{P}_{\text{train}}$ we have

$$\begin{aligned}\mathbb{E}_{\text{train}}\left[\begin{pmatrix}X\\Y\end{pmatrix}|A\right] &= \mathbf{D}\,\mathbb{E}_{\text{train}}[\varepsilon|A] + \mathbf{DM}\,\mathbb{E}_{\text{train}}[A|A]\\ &= \mathbf{DM}A,\end{aligned}$$

where $\mathbb{P}_{\text{train}}$ is the training distribution and $\mathbb{E}_{\text{train}}[.] = E_{(X,Y)\sim\mathbb{P}_{\text{train}}}[.]$ the expectation with regard to $\mathbb{P}_{\text{train}}$. As ε is mean centred and independent of A. Thus, we can write

$$f(C_{XY|A}) = f(\mathbf{DM}C_A\mathbf{M}^T\mathbf{D}^T)$$

by linearity of f. Taking its expectation over the training distribution, we get

$$\mathbb{E}_{\text{train}}[f(C_{XY|A})] = f(\mathbf{DM}\Sigma_A\mathbf{M}^T\mathbf{D}^T), \tag{19}$$

which is similar to the right term of Eq. (18) as $\mathbb{E}_{\text{train}}[f(C_{XY|A})] = f(\Sigma_{XY|A})$. A similar reasoning leads to

$$\mathbb{E}_{\text{train}}[f(C_{XY})] - \mathbb{E}_{\text{train}}[f(C_{XY|A})] = f(\mathbf{D}\Sigma_\varepsilon\mathbf{D}^T). \tag{20}$$

Pluging-in Eq. (19) and Eq. (20) in Eq. (18) we get

$$\begin{aligned}\sup_{\nu\in C^\gamma}\mathbb{E}_\nu[f(C_{XY})] &= \mathbb{E}_{\text{train}}[f(C_{XY})] + (\gamma-1)\,\mathbb{E}_{\text{train}}[f(C_{XY|A})]\\ &= f(\Sigma_{XY}) + (\gamma-1)f(\Sigma_{XY|A}),\end{aligned}$$

which concludes the proof. □

B.2 Proof of Proposition

Proof. The loss can easily decomposed as

$$\begin{aligned}\mathbb{E}[f_{\boldsymbol{\Theta}}(R(\mathbf{B}^\diamond)\otimes R(\mathbf{B}^\diamond))] &= \mathbb{E}[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))] + f_{\boldsymbol{\Theta}}((\mathbf{B}-\mathbf{B}^\diamond)\Sigma_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T)\\ &\quad - \mathbb{E}[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes(\mathbf{B}-\mathbf{B}^\diamond)\begin{pmatrix}X\\Y\\H\end{pmatrix})]\\ &\geq \mathbb{E}[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))] + f_{\boldsymbol{\Theta}}((\mathbf{B}-\mathbf{B}^\diamond)\Sigma_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T)\\ &\quad - \mathbb{E}[f_{\boldsymbol{\Theta}}((R(\mathbf{B})\otimes R(\mathbf{B})))]^{\frac{1}{2}}\mathbb{E}[f_{\boldsymbol{\Theta}}(((\mathbf{B}-\mathbf{B}^\diamond)\Sigma_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T))^{\frac{1}{2}}\end{aligned}$$

Where the first equality is a simple bias-variance decomposition and we use Cauchy-Schartz inequality for the inequality. By assumptions

$$\sup_{\nu\in C^\gamma}\mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B}^\diamond)\otimes R(\mathbf{B}^\diamond))] \leq \sup_{\nu\in C^\gamma}\mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))].$$

This leads to

$$\begin{aligned}\mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))] &\geq \mathbb{E}_\nu[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))] + f_{\boldsymbol{\Theta}}((\mathbf{B}-\mathbf{B}^\diamond)\Sigma^{do(A:=\nu)}_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T)\\ &\quad - \mathbb{E}_\nu[f_{\boldsymbol{\Theta}}((R(\mathbf{B})\otimes R(\mathbf{B})))]^{\frac{1}{2}}f_{\boldsymbol{\Theta}}(((\mathbf{B}-\mathbf{B}^\diamond)\Sigma^{do(A:=\nu)}_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T))^{\frac{1}{2}}\end{aligned}$$

which is equivalent to

$$\begin{aligned}f_{\boldsymbol{\Theta}}((\mathbf{B}-\mathbf{B}^\diamond)\Sigma^{do(A:=\nu)}_{XYH}(\mathbf{B}-\mathbf{B}^\diamond)^T) &\leq 4\mathbb{E}[f_{\boldsymbol{\Theta}}(R(\mathbf{B})\otimes R(\mathbf{B}))]\\ &\leq 4f_{\boldsymbol{\Theta}}(\Sigma_\varepsilon) + 4\gamma f_{\boldsymbol{\Theta}}(\mathbf{M}\Sigma_A\mathbf{M}^T).\end{aligned}$$

This concludes the proof. □

B.3 Proof of Equivalence for Transformed Data

Proof. From

$$\begin{aligned}\tilde{\mathbf{X}} &= (\mathbf{I} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{X} \\ \tilde{\mathbf{Y}} &= (\mathbf{I} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{Y},\end{aligned} \tag{21}$$

we can easily derive that

$$\begin{aligned}\mathbf{S}_{\tilde{X}\tilde{Y}} =& \frac{1}{n-1}(\tilde{\mathbf{X}}\tilde{\mathbf{Y}})^T(\tilde{\mathbf{X}}\tilde{\mathbf{Y}}) \\ =& \frac{1}{n-1}((\mathbf{XY} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}}\mathbf{XY})^T(\mathbf{XY} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}}\mathbf{XY}))\end{aligned}$$

As $\Pi_{\mathbf{A}}$ is an orthogonal projection, we have that $(\mathbf{XY})^T(\Pi_{\mathbf{A}}\mathbf{XY}) = \Pi_{\mathbf{A}}(\mathbf{XY})^T\mathbf{XY} = \mathbf{S}_{XY|A}$, leading to

$$\begin{aligned}\mathbf{S}_{\tilde{X}\tilde{Y}} =& \mathbf{S}_{XY} + 2(\sqrt{\gamma} - 1)\mathbf{S}_{XY|A} + (\sqrt{\gamma} - 1)^2\mathbf{S}_{XY|A} \\ =& \mathbf{S}_{XY} + (\gamma - 1)\mathbf{S}_{XY|A}.\end{aligned}$$

Thus, by linearity of $f_{\boldsymbol{\Theta}}$, the minimiser of $f_{\boldsymbol{\Theta}}(\mathbf{S}_{\tilde{X}\tilde{Y}})$ is given by

$$\begin{aligned}\boldsymbol{\Theta}^\gamma &= \operatorname{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\mathbf{S}_{\tilde{X}\tilde{Y}}) \\ &= \operatorname{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\mathbf{S}_{XY}) + (\gamma - 1)f_{\boldsymbol{\Theta}}(\mathbf{S}_{XY|A}).\end{aligned}$$

This conclude the proof. □

C ESTIMATORS: FURTHER DETAILS

Simpler Formulation For computational and practical reasons, $\widehat{\boldsymbol{\Theta}}^\gamma$ can be estimated by transforming the training data:

$$\tilde{\mathbf{X}} = (\mathbf{I} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{X} \tag{22}$$

$$\tilde{\mathbf{Y}} = (\mathbf{I} + (\sqrt{\gamma} - 1)\Pi_{\mathbf{A}})\mathbf{Y}, \tag{23}$$

where $\Pi_{\mathbf{A}} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T \in \mathbb{R}^{n\times n}$ projects onto the column space of $\mathbf{A}$, assuming that $\mathbf{A}^T\mathbf{A}$ is invertible and $\mathbf{X}$ and $\mathbf{Y}$ are centered. Finally, build the estimator $\widehat{\boldsymbol{\Theta}}^\gamma = \operatorname{argmin}_{\boldsymbol{\Theta}} f_{\boldsymbol{\Theta}}(\mathbf{S}_{\tilde{X}\tilde{Y}})$, where $\mathbf{S}_{\tilde{X}\tilde{Y}}$ is the empirical variance-covariance of the projected data defined in Eq. (22). Proof of the equivalence of using transformed data as in Eq. (22) is equivalent to Eq. (11) is available in Supplementary B.3. We give computational complexity and consistency results in Supplementary C.

Consistency From the law of large numbers, the empirical covariance matrix of (X, Y, A) converges to its empirical covariance matrix, i.e., $\mathbf{S}_{XY} \xrightarrow[n\to\infty]{} \Sigma_{XY}$. Thus, the consistency of the anchor-regularised estimator depends on the consistency of the original estimator. For continuous functions f, we have by continuity that $\widehat{\boldsymbol{\Theta}}^\gamma \xrightarrow[n\to\infty]{} \boldsymbol{\Theta}^\gamma$ (see Rothenhäusler et al., 2018, section 4.1).

Computational Complexity The computational cost of projecting X and Y into the span of A involves computing the covariance matrix of A (of cost $O(nr^2)$), inverting it (of cost $O(r^3)$), and two matrix products (of cost $O(nrp)$ and $O(nr^2)$), resulting in a total complexity of $O(nr^2 + nrp + r^3)$. Assuming the unregularised MVA algorithm's computational cost is c, the complexity of its anchor-regularised version is $O(c + nr^2 + nrp + r^3)$. We note that this generally incurs an affordable additive computational cost, as it is equivalent to two linear regressions. The only main concern arises with high-dimensional A, which is not addressed in this work.

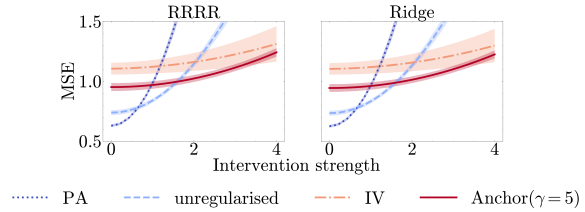

Figure 5: Experiments in high-dimensional setting with $d = p = 300$ and $n = 200$. We can see that both Multi-output Ridge Regression and Reduced rank Ridge Regression are optimal for a wide range of perturbation strength. Shaded areas represent the range of two standard errors of the mean for running $B = 20$ times the experiment.

D SIMULATION EXPERIMENTS: FURTHER DETAILS

D.1 Experiment Setting

The results of the simulation experiments are obtained as follows. The training data are sampled from the following SCM

$$\begin{aligned} \varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y &\sim \mathcal{N}(0,1) \\ A &\leftarrow \varepsilon_A \\ H &\leftarrow \varepsilon_H \\ X &\leftarrow A\mathbf{1}_p^\top + H\mathbf{1}_p^\top + \varepsilon_X \\ Y &\leftarrow \mathbf{W}^T X + H\mathbf{1}_d^\top + \varepsilon_Y, \end{aligned} \tag{24}$$

and the testing data are generated by modifying ε_A such that $\varepsilon_A \sim \mathcal{N}(0, t)$. The perturbation strength t is varied over a linear sequence $t \in [0, 4]$ with 20 steps. We repeat each experiment $B = 10$ times by sampling $n = 300$ training and testing samples of (A, X, Y). We plot the average Mean Squared Errors in Fig. 2 and Fig. 7. Both A and H are one-dimensional, while X and Y are 10-dimensional. The matrix $\mathbf{W}$ is generated as a low-rank (of rank ρ) matrix such that $\mathbf{W} = \mathbf{UV}$ with $\mathbf{A} \in \mathbb{R}^{d\times\rho}$ and $\mathbf{B} \in \mathbb{R}^{\rho\times p}$. The matrices $\mathbf{\Gamma}_i$ are hThe coefficients of $\mathbf{A}$ and $\mathbf{B}$ are sampled uniformly between 1 and 2 and are normalised such that their sum is 1.

We employ the algorithms CCA, PLS, and MLR implemented in the *scikit-learn* library (BSD 3-Clause License) Pedregosa et al. (2011), respectively `CCA`, `PLSRegression` and `LinearRegression` learners. We use the code available at https://github.com/rockNroll87q/RRRR (MIT License) Svanera et al. (2019); Mukherjee and Zhu (2011) for the Reduce Rank Regression algorithm and our implementation of OPLS based on Arenas-García and Gómez-Verdejo (2015) using an eigenvalue decomposition as we use the constraint $A^T A = \mathbf{I}$.

High-Dimensional Setting We also conduct a high-dimensional experiment to evaluate the performance of anchor-regularized algorithms when the dimensionality of X and Y exceeds the sample size. We generate data by sampling $n = 200$ instances of X and Y, each with a dimensionality of $d = p = 300$. The rank of $\mathbf{W}$ is set to $\rho = 100$. We compare the results between Ridge Regression and Reduced Rank Ridge Regression. As shown in Figure 5, both methods demonstrate robustness across a wide range of perturbation strengths.

Non-Gaussian Noise Experiments To assess the robustness of our results presented in paragraph 6.1, we conducted the same toy model experiments with noise ε_A, ε_H, ε_X, and ε_Y following an exponential distribution (Fig. 7.A) and a gamma distribution (Fig. 7.B) with scales 1.

As observed in Fig. 7, anchor-regularised models remain optimal for a wide range of perturbation strengths, except for anchor-regularised CCA. The behavior of anchor-regularised CCA is explained by its incompatibility with anchor regularisation.

Confounding Anchor Experiment We also reproduced experiments from equation 12 in Rothenhäusler et al. (2018) to demonstrate how anchor-regularised MVA algorithms exhibit interesting robustness properties when

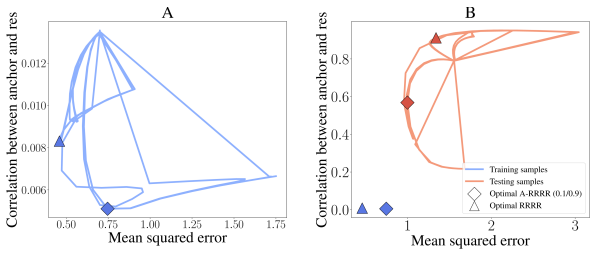

Figure 6: Pareto front for hyperparameters (γ, α, ρ) in RRRR and A-RRRR. For each pair (γ, α), ρ is selected to minimise the weighted sum of objectives (equal weights given here).

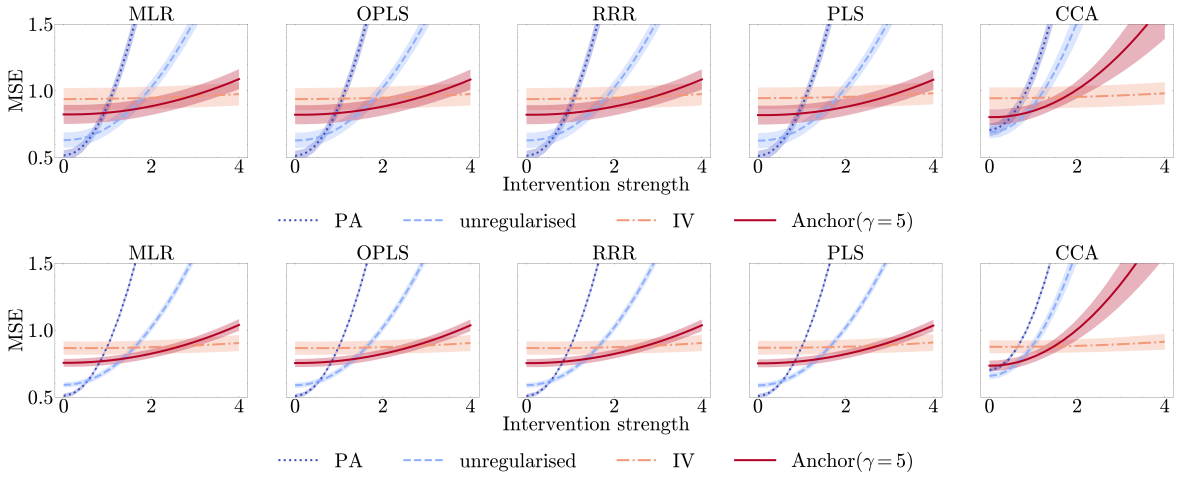

Figure 7: Similar experiments as in 6.1 with noise sampled from non-gaussian distributions. (Top) Exponential distribution of scale1, (Bottom) Poisson distribution of parameter 1. We see that all the *anchor-compatible* algorithms are robust to increasing perturbation strength. Shaded areas represent the range of two standard errors of the mean for running $B = 20$ times the experiment.

the anchor has a confounding effect. These results are illustrated in Fig. 8. Here, the training distribution is encapsulated in the following Structural Causal Model (SCM):

$$\begin{aligned} \varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y &\sim \mathcal{N}(0,1), \\ A &\leftarrow \varepsilon_A, \\ H &\leftarrow A + \varepsilon_H, \\ X &\leftarrow A + H + \varepsilon_X, \\ Y &\leftarrow \mathbf{W}^T X + H + \varepsilon_Y, \end{aligned} \tag{25}$$

with again testing distribution sampled by setting $\varepsilon_A \sim \mathcal{N}(0, t)$ where t is the perturbation strength and $\mathbf{W}$ being lower rank.

We can see that all algorithms exhibit robustness properties for a very large range of perturbation strength. It is interesting to note that CCA presents a similar behavior as the *anchor-compatible* algorithms. It would be interesting to investigate in which specific cases CCA gives robustness properties.

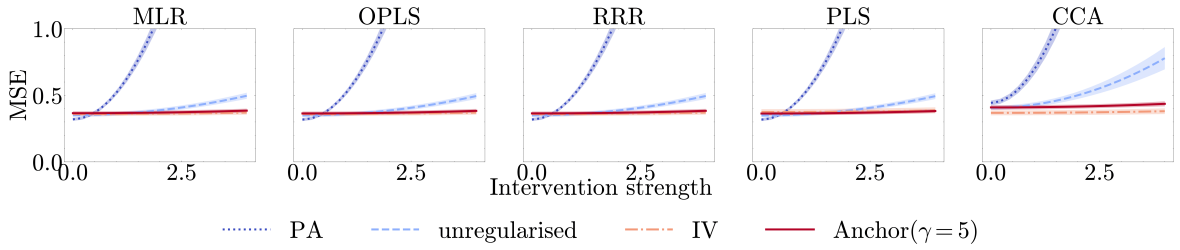

Figure 8: Experiment for anchor confounding X and Y. We can see that all algorithms present a very robust behavior to increasing perturbation strength (even CCA which is not *anchor-compatible*). Shaded areas represent the range of two standard errors of the mean for running $B = 20$ times the experiment.

D.2 Hyperparameters Selection Experiment

The results of the toy model experiments in the high-dimensional setting are obtained as follows. The training data are sampled following the DAG

$$\begin{aligned} \varepsilon_A, \varepsilon_H, \varepsilon_X, \varepsilon_Y &\sim \mathcal{N}(0,1), \\ A &\leftarrow \varepsilon_A, \\ H &\leftarrow \varepsilon_H, \\ X &\leftarrow \frac{1}{2}A + H + \varepsilon_X, \\ Y &\leftarrow 2A + \mathbf{W}^T X + H + \varepsilon_Y, \end{aligned} \tag{26}$$

and the testing data are generated by modifying ε_A such that $\varepsilon_A \sim \mathcal{N}(0, t)$. The perturbation strength t is set to 2. We repeat each experiment $B = 10$ times by resampling $n = 100$ training and validation samples and $n = 400$ testing samples of (A, X, Y). Both A and H are one-dimensional, while X and Y are 300-dimensional. The matrix $\mathbf{W}$ is generated as a low-rank (of rank ρ which is randomly sampled uniformly on $[10, 30]$) matrix such that $\mathbf{W} = \mathbf{UV}$ with $\mathbf{U} \in \mathbb{R}^{d\times\rho}$ and $\mathbf{V} \in \mathbb{R}^{\rho\times p}$. The coefficients of $\mathbf{U}$ and $\mathbf{V}$ are sampled uniformly between 1 and 3 and are normalised such that their sum is 1. Hyperparameters α ranging from 1 to 10^5 are chosen from a logarithmic scale with 20 candidates, while ρ ranges from 10 to 30 linearly with 10 candidates, and γ ranges from 1 to 10^4 logarithmically with 10 candidates. Performance, measured in terms of Mean Squared Error and Mean correlation between anchor variable and residuals, is evaluated on a validation set (not seen during training) and a perturbed testing set. We illustrate the results by selecting the optimal rank for each pair (γ, α) (see Figure D.1). Optimal parameters are chosen to minimize different objectives: for RRRR, parameters (α, ρ) that minimize training MSE are selected, while for A-RRRR, a convex combination of correlation between anchor and residuals $(g_1(\gamma, \alpha, \rho))$ and MSE $(g_2(\gamma, \alpha, \rho))$ at training is minimized, i.e., $(\gamma, \alpha, \rho) = \text{argmin} \sum_{i=1}^{2} w_i \frac{g_i(\gamma,\alpha,\rho)}{\eta_i}$, where η_1 and η_2 rescale the two objectives. In Figure D.1, we present the results for optimal RRRR and A-RRRR with weights $w_1 = 0.1$ and $w_2 = 0.9$, emphasizing independence of anchor and residuals.

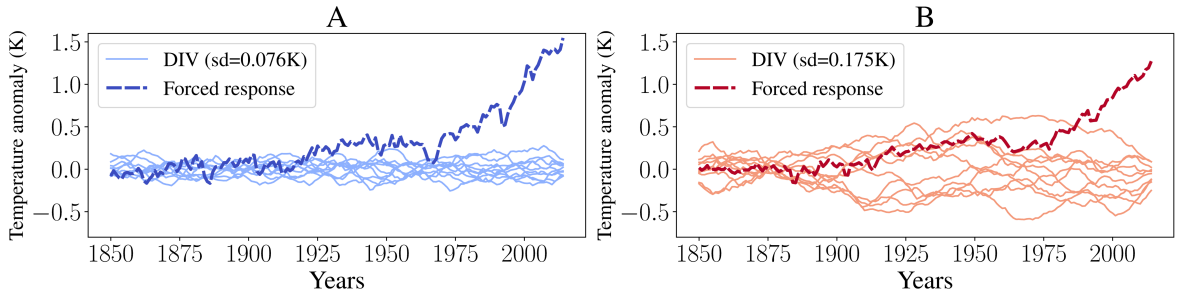

Figure 9: DIV and global forced response for a "low-variability" model (CESM2) (A) and for a "high-variability" model (CNRM-CM6-1) (B).

D.3 Anchor-MVA vs Univariate AR

As mentioned in the section, MVA methods offer a viable solution for handling correlated response variables within the context of the general multivariate linear model $Y = \mu + \mathbf{W}X + \varepsilon$. When $\mathbf{W}$ is not assumed to be full rank, the columns of Y exhibit interdependencies and often demonstrate discernible correlation patterns.

To clarify this point and the extent to which *anchor-regularised* MVA is advantageous compared to simply running multiple AR, we conducted experiments using a similar setup as in 6.1. We set the intervention strength $t = 2$, dimensions $n = 200, d = p = 400$ with $\mathbf{W}$ of rank $\rho = 10$, and used an anchor regularisation parameter $\gamma = 5$. We compared single variate AR (run for each output) and *anchor-regularised* Reduced Rank Regression with an oracle for the rank. The experiment was run $B = 50$ times, and we obtained the following results in terms of MSE:

Table 3: Comparaison of multiple AR and A-RRR with $\gamma = 5$. Results are reported in term of MSE with 95% confidence intervals.

	Multiple AR ($\gamma = 5$)	A-RRR ($\gamma = 5$)
Mean Squared Error	1.91 ± 0.26	$\mathbf{1.45 \pm 0.24}$

This showcases how the use of the *anchor-regularised* algorithm can outperform AR when faced with high-dimensional output variables.

E Details of Real-World Experiment

E.1 Robust Climate Prediction

Optimal Fingerprint for Detection of Forced Warming In D&A studies, the optimal fingerprinting process involves several steps. Firstly, the response of the climate system to an external forcing is extracted using a statistical learning model to predict the forced climate response, denoted as $Y_{\text{mod}}^{\text{forced}}$. This is achieved by utilising spatial predictors from a gridded field of climate variables X_{mod}, following the regression equation:

$$Y_{\text{mod}}^{\text{forced}} = X_{\text{mod}}\beta + \beta_0 + \varepsilon. \tag{27}$$

Here, the spatial fingerprint is represented by the regression coefficient (β, β_0). Subsequently, a detection metric is obtained by projecting observations (X_{obs}) and unforced simulations (X_{cntl}) (known as control scenarios) onto the extracted fingerprint $\hat{\beta}$. Practically, as multiple members and models are available for the unforced scenario, a distribution for $X_{\text{mod}}^{\text{cntl}}\hat{\beta} + \hat{\beta}_0$ can be obtained. Then, it is tested whether $X_{\text{obs}}\hat{\beta} + \hat{\beta}_0$ lies within the same distribution. If the test is rejected, a forced response is detected.

From Global to Regional Moving from a global to regional scale presents challenges in D&A studies. One of the current challenges is transitioning from detecting a global forced response to a regional or local forced response. This is because climate variability magnitude is larger at regional scales, and even larger at a local scale, leading to a lower signal-to-noise ratio of the forced warming response. As the signal-to-noise ratio is in this case much lower, it becomes much more challenging to detect the forced signal. Additionally, different climate models and observations exhibit different patterns of internal variability at regional scales, necessitating the training of statistical learning models robust to potential distribution shifts in internal variability.

Robustness to Multidecadal Internal Variability The ability to detect forced warming in Detection and Attribution studies is highly influenced by the level of internal variability (see Parsons et al., 2020). For instance, the observed 40-year Global Mean Temperature (GMT) trend of 0.76°C for the period $1980 - 2019$ would exceed the standard deviation of natural internal variability in CMIP5 and CMIP6 archive models classified as "low-variability" by a factor of 5 or more (see Fig. 1B in Sippel et al., 2021). However, for "high-variability" models, this trend exceeds the standard deviation only by a factor of 2. This becomes evident when comparing models with "low-variability" (Figure 9 A) to those with "high-variability" (Figure 9 B), where it is less clear that internal variability alone could generate the Global Mean Temperature. This discrepancy poses a significant challenge, as in the former case, the observed warming has an extremely low probability of being generated by

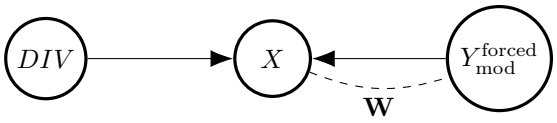

Figure 10: DAG considered for detection of forced warming response as proposed in Sippel et al. (2021).

Table 4: Climate models used in this work, their number of members, associated DIV magnitude and corresponding set (train or test).

Model	CMIP	♯members	DIV mag.	Set
CCSM4	5	8	0.052	Train
NorCMP1	6	30	0.064	Train
CESM2	6	11	0.076	Train
HadCM3	5	10	0.072	Train
CNRM-CM6-1	6	30	0.175	Test
CNRM-ESM2-1	6	10	0.174	Test
IPSL-CM6A-LR	6	32	0.141	Test

internal variability alone. Conversely, in the latter case, the rejection of the hypothesis that current temperature trends are solely attributable to internal variability is less robust. For this reason, it is important to develop tests that are robust to changes in the magnitude of internal variability, particularly with the increasing magnitude of DIV. Assuming the DAG depicted in Figure 10, anchor regularisation emerges as a suitable approach, enabling robust estimation robust to DIV increase.

Climate Models In this experiment, we utilised 7 climate models from the CMIP5 and CMIP6 archives, each having at least 8 members. The selection of these models was primarily practical, based on our access to them and their suitability for the train/test split procedure based on DIV magnitude.

Results The anchor-regularised version of RRRR generally exhibits superior performance compared to its unregularised counterpart (refer to Table 5 and Figure 12), both in terms of prediction performance (measured via R^2) and correlation between residuals and DIV at testing time. Notably, these results are achieved at a lower training time cost in terms of R^2, particularly when utilising $\gamma = 5$. This trade-off between performance during training and testing becomes clear when looking at the pareto front emerging (Figure 11) when considering different γ values for optimal hyperparameters (α, ρ) selected via the validation procedure detailed earlier. We can see that selecting γ such as it minimize an equally weighted ($\gamma 0.5/0.5$) objectives (prediction performances and correlation between residuals and DIV) lead to a very small decrease during training but induce strong robustness (being close to optimal) at testing in terms of predicting performance.

Regarding the spatial patterns of prediction performance, anchor-regularised RRRR outperforms the unregularised version, particularly in the northern hemisphere and notably at the North Pole, where internal climate variability is significantly higher than closer to the equator. It also exhibits superior performance in northern regions where unregularised RRRR performs poorly, albeit with suboptimal results. Conversely, it demonstrates slightly better performance in the southern hemisphere, especially in regions where the DIV and the residuals exhibit anticorrelation (e.g., central Africa and the southern region of South America). In terms of the correlation between DIV and residuals, anchor-regularised RRRR outperforms the unregularised version across most regions, except for central Africa and the southern regions of South America, where residuals and DIV exhibit anticorrelation. On average, it reduces the correlation between residuals and DIV by more than 15% for $\gamma = 5$ and close to 30% for $\gamma = 100$.

Table 5: Metrics for RRRR, and they anchor-regularised version with $\gamma = 5$ and $\gamma = 100$. We can see that both anchor-regularised approaches outperformed unregularised RRRR in term of R^2 score and mean correlation of residuals with DIV.

	Test		Train	
	R^2	Mean correlation	R^2	Mean correlation
RRRR	0.506	0.419	**0.510**	0.157
A-RRRR ($\gamma = 5$)	**0.537**	0.248	0.500	0.120
A-RRRR ($\gamma = 100$)	0.533	**0.123**	0.487	**0.096**

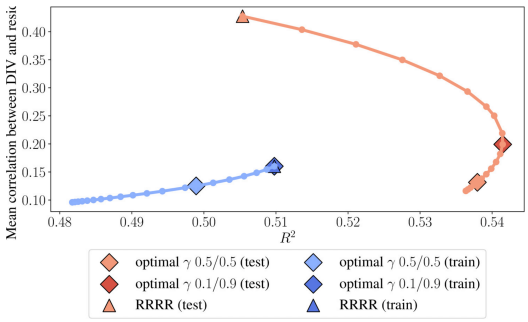

Figure 11: Pareto front for A-RRRR with optimal α, ρ, and $\gamma \in [1, 10^2]$ with optimal γ of 0.5/0.5 0.1/0.9 weighted objectives.

E.2 Robust air-quality prediction

IRM and CVP training IRM and CVP models are trained to MSE convergence with a learning rate of 0.1, patience of 200 epochs, tolerance of 10^{-4}, and a maximum of 50,000 epochs.

We provide pseudocodes for both CVP (Alg. E.2) and IRM (Alg. E.2) algorithms.

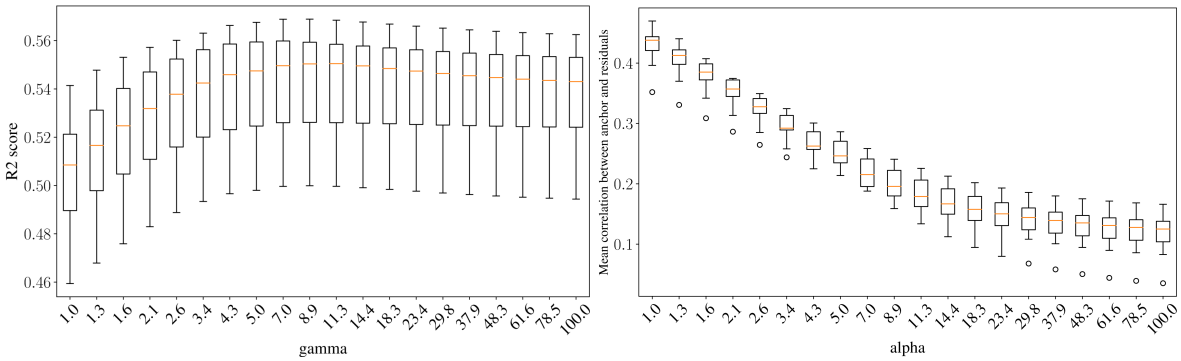

Figure 12: (Left) R^2 scores (averaged uniformly). (Right) Mean correlation between DIV and residuals for different values of γ.

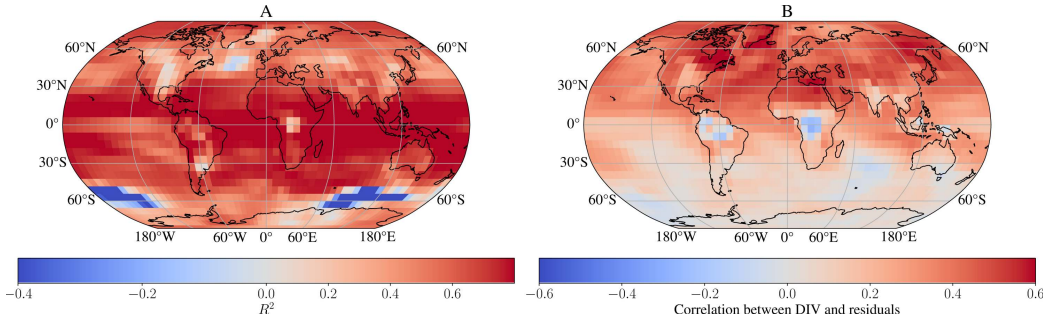

Figure 13: A-RRRR ($\gamma = 5$) scores. (A) R^2 scores. (B) Correlation between DIV and residuals.

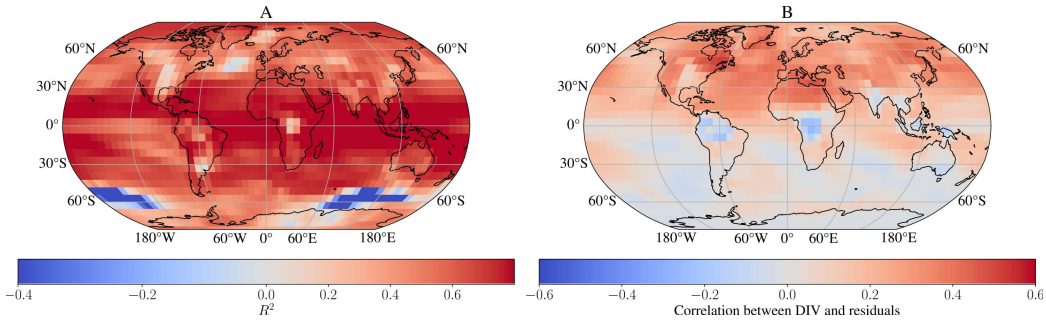

Figure 14: A-RRRR ($\gamma = 100$) scores. (A) R^2 scores. (B) Correlation between DIV and residuals.

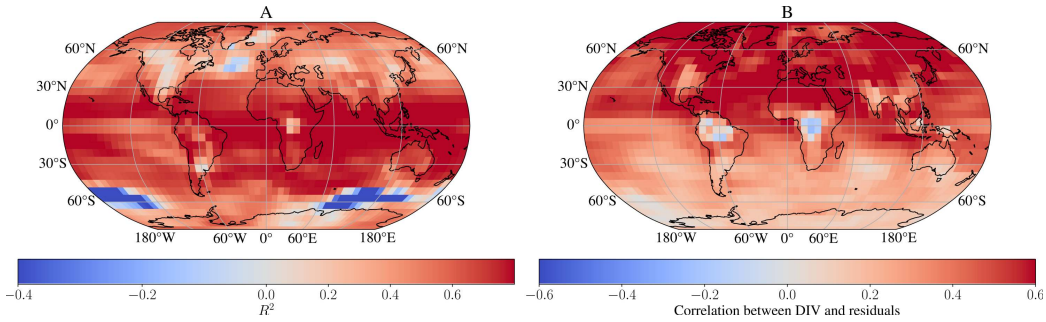

Figure 15: Unregularised RRRR scores. (A) R^2 scores. (B) Correlation between DIV and residuals.

Feature Name	Data Type	Description
CO(GT)	Integer	True hourly averaged concentration CO in mg/m^3
PT08.S1(CO)	Categorical	Hourly averaged sensor response
NMHC(GT)	Integer	True hourly averaged overall Non-Metanic HydroCarbons concentration in µg/m^3
C6H6(GT)	Continuous	True hourly averaged Benzene concentration in µg/m^3
PT08.S2(NMHC)	Categorical	Hourly averaged sensor response
NOx(GT)	Integer	True hourly averaged NOx concentration in ppb
PT08.S3(NOx)	Categorical	Hourly averaged sensor response
NO2(GT)	Integer	True hourly averaged NO2 concentration in µg/m^3

Table 6: Air Quality Dataset Outcome Description

Algorithm 1 Invariant Risk Minimization Linear Regression

Initialize: Parameters ϕ, w, learning rate η
Input: Training data X, Y, environment labels E, regularization λ_{irm}, epochs N, patience p
Set best loss to ∞, patience counter to 0
for *epoch* $= 1$ to N **do**
    Compute IRM loss:
        $\mathcal{L} = 0, P = 0$
    for each environment e in E **do**
        Extract X_e, Y_e where $E = e$
        Transform input: $X'_e = X_e\phi$
        Compute predictions: $\hat{Y}_e = X'_e w$
        Compute MSE loss: $\ell_e = \text{MSE}(\hat{Y}_e, Y_e)$
        Compute gradient penalty: $P+ = \left(\frac{d\ell_e}{dw}\right)^2$
        Update total loss: $\mathcal{L}+ = \ell_e$
    end for
    Compute final loss: $\mathcal{L} = \mathcal{L} + \lambda_{irm} P$
    Update ϕ and w using gradient descent
    Compute validation MSE for early stopping
    if validation MSE improves **then**
        Store best ϕ, w
        Reset patience counter
    else
        Increment patience counter
    end if
    if patience counter $\geq p$ **then**
        Stop training
    end if
end for
Return best ϕ and w

Algorithm 2 Conditional Variance Penalty Linear Regression

Initialize: Model parameters w, learning rate η
Input: Training data X, Y, environment labels E, regularization λ_{cvp}, epochs N, patience p
Set best loss to ∞, patience counter to 0
for *epoch* $= 1$ to N **do**
    Compute CVP loss:
        $\mathcal{L} = \text{MSE}(Xw, Y)$
    for each environment e in E **do**
        Extract X_e, Y_e where $E = e$
        Compute predictions: $\hat{Y}_e = X_e w$
        Compute variance of predictions: $V_e = \text{Var}(\hat{Y}_e)$
        Update total loss: $\mathcal{L} += \lambda_{cvp} \sum V_e$
    end for
    Update w using gradient descent
    Compute validation MSE for early stopping
    if validation MSE improves **then**
        Store best w
        Reset patience counter
    else
        Increment patience counter
    end if
    if patience counter $\geq p$ **then**
        Stop training
    end if
end for
Return best w