Causal Importance for Physics-Informed Machine Learning

Daniel F.T. Hagan DANIEL.HAGAN@UGENT.BE
Thomas Mortier THOMASF.MORTIER@UGENT.BE
Cas Decancq CAS.DECANCQ@UGENT.BE
Diego G. Miralles DIEGO.MIRALLES@UGENT.BE
Hydro-Climate Extremes Lab, Ghent University, Belgium

Editors: Bijan Mazaheri and Niels Richard Hansen

Abstract

Predictive modelling in complex dynamical systems often relies on machine learning (ML) models trained on correlated and partially redundant predictors. Standard feature importance measures are usually correlation-based and model-specific, providing limited guidance for disentangling mediators, confounders, and true drivers, and generally offering no principled route for causally motivated prediction. Here we bridge these by explicitly importing causal information, derived from the multivariate Liang–Kleeman information flow (LKIF) framework, which treats causality as a physical notion, into ML models. Here, we decompose the differential information flow into four conditioner-importance indices (Mediator Dominance Index, Moderation Gain, Confounding Pressure, and Causal Sufficiency Rate), then we construct a Causal Importance Score (CIS) that summarises the relevance of each conditioner to a given coupling. Finally, we use this CIS as a prior for two complementary ML strategies: (i) a baseline Random Forest (RF), and (ii) a neural network (NN) whose input-layer attention weights are regularised towards CIS-based priors. Using a real-world testbed with four interacting eco-hydrological variables and a target, we show that CIS-regularised NNs can closely align their learned feature usage with the physically motivated causal ranking, while retaining competitive predictive skill. This provides a concrete example of causally informed prediction, where causal diagnostics do not merely interpret an already-trained black box, but actively shape the hypothesis space explored by the model and offer a principled handle on feature selection and dimensionality reduction.

Keywords: information flow causality; causally-informed ML; feature selection; land–atmosphere interactions

1. Introduction

1.1. Towards causal conditioning

Many high-dimensional dynamical systems are governed by a web of interacting variables in which a small set of core drivers co-exist with numerous mediators and confounders (Liang, 2014; Zhou et al., 2024). In such settings, predictors may display strong co-variability and regime dependence, where the apparent effect of a given driver on a target can change substantially once additional variables are included, or when attention is restricted to specific subsets of the state space (Ebert-Uphoff and Deng, 2012; Runge et al., 2019).

Traditional correlation-based tools capture co-variability but cannot distinguish cause from effect, nor can they cleanly separate direct influences from those that operate via mediators or shared confounders (Ebert-Uphoff and Deng, 2012; Wang and Culotta, 2020). To capture directionality

and feedbacks, causal inference methods such as Granger causality (Granger, 1969), constraint-based approaches, and information-theoretic measures (Liang, 2014), have been increasingly applied across the physical and environmental sciences (Runge et al., 2019). Ecosystem interactions among variables such as soil moisture (SM), atmospheric moisture demand (or vapour pressure deficit, VPD), temperature (T), surface net shortwave radiation (SSR), and vegetation regulate key Earth system processes, including water availability, carbon uptake, droughts, and heat extremes, with direct societal and environmental consequences (Chen et al., 2021). These interactions are inherently complex because variables can switch roles between drivers, mediators, and confounders depending on system state, time scale, and background conditions (Zhou et al., 2024). Strong covariability (driven by shared external forcing) and nonlinear interactions further obscure causal relationships, making it difficult to distinguish correlation from causality. As a result, predictive models that ignore causal structure may perform well statistically yet remain ungeneralizable, poorly interpretable, and unreliable when applied outside the conditions under which they were trained (Mooij et al., 2016).

Information-flow (IF) formalisms and, in particular, the Liang–Kleeman information flow (LKIF), summarized in Appendix A, offer a physically grounded route to quantify directed interactions (Liang, 2016). It is derived from first principles, linking information transfer to the governing dynamical equations. Recently, it has been extended to multivariate dimensions which allows mediators and confounders—collectively referred to here as conditioners—to be included explicitly as conditioning variables, enabling more realistic representations of causal networks in the presence of overlapping drivers and feedbacks (Liang et al., 2021b). Further recent developments have adapted LKIF to non-stationary, multivariate networks, providing time-varying causal estimates between interacting variables under changing conditions (Zhou et al., 2024). Maiden applications of this framework to ecosystem interactions have provided time-varying causal estimates between climate variables (Zhou et al., 2024; Siddique et al., 2026). The studies revealed that the discrepancy between bivariate and fully conditioned IF estimates—summarised by a differential information flow (ΔIF)—depends strongly on environmental regimes (e.g., SM–VPD quartiles) and the role of additional variables such as T and SSR. In particular, the importance of a given conditioner is not static: it varies across space, time, and hydroclimatic states (Appendix C). This led to a set of four conditioner-importance indices that quantify how mediators and confounders restructure an inferred coupling (Appendix B). Together, these indices provide a compact summary of where and when additional variables are essential to obtain physically meaningful causal estimates (Siddique et al., 2026).

1.2. Machine learning, feature importance, and dimensionality reduction

In recent years, ML methods have become standard tools for prediction in complex systems such as ecosystems (Reichstein et al., 2019). Tree-based methods and neural networks are widely used to approximate complex mappings from multi-dimensional inputs to targets such as vegetation dynamics, taking advantage of large spatio-temporal datasets(Papagiannopoulou et al., 2017). These models typically treat inputs as flexible predictors, optimised to minimise a chosen loss function, without explicit knowledge of causal structure or of the role of individual conditioners (Ross et al., 2017). Feature importance measures (e.g., RF Gini importances, permutation importance, SHAP values) are then used post hoc to infer which predictors the model relied on (Aas et al., 2021). While informative, these measures are generally (1) inherently tied to the hypothesis class and learning al-

gorithm; (2) agnostic to the distinction between direct causes, mediators, and confounders; and (3) largely blind to the causal structure, underlying the data that matters for principled dimensionality reduction (Toms et al., 2020; Falasca et al., 2024; Schölkopf et al., 2021).

Dimensionality reduction in such contexts is often performed via generic tools; e.g., principal component analysis, regularisation, automatic relevance determination, or simple thresholding of feature importances (Zhang et al., 2023). These procedures can successfully reduce variance and computational complexity; however, they rarely answer the causal question: Which variables are necessary to prevent systematic misinterpretation of the system's response? A predictor that appears redundant in a purely predictive sense may be essential to avoid confounding or to maintain the correct causal ordering (Falasca et al., 2024).

Causal diagnostics such as information flow provide system-level estimates where they quantify how changes in one variable propagate, conditional on others (Tyrovolas et al., 2023). If these diagnostics could be translated into constraints for ML models (priors), they would offer a route to (1) causally informed prediction (Ross et al., 2017) and (2) causally motivated feature selection and dimensionality reduction (Falasca et al., 2024). Where the latter means that instead of discarding variables solely because they appear dispensable to an unconstrained model, we could favour compact predictor sets that remain faithful to the system's causal structure.

1.3. From conditioner importance to a Causal Importance Score

Building on differential information flow, an earlier work introduced four indices that summarise the importance of mediators and confounders for a given driver–response coupling (such as how SM controls T variability) (Siddique et al., 2026):

- the Mediator Dominance Index (MDI), which measures how strongly each conditioner contributes to the magnitude of IF changes;
- the Moderation Gain (MG), quantifying how much a conditioner enhances or suppresses the magnitude of the coupling;
- the Confounding Pressure (CP), capturing how strongly a conditioner corrects inflated bivariate causality estimates; and
- the Causal Sufficiency Rate (CSR), which measures how often the bivariate coupling is sufficiently close to the fully conditioned estimate.

We derived these indices from the temporal evolution of ΔIF as conditioners are added sequentially, capturing not only whether mediators/confounders matter, but in what way they alter the inferred causal structure. Large MDI and MG values indicate strong mediator effects, large CP reveals substantial confounding, and high CSR suggests that a low-dimensional, bivariate representation is often sufficient to extrapolate causality.

While each index is informative, using them jointly in ML pipelines calls for a more compact summary that can play the role of a prior over variables. In this study, we propose a Causal Importance Score (CIS) for each conditioner, defined as a weighted combination of MDI, MG, CP, and $1 - \mathrm{CSR}$, capturing in a single scalar the degree to which that conditioner is both necessary and effective in restructuring the coupling. Normalised CIS values can then be interpreted as a causal importance distribution over conditioners. This distribution not only guides the interpretation of

causal diagnostics, but also naturally suggests a ranking for causality-aware prediction and feature selection, and a criterion for dimension reduction. Thus, low-CIS variables are prime candidates for exclusion when constructing compact, yet causally faithful ML models (Schölkopf et al., 2021).

1.4. Aims of this study

In this study, we aim to integrate conditioner-aware causal diagnostics into ML models, and to clarify their implications for feature selection and dimensionality reduction. We pursue three specific objectives:

1. To construct a physically grounded Causal Importance Score (CIS) from the four conditioner-importance indices (MDI, MG, CP, CSR), and interpret it as a prior over the relative importance of conditioners for a given coupling. Here, we provide an application to explore the importance of SM, VPD, T and SSR for changes in the ecosystem.

2. To empirically investigate the potential of CIS-regularised neural networks—that differ in how strongly they encode this causal prior, and to examine how the learned input weights (denoted $\boldsymbol{\alpha}$) evolve as a function of the regularisation strength λ.

3. To explore the implications of causally informed prediction for feature selection and dimensionality reduction, showing how CIS can be used both to diagnose when variables are essential conditioners and to design compact predictor sets that remain consistent with causal diagnostics.

In the following, we firstly describe the causality formalism leading to the four indices and CIS, then detail the RF and NN architectures and their CIS-based regularisation. Finally, we discuss the broader relevance of this approach for building causally aware ML models in complex systems such as the climate system.

2. Methods

2.0.1. DIFFERENTIAL INFORMATION FLOW

To avoid notational ambiguity, we restate the time-dependent conditional information flow (IF) in full generality. For a d-dimensional system with state vector $\mathbf{X}(t) = (X_1(t), \ldots, X_d(t))^\top$, the IF from a source variable X_j to a target variable X_i, conditioned on all remaining variables $\mathcal{C} = \{X_1, \ldots, X_d\} \setminus \{X_i, X_j\}$, is given by

$$\mathcal{T}^{(t)}_{j\to i\,|\,\mathcal{C}} \;=\; \frac{1}{\det\big(\mathbf{P}_t\big)} \left(\sum_{k=1}^{d} \dot{\Delta}_{jk}(t)\, P_{k,d+i}(t) \right) \frac{P_{ij}(t)}{P_{ii}(t)}. \tag{1}$$

Here, $\mathbf{P}_t = [P_{mn}(t)] \in \mathbb{R}^{d\times d}$ denotes the time-dependent covariance matrix of $\mathbf{X}(t)$, with entries $P_{mn}(t) = \mathrm{cov}\big(X_m, X_n\big)\big|_t$, and $\det(\mathbf{P}_t)$ is its determinant. The quantity $\dot{\Delta}_{jk}(t)$ is the (j,k) cofactor of $\mathbf{P}_t$, i.e., $\dot{\Delta}_{jk}(t) = (-1)^{j+k} \det\big(\mathbf{P}^{(jk)}_t\big)$ where $\mathbf{P}^{(jk)}_t$ is the minor obtained by removing row j and column k from $\mathbf{P}_t$. The term $P_{k,d+i}(t)$ denotes the time-dependent cross-covariance between X_k and the estimated tendency of the target, $\dot{X}_i$, namely $P_{k,d+i}(t) = \mathrm{cov}\big(X_k, \dot{X}_i\big)\big|_t$. In discrete time, $\dot{x}_i(n) \approx \big(x_i(n+\kappa) - x_i(n)\big)/(\kappa\Delta t)$ for sampling interval Δt and integer step $\kappa \geq 1$.

The ratio $P_{ij}(t)/P_{ii}(t)$ rescales the flow by the target variance, consistent with the LKIF local-linearization framework (Hagan et al., 2019; Zhou et al., 2024). At each time step, the covariances $\mathbf{P}_t$ and $P_{k,d+i}(t)$ are estimated sequentially using a (square-root) Kalman filter, yielding a temporally resolved measure of directed, conditional causality in nats per unit time (Hagan et al., 2019; Zhou et al., 2024). A brief introduction to the LKIF formulation is provided in Appendix A, while further details can also be found in Zhou et al. (2024) and Liang et al. (2021b).

Let $\mathcal{T}_{j\to i}(t)$ denote the bivariate information flow from driver j (SM) to response i (LAI), and let $\mathcal{T}_{j\to i|\mathcal{C}}(t)$ be the corresponding multivariate information flow conditioned on a set of mediators/confounders $\mathcal{C}$ (e.g., VPD, T, SSR).

The differential information flow is defined as

$$\Delta IF_{j\to i|\mathcal{C}}(t) \;=\; \mathcal{T}_{j\to i|\mathcal{C}}(t) \;-\; \mathcal{T}_{j\to i}(t), \tag{2}$$

which can be considered either as a time series or in terms of its magnitude $\left|\Delta IF_{j\to i|\mathcal{C}}(t)\right|$. Small values indicate that conditioning on $\mathcal{C}$ does not substantially alter the coupling, whereas large values reveal strong mediator/confounder effects. When useful, we retain the signed $\Delta IF_{j\to i|\mathcal{C}}(t)$ to distinguish amplification ($\Delta IF > 0$) from suppression ($\Delta IF < 0$) of the bivariate coupling after conditioning.

2.0.2. SEQUENTIAL DECOMPOSITION AND FOUR INDICES

Consider a driver–response pair (j, i) and an ordered sequence of conditioners

$$\mathcal{C}^{(1)} \subset \mathcal{C}^{(2)} \subset \cdots \subset \mathcal{C}^{(M)},$$

where the final set includes all considered conditioners (e.g., $\mathcal{C}^{(M)} = \{\text{VPD}, T, \text{SSR}\}$ for $j = \text{SM}$, $i = \text{LAI}$). At each step m, the incremental contribution of adding conditioner c_m is

$$\Delta IF^{(m)}_{j\to i}(t) \;=\; \mathcal{T}_{j\to i|\mathcal{C}^{(m)}}(t) \;-\; \mathcal{T}_{j\to i|\mathcal{C}^{(m-1)}}(t), \tag{3}$$

with $\mathcal{C}^{(0)} = \emptyset$ and $\mathcal{T}_{j\to i|\mathcal{C}^{(0)}} = \mathcal{T}_{j\to i}$.

From these incremental contributions, the four indices are defined. See Appendix A for more details.

2.1. Causal Importance Score (CIS)

The four indices provide complementary views of conditioner relevance. To obtain a single measure per conditioner that can be used as a prior in ML models, we define a Causal Importance Score (CIS) as

$$\text{CIS}_c \;=\; w_{\text{MG}}\,\text{MG}_c \;+\; w_{\text{CP}}\,\text{CP}_c \;+\; w_{\text{MDI}}\,\text{MDI}_c \;+\; w_{\text{S}}\,\big(1-\text{CSR}\big), \tag{4}$$

where $w_{\text{MG}}, w_{\text{CP}}, w_{\text{MDI}}, w_{\text{S}}$ are non-negative weights. In the simplest case, we set all weights to unity, giving each aspect equal influence, but other choices are possible (e.g., emphasising confounding correction by increasing w_{CP}).

For a given driver–response pair (here SM→LAI), we compute CIS_c for each conditioner $c \in \{\text{VPD}, T, \text{SSR}\}$, then normalise:

$$p_c \;=\; \frac{\text{CIS}_c}{\sum_{c'} \text{CIS}_{c'}}. \tag{5}$$

The vector $\mathbf{p} = (p_{\text{VPD}}, p_T, p_{\text{SSR}})$ is interpreted as a causal prior over conditioners for the coupling. To extend this to the full input set $\{\text{SM}, \text{VPD}, T, \text{SSR}\}$, we can incorporate physical knowledge that SM is the upstream driver (assigning a fixed prior p_{SM}) and renormalise the full four-dimensional prior

$$\mathbf{p}_{\text{all}} = (p_{\text{SM}}, p_{\text{VPD}}, p_T, p_{\text{SSR}}), \tag{6}$$

where p_{SM} is chosen a priori (e.g., 0.4) and the remaining mass is split among VPD, T, and SSR according to their CIS.

2.2. Baseline Random Forest model

2.2.1. MODEL FORMULATION

The baseline ML model selected here is a Random Forest (RF) where LAI is the target and SM, VPD, T, and SSR as predictors. For each region, we construct a feature matrix

$$\mathbf{x}(t) = \big[\text{SM}(t), \text{VPD}(t), T(t), \text{SSR}(t)\big], \qquad y(t) = \text{LAI}(t).$$

The time series is split into training and test sets using a chronological split (e.g., 75% for training, 25% for testing) to respect temporal ordering. An RF with n_{trees} estimators is trained on the training set, with standard hyperparameters (maximum depth, minimum samples per split/leaf) chosen to balance bias and variance.

2.2.2. PERFORMANCE METRICS AND FEATURE IMPORTANCE

Predictive performance is evaluated on the test set using the root mean squared error (RMSE) and the coefficient of determination R^2. The RF also provides a built-in measure of model-defined feature importance, often computed as the average decrease in impurity (Gini importance) or via permutation importance (Yuan et al., 2021). Denote the RF importance of input variable x_k as I_k^{RF} for $k \in \{\text{SM}, \text{VPD}, T, \text{SSR}\}$.

To compare RF importance with the causal prior, we normalise the importances of the conditioners:

$$\tilde{I}_c^{\text{RF}} = \frac{I_c^{\text{RF}}}{\sum_{c' \in \{\text{VPD},T,\text{SSR}\}} I_{c'}^{\text{RF}}}. \tag{7}$$

We then compare the distributions $\mathbf{p} = (p_{\text{VPD}}, p_T, p_{\text{SSR}})$ and $\tilde{\mathbf{I}}^{\text{RF}} = (\tilde{I}_{\text{VPD}}^{\text{RF}}, \tilde{I}_T^{\text{RF}}, \tilde{I}_{\text{SSR}}^{\text{RF}})$ using an alignment metric and scatter plots.

2.3. Alignment metric between causal prior and ML importance

To quantify how closely an ML model's feature usage mirrors the CIS-based causal prior, we compute a simple alignment index. Let $\mathbf{p} = (p_c)_c$ denote the normalised CIS-derived causal prior over the conditioners (Eq. 5), and let $\tilde{\mathbf{I}} = (\tilde{I}_c)_c$ denote the corresponding normalised ML importance distribution over the same conditioners (e.g., Eq. 7 for RF or Eq. 16 for NN). By construction, $\sum_c p_c = \sum_c \tilde{I}_c = 1$. We define the alignment index as

$$A \;=\; 1 - \frac{\left\|\mathbf{p} - \tilde{\mathbf{I}}\right\|_2}{\sqrt{2}}, \tag{8}$$

where $\|\cdot\|_2$ is the Euclidean norm. The denominator $\sqrt{2}$ is the maximum possible distance between two three-element probability vectors (e.g., $(1,0,0)$ versus $(0,1,0)$), so $A \in [0,1]$, with $A = 1$ indicating perfect agreement and smaller values indicating increasing mismatch.

We also compute per-conditioner mismatches

$$\Delta_c \;=\; \tilde{I}_c - p_c, \tag{9}$$

which show whether the ML model over- or under-emphasises each conditioner relative to the causal prior.

2.4. CIS-regularised neural network

2.4.1. ARCHITECTURE AND ATTENTION-LIKE INPUT WEIGHTS

To incorporate the causal prior into a flexible, differentiable model, we design a neural network mapping $\mathbf{X}(t)$ to $\hat{y}(t)$, with an attention-like input layer that learns non-negative weights $\boldsymbol{\alpha}$ (Toms et al., 2020). Concretely, let the four inputs be denoted $x_k(t)$ for $k = 1, \ldots, 4$ corresponding to (SM, VPD, T, SSR). The network begins with a learned, normalised weighting:

$$\tilde{x}_k(t) = \alpha_k \, x_k(t), \qquad \alpha_k \geq 0, \quad \sum_{k=1}^{4} \alpha_k = 1, \tag{10}$$

implemented via a softmax transformation of unconstrained parameters θ_k:

$$\alpha_k = \frac{\exp(\theta_k)}{\sum_{\ell=1}^{4} \exp(\theta_\ell)}. \tag{11}$$

Here, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)$ are unconstrained trainable logits with $\theta_k \in \mathbb{R}$. The attention/mixture weights are obtained via $\boldsymbol{\alpha} = \mathrm{softmax}(\boldsymbol{\theta})$. This parameterization guarantees $\alpha_k \geq 0$ and $\sum_{k=1}^{K} \alpha_k = 1$ (Eq. 10) by construction, enabling unconstrained gradient-based optimization over $\boldsymbol{\theta}$. The transformed inputs $\tilde{\mathbf{x}}(t)$ are then fed into one or more hidden layers with nonlinear activation functions, followed by a linear output layer producing $\hat{y}(t)$.

This architecture not only predicts LAI but also yields a set of interpretable weights $\boldsymbol{\alpha}$ that summarise the model's learned attribution of importance across SM, VPD, T, and SSR. More details of the neural network architecture used here is provided in Appendix D.

2.4.2. LOSS FUNCTION AND λ-CONTROLLED REGULARISATION

The CIS-based causal prior $\mathbf{p}_{\text{all}}$ is incorporated via a regularisation term that penalises deviations of $\boldsymbol{\alpha}$ from $\mathbf{p}_{\text{all}}$ (Beucler et al., 2021). The total loss function is

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \lambda \, \mathcal{L}_{\text{CIS}}, \tag{12}$$

where

$$\mathcal{L}_{\text{pred}} = \frac{1}{N} \sum_{t=1}^{N} \big(\hat{y}(t) - y(t)\big)^2 \tag{13}$$

is the mean squared error over the training samples, and

$$\mathcal{L}_{\text{CIS}} = \sum_{k=1}^{4} \left(\alpha_k - p_{\text{all},k}\right)^2 \tag{14}$$

is a quadratic penalty driving $\boldsymbol{\alpha}$ towards the prior.

The scalar hyperparameter $\lambda \geq 0$ governs the trade-off between predictive flexibility and causal adherence (Tyrovolas et al., 2023):

- $\lambda = 0$ removes the causal constraint, yielding a purely data-driven NN whose $\boldsymbol{\alpha}$ reflect the loss-minimising allocation of importance.
- Small to moderate λ values encourage the NN to respect the causal ordering and relative importance suggested by CIS, while still allowing deviations where data strongly disagree.
- Large λ values force $\boldsymbol{\alpha}$ to closely match $\mathbf{p}_{\text{all}}$, potentially at the cost of predictive performance if the causal prior is imperfect or if additional patterns in the data are not captured by the prior.

We train separate networks for a range of λ values (e.g., $\lambda \in \{0, 0.1, 0.5, 1, 2, ..., 7\}$), each with multiple epochs (e.g., up to 5000) of stochastic gradient descent (Adam optimisation)(Kingma and Ba, 2017). At selected epochs, we track the evolution of $\boldsymbol{\alpha}$, the training and validation losses, the RMSE/R^2 on test data, and the alignment A (Eq. **??**) between the NN-implied conditioner importance and CIS.

2.4.3. NN-BASED FEATURE IMPORTANCE AND ALIGNMENT

For each trained NN, we derive a model-defined importance distribution over conditioners, analogous to the RF case. Here, we use the squared input weights:

$$I_k^{\text{NN}} \propto \alpha_k^2, \tag{15}$$

normalised across VPD, T, and SSR. Alternatives include gradient-based sensitivity measures or permutation importance applied to the trained NN.

We then compute

$$\tilde{I}_c^{\text{NN}} = \frac{I_c^{\text{NN}}}{\sum_{c'} I_{c'}^{\text{NN}}}, \tag{16}$$

and evaluate the alignment index A between $\tilde{\mathbf{I}}^{\text{NN}}$ and the CIS-based prior $\mathbf{p}$. By comparing A and RMSE as functions of λ, we characterise the trade-offs between causal alignment and predictive accuracy.

2.5. Data and study regions

We focus on two climate regions that have been central in previous causality analyses: a monsoonal region in China (Huanan) and a portion of the Amazon Basin. For both regions, the target variable is LAI (Liang et al., 2021a), and the driving/conditioning variables are SM, VPD, T, and SSR. Monthly anomalies are used throughout, where the seasonal expectation at each month has been removed. Data typically spans 2004–2018, with time series derived from gridded satellite and

reanalysis products (e.g., GLASS LAI/GPP and ERA5-Land hydroclimatic fields). All datasets are obtained at a spatial resolution of 0.25^o. The VPD (derived), T and SSR used in this study are obtained from the data sets from the reanalysis of the European Center for Medium-Range Weather Forecasts' (ECMWF) and SM taken from ERA5-Land (Hersbach et al., 2020). Here, SM is derived from weighted average of SM, computed between 0 and 289 cm (all four layers of SM) is used here.

For each region, spatially averaged IF time series are constructed by taking the area mean over selected grid cells. This yields a set of regional anomalies:

$$\{\text{LAI}(t), \text{SM}(t), \text{VPD}(t), T(t), \text{SSR}(t)\}, \quad t = 1, \ldots, T.$$

These time series serve both as inputs to the information-flow analysis and as features/targets for the ML models.

3. Results

3.1. ΔIF and Conditioner Importance

Figure 1 shows the results of the four indices applied to the SM→LAI of the two regions, revealing both region-specific differences and similarities. The absolute MG in Figure 1a shows that Amazon exhibits substantially larger contributions from all three conditioners, notably from VPD, which is significantly larger than in China. This highlights that conditioning on atmospheric demand and radiation in the Amazon more strongly impacts absolute changes in the SM→LAI information flow. Additionally it addresses the tug-of-war between water supply(SM) and demand(VPD) to influence vegetation dynamics. The MG values in China, on the other hand, imply modest reshaping of the coupling under the influence of the conditioners, although they are still essential. Assessing relative importance demonstrates that the coupling in both regions are affected by VPD. Unlike the MGI, secondary and tertiary conditioners are region specific differences hinged on balances between thermodynamic and radiative controls, with VPD remaining the primary factor which modulates the coupling in both regions. Nonetheless, both regions show that the residual difference in ΔIF for this coupling is generally negligible.

3.2. Model Comparisons

In order to optimise the CIS regularisation strength λ and identify an appropriate trade-off between predictive accuracy and causal interpretability, we systematically analysed the regularisation path of the CIS-guided neural network. Our results show that increasing λ progressively strengthens the influence of the CIS prior, leading to a monotonic reduction in the prior-matching diagnostic $\|\boldsymbol{\alpha} - \mathbf{p}\text{CIS}\|_2$ and clear convergence of the global feature-gating weights $\boldsymbol{\alpha}$ (Figure 2). Predictive performance (RMSE) remains stable and slightly improves for intermediate values of λ, indicating that causal regularisation does not incur a substantial performance penalty. For larger λ, the regularisation term $\lambda L\text{reg}$ increasingly dominates the loss function, enforcing strong alignment between the learned feature-gating weights and the CIS prior. Based on this trade-off, $\lambda = 4$ emerges as a suitable operating point, providing near-best prior alignment while maintaining near-optimal predictive skill. While Figure 2 shows the results for the China domain only, we note that the Amazon domain shows a similar outcome.

For each model, we compute a normalised "ML importance" distribution over the conditioners (SM, VPD, T, SSR) and compare it with the normalised CIS distribution (Figure 3). In line with

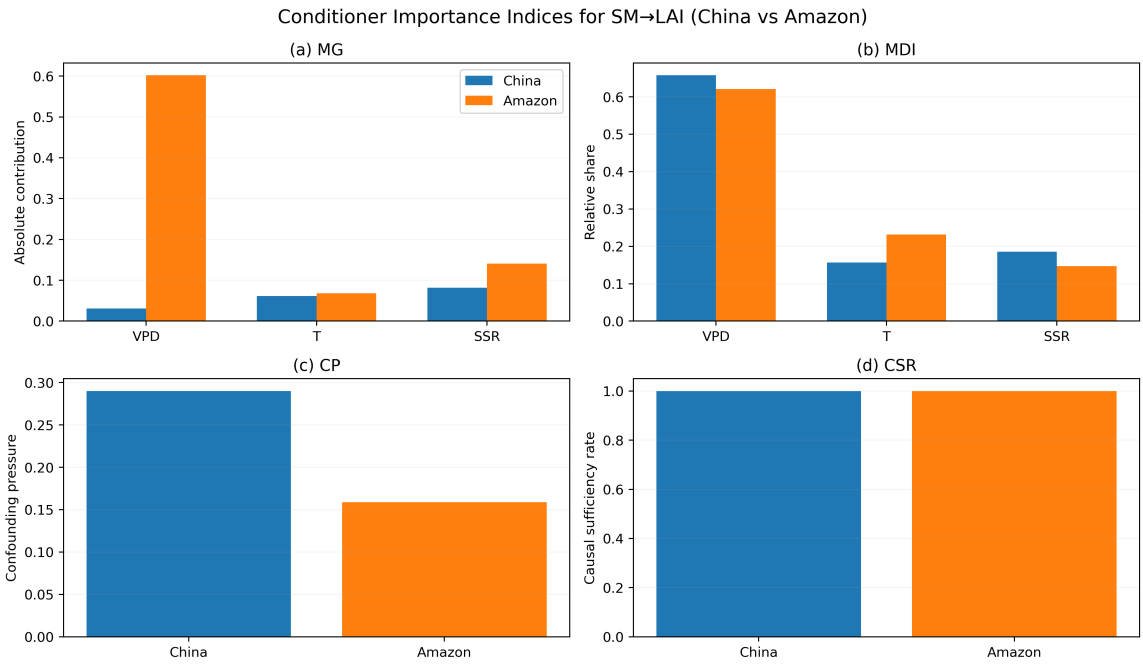

Figure 1: Conditioner importance indices for the SM→LAI coupling in China and the Amazon, showing (a) mediator gain (MG), (b) mediator dominance index (MDI), (c) confounding pressure (CP), and (d) causal sufficiency rate (CSR) computed using VPD, T, and SSR as conditioners.

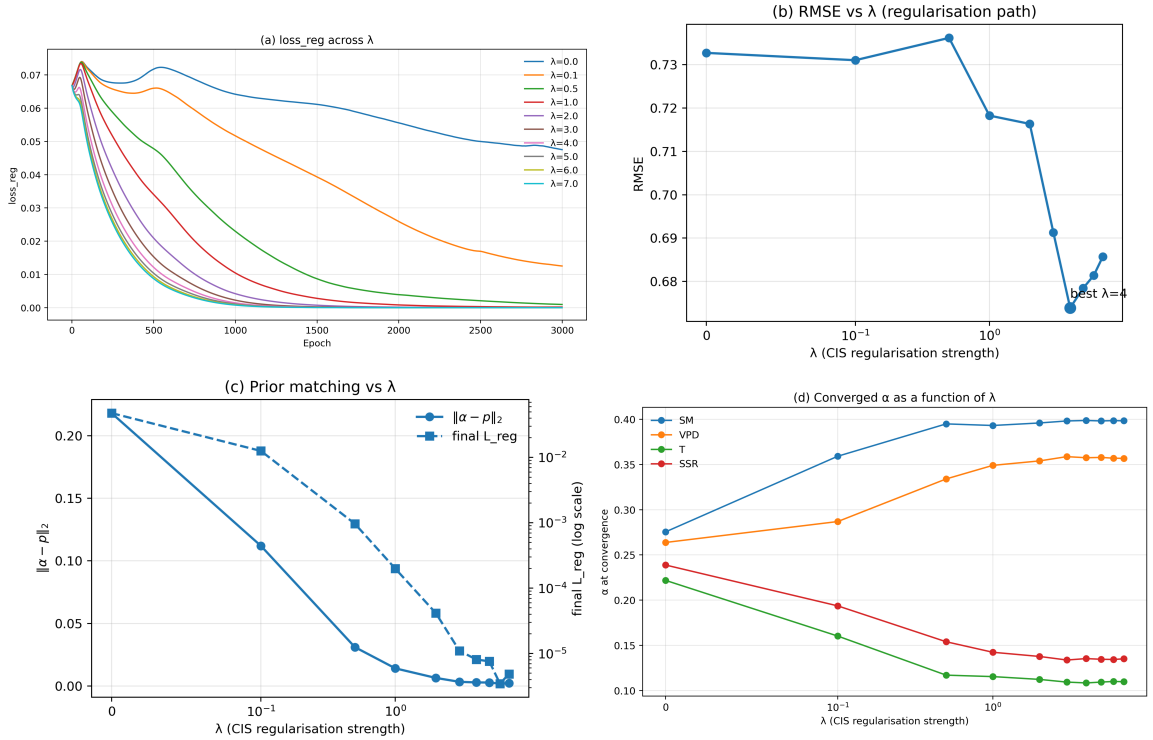

Figure 2: Regularisation and prior alignment of the CIS-guided neural network. (a) CIS regularisation term λL_{reg} as a function of the regularisation strength λ across training. (b) RMSE of predictive performance for the baseline CIS-aware neural network as a function of λ. (c) Prior-matching diagnostic $\|\boldsymbol{\alpha} - \mathbf{p}_{\text{CIS}}\|_2$, illustrating alignment of the learned global feature-gating weights with increasing λ. (d) Convergence of the global feature-gating weights $\boldsymbol{\alpha}$ as a function of λ.

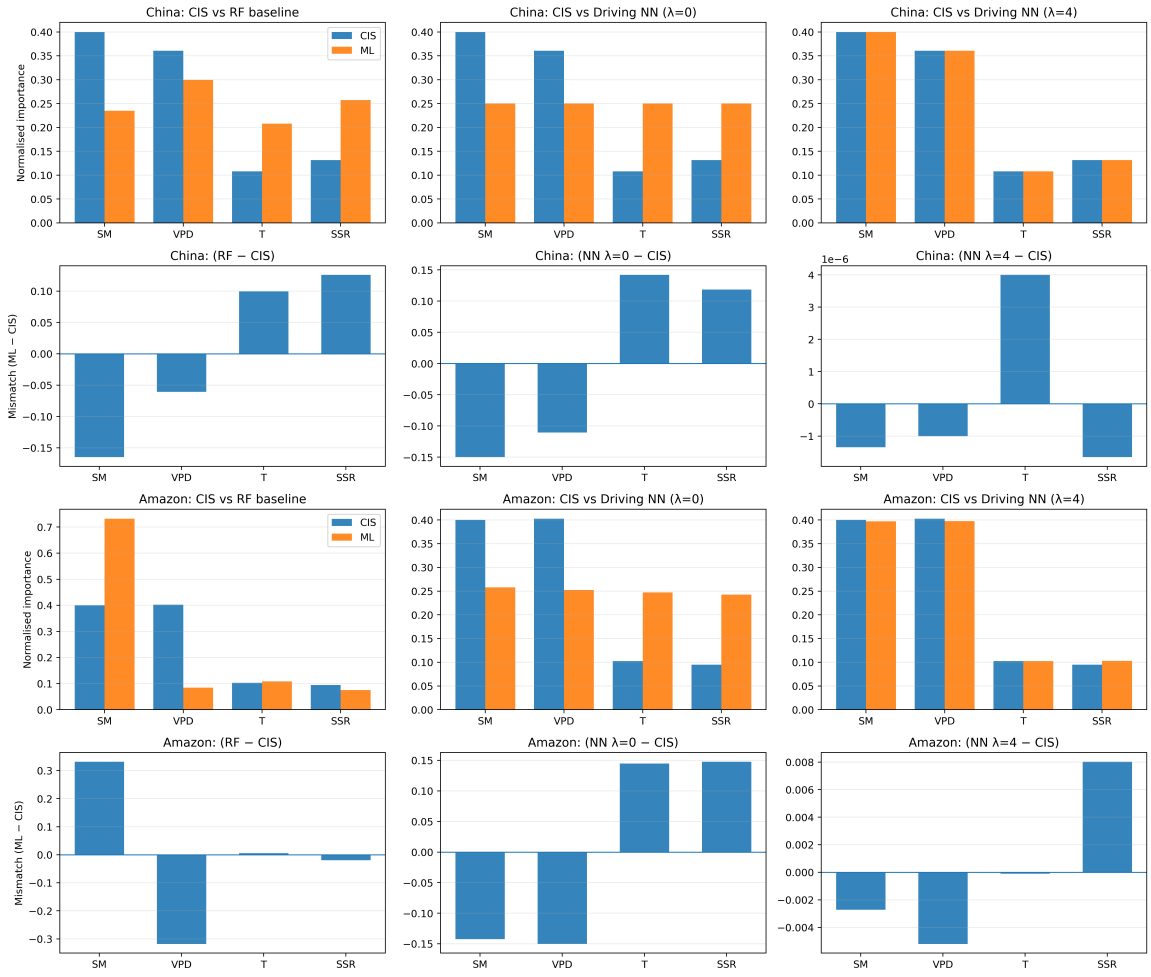

Figure 3: Comparison between causally uninformed and CIS-aware machine-learning feature attribution for SM→LAI in China and the Amazon. Top panels show results for China, while bottom panels show the corresponding results for the Amazon. In the first and third rows, normalised feature attributions from a baseline Random Forest (RF), a neural network without causal regularisation ($\lambda = 0$), and a CIS-aware neural network ($\lambda = 4$) are compared against the Causal Importance Score (CIS) derived from information-flow–based conditioner importance indices. The second and fourth rows show the corresponding differences between machine-learning attributions and CIS (ML−CIS), highlighting systematic mismatches in inferred feature importance.

the above analyses, Figure 3 illustrates significant alignments of the CIS-aware NN model at λ=4 for both regions, where mismatches are significantly reduced, implying causal knowledge in the CIS-aware NN model.

4. Conclusions

This study presents a unified framework for integrating causal diagnostics into machine-learning models of complex dynamical systems. Using the Multivariate Liang–Kleeman information-flow formulation, we derived a Causal Importance Score (CIS) that quantifies how individual variables mediate, moderate, or confound a given coupling under full multivariate conditioning. Thus, CIS provides a physically interpretable prior that extends beyond correlation-based relevance by explicitly encoding causal necessity and near-sufficiency. We then applied this framework to land–atmosphere interactions, where CIS revealed regional differences in how atmospheric demand, temperature, and radiation condition the SM→LAI coupling across China and the Amazon (Figures 1,4). These contrasts expose the limitations of bivariate and correlation-based metrics, which conflate direct influence with conditioning effects and obscure the causal roles of additional drivers.

From here, we demonstrated that CIS can be operationalised within machine learning. While Random Forest feature importances only partially reproduced the CIS structure, a neural network which was equipped with CIS-based regularisation learned global feature-gating weights that progressively aligned with the causal prior as the regularisation strength λ increases (Beucler et al., 2021). By jointly analysing predictive performance and prior-matching diagnostics, we identified an operating regime in which causal interpretability is substantially improved with minimal loss in predictive skill, illustrating a clear and controllable trade-off between model flexibility and causal adherence (Ross et al., 2017). Beyond prediction, the framework offers a causally grounded approach to feature attribution and selection, such that variables with low CIS contribute little to the fully conditioned coupling, while high-CIS variables remain essential even when their marginal predictive contribution is modest. Although the implementation in this study employs a global gating mechanism and quadratic regularisation, the framework is readily extensible to regime-dependent priors and more expressive architectures, especially since the CIS can also be localized in time and space.

Several assumptions were made to simplify the design of the study in order to track the transfer of causal knowledge from the IF estimates to the ML implementation. Firstly, the goal of the study was not to predict LAI from all possible permutations of predictors. Here, we asked a targeted question such as: How does SM control on LAI, influenced by potential confounders/mediators, inform us about how to predict LAI? For this reason, we focused on the SM→LAI coupling. The IF budget here only focused on information from other sources to LAI, which also excluded the self-IF of LAI. Additionally, this specific coupling was chosen because we had empirically determined that ΔIF is consistent over time, such that making a stationary assumption about the conditioners would deteriorate the causal structure (Appendix B). Nevertheless, these assumptions can be relaxed and further explored in a more detailed future study without coming to largely different conclusions than we have derived here. Overall, this work demonstrates a basis for causally informed machine learning, in which information-theoretic causality analysis directly shapes model training to yield predictions that are not only accurate but also physically and causally consistent.

Acknowledgments

This research was funded by the European Research Council (ERC) under grant agreement 101088405 (HEAT), the research Foundation Flanders under the FWO-NSFC bilateral research project funding (CausalHeat, G0AOO25N).

References

Kjersti Aas, Martin Jullum, and Anders Løland. Explaining individual predictions when features are dependent: More accurate approximations to shapley values. *Artificial Intelligence*, 298:103502, 2021. ISSN 0004-3702. doi: https://doi.org/10.1016/j.artint.2021.103502. URL https://www.sciencedirect.com/science/article/pii/S0004370221000539.

Tom Beucler, Michael Pritchard, Stephan Rasp, Jordan Ott, Pierre Baldi, and Pierre Gentine. Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.*, 126:098302, Mar 2021. doi: 10.1103/PhysRevLett.126.098302. URL https://link.aps.org/doi/10.1103/PhysRevLett.126.098302.

Ning Chen, Chao Song, Xufeng Xu, Yan Liu, Xiangjun Tian, and Yiqiang Zhang. Divergent impacts of atmospheric water demand on gross primary productivity in three typical ecosystems in china. *Agric. For. Meteorol.*, 307:108527, 2021. doi: 10.1016/j.agrformet.2021.108527.

Peter Bühlmann Jonas Peters Dominik Rothenhäusler, Nicolai Meinshausen. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2021. URL https://doi.org/10.1111/rssb.12398.

Imme Ebert-Uphoff and Yi Deng. Causal discovery for climate research using graphical models. *J. Climate*, 25(17):5648–5665, 2012. doi: 10.1175/JCLI-D-11-00387.1.

Fabrizio Falasca, Pavel Perezhogin, and Laure Zanna. Data-driven dimensionality reduction and causal inference for spatiotemporal climate fields. *Phys. Rev. E*, 109:044202, Apr 2024. doi: 10.1103/PhysRevE.109.044202. URL https://link.aps.org/doi/10.1103/PhysRevE.109.044202.

Clive W. J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969. doi: 10.2307/1912791.

Daniel Fiifi Tawia Hagan, Guojie Wang, X. San Liang, and Han A. J. Dolman. A time-varying causality formalism based on the liang–kleeman information flow for analyzing directed interactions in nonstationary climate systems. *Journal of Climate*, 32(21):7521 – 7537, 2019. doi: 10.1175/JCLI-D-18-0881.1. URL https://journals.ametsoc.org/view/journals/clim/32/21/jcli-d-18-0881.1.xml.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, Adrian Simmons, Cornel Soci, Saleh Abdalla, Xavier Abellan, Gianpaolo Balsamo, Peter Bechtold, Gionata Biavati, Jean Bidlot, Massimo Bonavita, Giovanna De Chiara, Per Dahlgren, Dick Dee, Michail Diamantakis, Rossana Dragani, Johannes Flemming, Richard Forbes, Manuel Fuentes, Alan Geer, Leo Haimberger, Sean Healy, Robin J. Hogan, Elías Hólm, Marta Janisková, Sarah Keeley, Patrick Laloyaux, Philippe Lopez, Cristina Lupu, Gabor Radnoti, Patricia de Rosnay, Iryna Rozum, Freja Vamborg, Sebastien Villaume, and Jean-Noël Thépaut. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020. doi: https://doi.org/10.1002/qj.3803. URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Lucas Kania and Ernst Wit. Causal regularization: On the trade-off between in-sample risk and out-of-sample risk guarantees, 2025. URL https://arxiv.org/abs/2205.01593.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. 2017. URL https://arxiv.org/abs/1412.6980.

Zhao-Rong Lai and Weiwen Wang. Invariant risk minimization is a total variation model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 25913–25935. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/lai24c.html.

Shunlin Liang, Jie Cheng, Kun Jia, Bo Jiang, Qiang Liu, Zhiqiang Xiao, Yunjun Yao, Wenping Yuan, Xiaotong Zhang, Xiang Zhao, and Ji Zhou. The global land surface satellite (glass) product suite. *Bulletin of the American Meteorological Society*, 102(2):E323 – E337, 2021a. doi: 10.1175/BAMS-D-18-0341.1. URL https://journals.ametsoc.org/view/journals/bams/102/2/BAMS-D-18-0341.1.xml.

X. San Liang. Information flow and causality as rigorous notions ab initio. *Phys. Rev. E*, 94:052201, Nov 2016. doi: 10.1103/PhysRevE.94.052201. URL https://link.aps.org/doi/10.1103/PhysRevE.94.052201.

Xiang S. Liang, Xiaojun Yu, and Zeng-Zhen Li. Normalized multivariate time series causality analysis and causal graph reconstruction. *Entropy*, 23(5):679, 2021b. doi: 10.3390/e23050679.

Xiang San Liang. Unraveling the cause–effect relation between time series. *Phys. Rev. E*, 90(5): 052150, 2014. doi: 10.1103/PhysRevE.90.052150.

Joris M. Mooij, Dominik Janzing, Jan Peters, and et al. Distinguishing cause from effect using observational data: Methods and benchmarks. *J. Mach. Learn. Res.*, 17(32):1–102, 2016.

C. Papagiannopoulou, D. G. Miralles, S. Decubber, M. Demuzere, N. E. C. Verhoest, W. A. Dorigo, and W. Waegeman. A non-linear granger-causality framework to investigate climate–vegetation dynamics. *Geoscientific Model Development*, 10(5):1945–1960, 2017. doi: 10.5194/gmd-10-1945-2017. URL https://gmd.copernicus.org/articles/10/1945/2017/.

Markus Reichstein, Gustau Camps-Valls, Bjorn Stevens, Martin Jung, Joachim Denzler, Nuno Carvalhais, and Prabhat. Deep learning and process understanding for data-driven earth system science. *Nature*, 566:195–204, 2019. doi: 10.1038/s41586-019-0912-1.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations, 2017. URL https://arxiv.org/abs/1703.03717.

Jakob Runge, Sebastian Bathiany, Erik Bollt, Inderjit Dokuchaev, Vahid Dojani, Jonas Gilson, Marleen Kretschmer, Christopher C. Lovell, Kun Zhang, et al. Inferring causation from time series in earth system sciences. *Nature Communications*, 10:2553, 2019. doi: 10.1038/s41467-019-10105-3.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9363924. *equal contribution.

Buhlmann P. Taeb A. Shen, X. Causality-oriented robustness: Exploiting general noise interventions. *Journal of the American Statistical Association*, 2026. URL https://doi.org/10.1080/01621459.2025.2544365.

Fareeha Siddique, Daniel Fiifi Tawia Hagan, Guojie Wang, David Docquier, and Stéphane Vannitsem. Benchmarking conditioners in liang–kleeman information flow: Application to land–atmosphere interactions. *EGUsphere [preprint]*, 2026. URL https://doi.org/10.5194/egusphere-2026-267.

Benjamin A. Toms, Elizabeth A. Barnes, and Imme Ebert-Uphoff. Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems*, 12(9):e2019MS002002, 2020. doi: https://doi.org/10.1029/2019MS002002. URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002. e2019MS002002 10.1029/2019MS002002.

Marios Tyrovolas, X. San Liang, and Chrysostomos Stylios. Information flow-based fuzzy cognitive maps with enhanced interpretability. *Granular Computing*, 8(6):2364–4974, 2023. doi: 10.1007/s41066-023-00417-7. URL https://doi.org/10.1007/s41066-023-00417-7.

Isabella Verdinelli and Larry Wasserman. Decorrelated variable importance. *Journal of Machine Learning Research*, 25(7):1–27, 2024. URL http://jmlr.org/papers/v25/22-0801.html.

Zhao Wang and Aron Culotta. Robustness to spurious correlations in text classification via automatically generated counterfactuals, 2020. URL https://arxiv.org/abs/2012.10040.

Ye Yuan, Liji Wu, and Xiangmin Zhang. Gini-impurity index analysis. *IEEE Transactions on Information Forensics and Security*, 16:3154–3169, 2021. doi: 10.1109/TIFS.2021.3076932.

Hao-Tian Zhang, Wen-Yong Guo, and Wen-Ting Wang. The dimensionality reductions of environmental variables have a significant effect on the performance of species distribution models. *Ecology and Evolution*, 13(11):e10747, 2023. doi: https://doi.org/10.1002/ece3.10747. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/ece3.10747.

Feihong Zhou, Daniel Fiifi Tawia Hagan, Guojie Wang, X. San Liang, Shijie Li, Yuhao Shao, Emmanuel Yeboah, and Xikun Wei. Estimating time-dependent structures in a multivariate causality for land–atmosphere interactions. *Journal of Climate*, 37(6):1853 – 1876, 2024. doi: 10.1175/JCLI-D-23-0207.1. URL https://journals.ametsoc.org/view/journals/clim/37/6/JCLI-D-23-0207.1.xml.

Appendix A. The Liang–Kleeman Information Flow Theory

A.1. From first principles

For a multivariate dynamical system, causal influence is naturally directional: the effect of X on Y need not equal the effect of Y on X. The Liang–Kleeman Information Flow (LKIF) formalism quantifies this asymmetry by defining causality as the rate of information transfer from one component of a system to another, derived from the underlying system dynamics or first principles (Liang, 2016, 2014). In this Newtonian sense, causality is treated as a physical notion encoded in the evolution equations of the system. Accordingly, LKIF characterizes directed influence via how the dynamics propagate uncertainty (Shannon entropy), rather than relying on purely correlation-based or model-specific statistical ansätze.

A key philosophical distinction from purely statistical approaches is that LKIF starts from (i) the evolution law of the system state (deterministic or stochastic), (ii) derives the evolution of the probability density (Liouville/Fokker–Planck), and (iii) quantifies how the marginal entropy of a target component changes due to other components (Liang, 2016).

We begin with the continuous-time stochastic system

$$d\mathbf{X} = \mathbf{F}(\mathbf{X}, t)\, dt + \mathbf{B}(\mathbf{X}, t)\, d\mathbf{W}, \tag{17}$$

where $\mathbf{X} = (X_1, \ldots, X_d)^\top$ is the state vector, $\mathbf{F}$ is the (possibly nonlinear) drift, $\mathbf{B}$ is the diffusion amplitude matrix, and $\mathbf{W}$ is a vector of standard Wiener processes. Equation (17) is the starting point for LKIF: causality is treated as a dynamical notion encoded in how the system evolves in time.

A.2. Dynamic density evolution (Liouville / Fokker–Planck)

Equation (17) implies an evolution equation for the joint probability density $\rho(\mathbf{x}, t)$. For deterministic flows ($\mathbf{B} \equiv 0$), this is the Liouville equation; for SDEs it becomes the Fokker–Planck equation. This step is the bridge from "physics" (state evolution) to "information" (entropy evolution): once ρ evolves, all marginals and entropies evolve.

Given $\rho(\mathbf{x}, t)$, the marginal density of a component X_i is

$$\rho_i(x_i, t) = \int \rho(\mathbf{x}, t)\, d\mathbf{x}_{-i},$$

and its marginal Shannon entropy is

$$H_i(t) = -\int \rho_i(x_i, t) \log \rho_i(x_i, t)\, dx_i.$$

LKIF asks: how much of $\frac{dH_i}{dt}$ is attributable to another component X_j through the system dynamics?

A.3. LKIF as the entropy-transfer rate

The central LKIF idea is a decomposition of the marginal entropy tendency of a target variable X_i:

$$\frac{dH_i}{dt} = \Big(\text{entropy tendency of } X_i \text{ with } X_j \text{ excluded}\Big) + \Big(\text{information transferred from } X_j \text{ to } X_i\Big).$$

Operationally, LKIF defines the transfer ($\mathcal{T}_{j\to i}$) as the portion of the marginal-entropy rate that disappears if the source variable is (instantaneously) removed from the dynamics (Liang, 2016).

For the continuous-time setting above, the LKIF from a source X_j to a target X_i can be written in an integral form involving the drift and densities (Liang, 2016):

$$\mathcal{T}_{j\to i} = -\int_{\mathbb{R}^2} \frac{1}{\rho_i(x_i,t)} \frac{\partial(F_i\,\rho_i(x_i,t))}{\partial x_i}\,\rho_{j|i}(x_j|x_i,t)\,dx_i\,dx_j. \tag{18}$$

Here, F_i is the drift component governing X_i, ρ_i is the marginal density of X_i, and $\rho_{j|i}$ is the conditional density of X_j given X_i. (In the most general stochastic setting, an additional diffusion contribution can appear; in the formulations used here, this contribution is either negligible or does not depend on the source variable and is omitted for clarity; see Liang, 2016.)

Interpretation. $\mathcal{T}_{j\to i}$ has units of nats per unit time and can be read as the rate at which the variability of X_j changes the uncertainty (entropy) of X_i through the system dynamics. A positive value indicates that X_j increases the marginal uncertainty of X_i; a negative value indicates a stabilizing (uncertainty-reducing) influence.

A.4. Towards a practical estimation

While Eq. (18) is fully general, direct estimation is difficult in high dimensions. A widely used and effective specialization assumes a locally linear drift:

$$\mathbf{F}(\mathbf{X},t) \approx \mathbf{A}\mathbf{X}, \qquad \mathbf{A} = (a_{ij}), \tag{19}$$

under which the information flow simplifies to (**?**):

$$\mathcal{T}_{j\to i} = a_{ij}\,\frac{\sigma_{ij}}{\sigma_{ii}}, \tag{20}$$

where $\sigma_{ij} = \mathrm{cov}(X_i, X_j)$ and $\sigma_{ii} = \mathrm{var}(X_i)$ are population covariances (Liang, 2014).

To estimate a_{ij} from discrete time series $\{x_i(n)\}_{n=1}^{N}$ with step size Δt, one approximates $\dot{x}_i$ using an Euler forward difference and solves a least-squares system. Specifically, for the tendency of a generic target x_i one writes:

$$\sum_{m=1}^{d} a_{im}\,x_m(n) = \dot{x}_i(n), \qquad n = 1,\ldots,N, \tag{21}$$

with $\dot{x}_i(n) \approx \big(x_i(n+k) - x_i(n)\big)/(k\Delta t)$. This provides an empirically tractable route from data to the coupling coefficients and, through Eq. (20), to directed information flow.

A.5. Multivariate conditioning: causal flow "given" other variables

In multivariate systems, a bivariate flow $\mathcal{T}_{j\to i}$ may be distorted by additional variables that act as mediators or confounders. Liang's multivariate extension estimates the conditional flow from X_j to X_i while accounting for the remaining variables $\mathcal{C} = \{X_1,\ldots,X_d\} \setminus \{X_i, X_j\}$ (Liang et al., 2021b).

In Zhou et al. (2024), the stationary conditional flow can be written in a sample-covariance (cofactor) form:

$$\mathcal{T}_{j\to i\,|\,\mathcal{C}} = \frac{1}{\det(\mathbf{C})}\left(\sum_{m=1}^{d}\Delta_{jm}\,C_{m,d+i}\right)\frac{C_{ij}}{C_{ii}}, \tag{22}$$

where $\mathbf{C}$ is the $d \times d$ sample covariance matrix with entries $C_{ab} = \mathrm{cov}(X_a, X_b)$, $\det(\mathbf{C})$ is its determinant, and Δ_{jm} is the (j,m) cofactor of $\mathbf{C}$. The term $C_{m,d+i}$ denotes the covariance between x_m and the derived series $\dot{x}_i$ (i.e., the estimated tendency of the target). Equation (22) reduces to the familiar $(2 \to 1)$ form by setting $(i,j) = (1,2)$.

A.6. Time-varying estimation via the Kalman-filter

A key limitation of Eq. (22) is its reliance on a stationary covariance matrix computed over a window sufficiently long for statistical stability. However, land–atmosphere interactions often exhibit seasonal and intraseasonal changes in coupling structure. To track such time variation, Zhou et al. (2024) estimate a time-evolving covariance matrix (and, where required, time-varying local linear coefficients) sequentially using a (square-root) Kalman filter, and then substitute the evolving covariance into the LKIF estimator (Zhou et al., 2024). In other words, the role of the Kalman filter in this framework is to provide $\mathbf{P}(t)$ (and optionally $a_{ij}(t)$) at each time step, enabling a temporally resolved LKIF.

A.7. Kalman filter formulation for $\mathbf{P}(t)$

The Kalman filter assumes a linear stochastic state-space model (**??**):

$$\mathbf{x}_k = \mathbf{A}\mathbf{x}_{k-1} + \mathbf{B}\mathbf{u}_k + \mathbf{w}_{k-1}, \tag{23}$$

$$\mathbf{y}_k = \mathbf{H}\mathbf{x}_k + \mathbf{v}_k, \tag{24}$$

where $\mathbf{x}_k$ is the (possibly augmented) state vector, $\mathbf{y}_k$ is the measurement vector, $\mathbf{w}_k$ and $\mathbf{v}_k$ are zero-mean process and measurement noises with covariances $\mathbf{Q}_k$ and $\mathbf{R}_k$, respectively, and $\mathbb{E}[\mathbf{w}_k\mathbf{v}_i^\top] = \mathbf{0}$. In the time-varying LKIF application, the filter is used to stably update the evolving covariance structure $\mathbf{P}_k$ (and, in the time-varying formulation, the locally linear coefficients $a_{ij}(t)$; see Zhou et al. (2024) for implementation details).

The recursion alternates between prediction and correction. The predicted state and predicted error covariance are:

$$\hat{\mathbf{x}}_k^- = \mathbf{A}\hat{\mathbf{x}}_{k-1} + \mathbf{B}\mathbf{u}_k, \tag{25}$$

$$\mathbf{P}_k^- = \mathbf{A}\mathbf{P}_{k-1}\mathbf{A}^\top + \mathbf{Q}_k. \tag{26}$$

The Kalman gain then weights the innovation to update the state estimate and covariance:

$$\mathbf{K}_k = \mathbf{P}_k^-\mathbf{H}^\top\left(\mathbf{H}\mathbf{P}_k^-\mathbf{H}^\top + \mathbf{R}_k\right)^{-1}, \tag{27}$$

$$\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k\left(\mathbf{y}_k - \mathbf{H}\hat{\mathbf{x}}_k^-\right), \tag{28}$$

$$\mathbf{P}_k = \left(\mathbf{I} - \mathbf{K}_k\mathbf{H}\right)\mathbf{P}_k^-. \tag{29}$$

Zhou et al. (2024) further employ a square-root Kalman filter (Bierman–Thornton) for numerical stability, particularly important for highly correlated multivariate geophysical variables.

A.8. Time-varying multivariate LKIF via the evolving covariance

With the Kalman-filtered covariance $\mathbf{P}(t)$ available at each time step, the multivariate LKIF estimator is evaluated by replacing the stationary sample covariance $\mathbf{C}$ with $\mathbf{P}(t)$. Zhou et al. (2024) express the time-varying conditional flow in an analogous cofactor form:

$$\mathcal{T}^{(t)}_{j\to i\,|\,\mathcal{C}} = \frac{1}{\det(\mathbf{P}_t)} \left(\sum_{m=1}^{d} \dot{\Delta}_{jm}(t)\, P_{m,d+i}(t) \right) \frac{P_{ij}(t)}{P_{ii}(t)}, \tag{30}$$

where $\dot{\Delta}_{jm}(t)$ denotes the (j,m) cofactor of $\mathbf{P}_t = (P_{ab}(t))$, and $P_{m,d+i}(t)$ is the covariance between X_m and the derived series $\dot{X}_i$ estimated at time t. Equation (30) reduces to the familiar $(2 \to 1)$ form by setting $(i,j) = (1,2)$.

Appendix B. Conditioner Importance indices

Mediator Dominance Index (MDI). For each conditioner c, MDI measures how strongly c contributes to the overall restructuring:

$$\mathrm{MDI}_c \;=\; \frac{\sum_t \left|\Delta IF^{(c)}_{j\to i}(t)\right|}{\sum_{c'}\sum_t \left|\Delta IF^{(c')}_{j\to i}(t)\right|}. \tag{31}$$

Here $\Delta IF^{(c)}$ denotes the incremental IF change when c is added. MDI values sum to 1 across conditioners.

Moderation Gain (MG). MG captures how much the conditioner increases (or decreases) the magnitude of the coupling:

$$\mathrm{MG}_c \;=\; \frac{\sum_t \Big[\left|T_{j\to i|\mathcal{C}^{(c)}}(t)\right| - \left|T_{j\to i|\mathcal{C}^{(c-1)}}(t)\right|\Big]_+}{\sum_{c'}\sum_t \Big[\left|T_{j\to i|\mathcal{C}^{(c')}}(t)\right| - \left|T_{j\to i|\mathcal{C}^{(c'-1)}}(t)\right|\Big]_+}, \tag{32}$$

where $[x]_+ = \max(x, 0)$. Large MG values signal strong enhancement or restructuring by c.

Confounding Pressure (CP). CP measures the degree to which a conditioner reduces an inflated bivariate coupling:

$$\mathrm{CP}_c \;=\; \frac{\sum_t \Big[\left|T_{j\to i}(t)\right| - \left|T_{j\to i|\mathcal{C}^{(c)}}(t)\right|\Big]_+}{\sum_{c'}\sum_t \Big[\left|T_{j\to i}(t)\right| - \left|T_{j\to i|\mathcal{C}^{(c')}}(t)\right|\Big]_+}. \tag{33}$$

High CP indicates that removing c from the model would lead to significant overestimation of causal strength.

Causal Sufficiency Rate (CSR). CSR quantifies the fraction of time steps where the bivariate and fully conditioned couplings are sufficiently close:

$$\mathrm{CSR} \;=\; \frac{1}{T}\sum_{t=1}^{T}\mathbb{I}\Big(\big|T_{j\to i|\mathcal{C}^{(M)}}(t) - T_{j\to i}(t)\big| < \tau\Big), \tag{34}$$

where τ is a tolerance chosen on a physically meaningful scale of IF variation and $\mathbb{I}(\cdot)$ is the indicator function. High CSR implies a largely sufficient bivariate representation; low CSR indicates consistent divergence. For more details, readers are referred to Siddique et al. (2026).

Appendix C. Information Flow estimations of the SM→LAI couplings

Figure 4 shows the monthly IF time series for SM→LAI for the two regions (top and middle panels) between 2004 and 2019. Overall, the results show that the impact of the conditioners (VPD, T and SSR) are both time and space dependent. Furthermore, different combinations of conditioners impact the coupling differently. These differences are summarized in the ΔIF time series in the bottom panel, which also shows that for the same variables within the same time window, they are affected differently by the conditioners in the two different regions. From these results, it is clear that making a non-physical assumption about how the variables interact, such as the statistical assumptions used in constructing ML models may ignore how predictors interact with the target in reality. For instance, the ΔIF time series show (a) a stronger dependency on the conditioners in the Amazon region than in the China region; and (b) the conditioners affect the coupling differently at different times owing to physical stress at a particular time point.

Appendix D. Neural Network Architecture and Training Details

Here, we provide details of the implementation of the neural network (NN) architecture and training procedure used in Section 2.5.1 to incorporate the Causal Importance Score (CIS) into a predictive model.

D.1. Neural network architecture

The CIS-regularised neural network is designed to predict a target variable $y(t)$ (LAI or GPP) from four input variables

$$x(t) = [\,\mathrm{SM}(t), \mathrm{VPD}(t), \mathrm{T}(t), \mathrm{SSR}(t)\,].$$

The network consists of two conceptual components:

Global feature-gating layer. A learnable, normalised weight vector

$$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4), \qquad \alpha_k \geq 0, \quad \sum_{k=1}^{4}\alpha_k = 1,$$

is applied multiplicatively to the inputs:

$$\tilde{x}_k(t) = \alpha_k\, x_k(t).$$

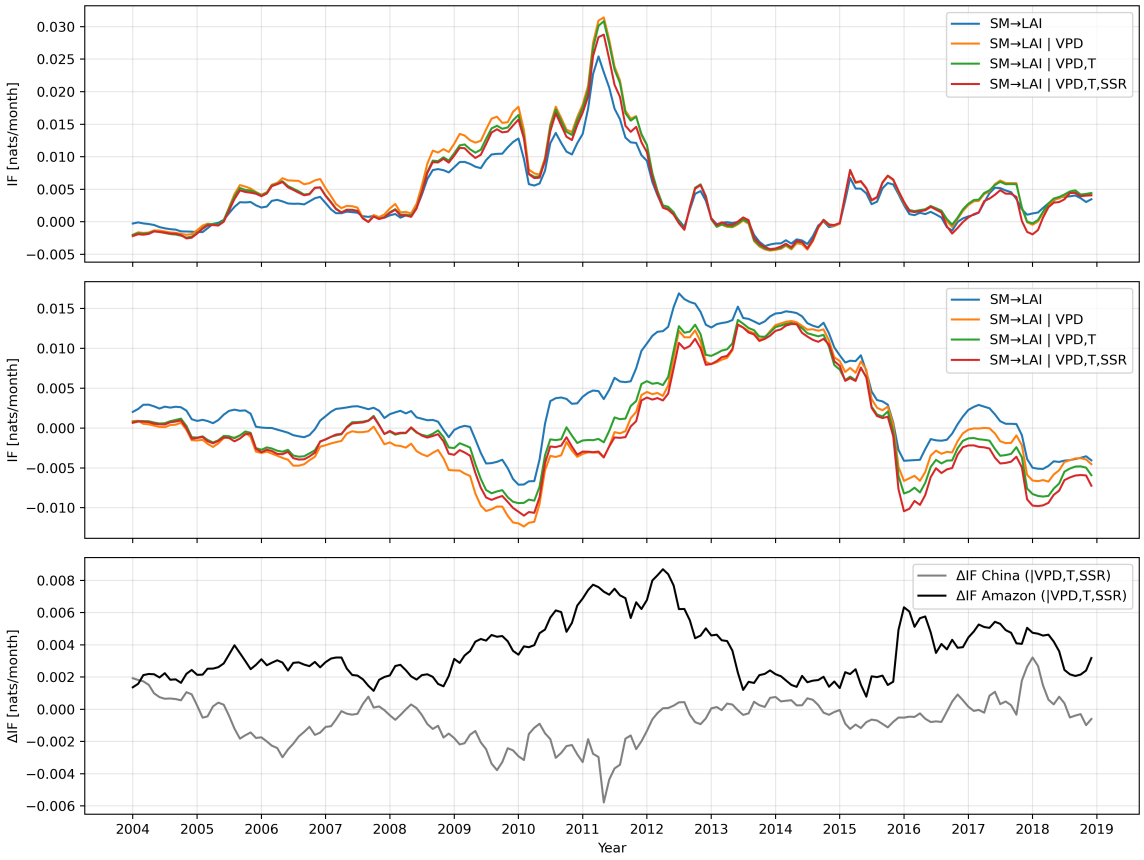

Figure 4: Information flow time series of the SM→LAI couplings based on different conditioner combinations for the selected region in China (top panel) and the Amazon basin (middle panel). the differential IFs (ΔIF) are shown in the bottom panel for the two regions in the instance where all the conditioners are used.

The gating weights are parameterised through unconstrained logits θ_k using a softmax transformation

$$\alpha_k = \frac{\exp(\theta_k)}{\sum_{\ell=1}^{4} \exp(\theta_\ell)},$$

which guarantees positivity and unit sum.

Feed-forward prediction network. The gated inputs $\tilde{\mathbf{x}}(t)$ are passed through a small multilayer perceptron (MLP) with two hidden layers:

$$\begin{aligned} \mathbf{h}_1 &= \mathrm{ReLU}(W_1 \tilde{\mathbf{x}} + b_1), \\ \mathbf{h}_2 &= \mathrm{ReLU}(W_2 \mathbf{h}_1 + b_2), \\ \hat{y}(t) &= W_3 \mathbf{h}_2 + b_3. \end{aligned}$$

In all experiments the hidden layers contained 32 neurons each, resulting in the architecture

$$4 \to 32 \to 32 \to 1.$$

This deliberately simple design was selected to prioritise interpretability of the input-layer gating weights rather than model complexity.

D.2. Loss Function and Regularisation

As noted above, the total loss function is computed as

$$\mathcal{L} = \mathcal{L}_{\mathrm{pred}} + \lambda \, \mathcal{L}_{\mathrm{CIS}},$$

where

- $\mathcal{L}_{\mathrm{pred}}$ is the mean squared error between predictions and observations,
- $\mathcal{L}_{\mathrm{CIS}} = \sum_{k=1}^{4} (\alpha_k - p_{\mathrm{CIS},k})^2$ penalises deviations of the learned gating weights from the CIS prior,
- $\lambda \geq 0$ controls the strength of causal regularisation.

Setting $\lambda = 0$ yields a purely data-driven network, while increasing λ enforces progressively stronger adherence to the causal prior.

D.3. Training Protocols used here

All networks were trained using the following protocol:

- Optimiser: Adam
- Learning rate: 10^{-3}
- Batch size: 256
- Number of epochs: 3000

- Weight decay: 0
- Data split: 75% training, 25% testing: we note that since these are time series analyses, the split is done chronologically
- Input preprocessing: Each predictor was standardised to zero mean and unit variance using statistics from the training set only.

To ensure robustness to random initialisation, each configuration was repeated for 20 independent seeds uniformly sampled between 80 and 150. Reported results represent the mean and standard deviation across these seeds.

D.4. Hyperparameter Selection

The regularisation strength λ was tuned through a systematic regularisation-path analysis (Figure 2), evaluating predictive performance and prior alignment for

$$\lambda \in \{0, 0.1, 0.5, 1, 2, 3, 4, 5\}.$$

Based on this analysis, $\lambda = 4$ was selected as the operating point providing an effective balance between

- high predictive skill (near-optimal RMSE and R^2),
- strong alignment of $\boldsymbol{\alpha}$ with $\mathbf{p}_{\mathrm{CIS}}$,
- stability of results to small perturbations of λ.

D.5. Implementation

The network was implemented in `PyTorch`. The global gating mechanism was realised as a dedicated learnable parameter vector, distinct from the MLP weights, enabling direct extraction and interpretation of $\boldsymbol{\alpha}$ after training.

This architecture was used consistently for all regions and targets to ensure comparability across experiments.

Appendix E. Practical Selection of the Regularisation Parameter λ

In this section, we provide details on how an appropriate value of the CIS regularisation parameter (λ) can be selected when applying the proposed methodology to new datasets or other applications.

E.1. Guiding Principles

λ controls the trade-off between two competing objectives:

- predictive accuracy, governed by the data-driven loss $\mathcal{L}_{\mathrm{pred}}$, and
- causal interpretability, enforced through the prior-alignment term $\mathcal{L}_{\mathrm{CIS}}$.

We note that there is no universally optimal value of λ; instead, it should be selected based on the goals of the analysis. Applications prioritising pure prediction may favour small values of λ, while applications seeking interpretable and causally consistent models should adopt larger values(Verdinelli and Wasserman, 2024; Lai and Wang, 2024; Kania and Wit, 2025).

To make this choice systematic and reproducible, we propose the following general procedure.

E.2. Recommended Tuning Procedure

For a new problem, we recommend the following steps:

Step 1: Define a candidate grid. Select a moderate range of candidate values spanning several orders of magnitude, for example

$$\lambda \in \{0, 0.1, 0.5, 1, 2, 3, 4, 5\}.$$

The range can be widened if strong sensitivity is observed.

Step 2: Train models along a regularisation path. For each candidate λ, train the CIS-guided neural network using identical data splits and hyperparameters. To reduce variance from random initialisation, repeat training across multiple random seeds and average the results.

Step 3: Evaluate three diagnostics. For each λ, compute:

1. Predictive performance (e.g., RMSE and R^2 on held-out data),
2. Prior-matching error $\|\boldsymbol{\alpha} - \mathbf{p}_{\text{CIS}}\|_2$,
3. The magnitude of the regularisation contribution λL_{reg}.

These quantities jointly characterise the accuracy–interpretability trade-off. See Figure 2 for our implementation in this study.

Step 4: Identify the Pareto-optimal region. Plotting the diagnostics against λ typically reveals three regimes:

- **Small** λ: high predictive skill but weak adherence to the causal prior,
- **Large** λ: strong prior alignment but deteriorating predictive accuracy,
- **Intermediate** λ: a plateau where predictive skill is near-optimal while prior alignment is substantially improved.

The recommended operating point lies in this intermediate plateau region.

E.3. Quantitative Selection Criteria

The above procedure may be formalised using one of the following criteria:

- Select the smallest λ such that
$$\|\boldsymbol{\alpha} - \mathbf{p}_{\text{CIS}}\|_2 \leq \tau,$$
where τ is an acceptable misalignment tolerance.

- Select the largest λ for which RMSE remains within a user-specified percentage (e.g., 1–2%) of its minimum value.
- Use a multi-objective score such as

$$S(\lambda) = \mathrm{RMSE}(\lambda) + \beta \, \|\boldsymbol{\alpha} - \mathbf{p}_{\mathrm{CIS}}\|_2,$$

  where β encodes a preference for interpretability.

These criteria are simple to compute and can be automated.

E.4. Robustness Considerations

In practice, the optimal region is often broad rather than a single sharp point. Once a suitable λ has been identified, we recommend verifying that:

- model conclusions are qualitatively stable to small perturbations of λ (e.g., $\pm 20\%$),
- the ordering of feature importances remains consistent across nearby values,
- predictive performance does not degrade abruptly near the chosen operating point.

If these conditions hold, the chosen λ can be considered robust.

E.5. Application to the Present Study

In this study, the above procedure led to the selection of $\lambda = 4$, which lies within the plateau where:

- predictive skill is effectively optimal,
- $\|\boldsymbol{\alpha} - \mathbf{p}_{\mathrm{CIS}}\|_2$ is small,
- results are stable for neighbouring values (e.g., $\lambda = 3$ or 5).

However, we emphasise that other applications may favour different operating points depending on data characteristics and analysis objectives.

E.6. Summary

When applying the framework, select λ using a regularisation-path analysis that explicitly balances predictive performance and prior alignment, and choose a value within the broad, robust plateau where both objectives are simultaneously satisfied.

This procedure ensures that the methodology can be transferred to new domains in a transparent and reproducible manner(Dominik Rothenhäusler, 2021; Shen, 2026).

Appendix F. Linear toy-model validation of the Causal Importance framework

In this section, we provide a simulation-based validation of the CIS framework for a linear dynamical system where the causal structure is known. We use a linear vector autoregressive (VAR(1)) system with explicitly prescribed mediation and confounding pathways. Thus, this experiment is designed to test whether (i) bivariate LKIF can exhibit inflated coupling when mediators/confounders exist, and (ii) conditioning and the CIS components correctly attribute the restructuring to the intended variables.

F.1. Ground-truth linear system with mediation and confounding

We consider a four-variable linear dynamical system with state vector $\mathbf{X}_t = (Z_t, X_t, M_t, Y_t)^\top$ evolving as a VAR(1) process:

$$\begin{aligned} Z_t &= \alpha_Z Z_{t-1} + \varepsilon_t^Z, \\ X_t &= \alpha_X X_{t-1} + \beta_{Z\to X} Z_{t-1} + \varepsilon_t^X, \\ M_t &= \alpha_M M_{t-1} + \beta_{X\to M} X_{t-1} + \varepsilon_t^M, \\ Y_t &= \alpha_Y Y_{t-1} + \beta_{M\to Y} M_{t-1} + \beta_{Z\to Y} Z_{t-1} + \varepsilon_t^Y, \end{aligned} \tag{35}$$

where $\varepsilon_t^Z, \varepsilon_t^X, \varepsilon_t^M, \varepsilon_t^Y$ are mutually independent, zero-mean Gaussian noises. This system contains (i) a mediated pathway $X \to M \to Y$ and (ii) a confounding pathway $Z \to X$ and $Z \to Y$, while the direct link $X \to Y$ is intentionally absent. Therefore, any apparent bivariate coupling $X \to Y$ must be attributable to mediation and/or confounding.

In the simulation presented here, we set $\alpha_Z = 0.6$, $\alpha_X = 0.7$, $\alpha_M = 0.5$, $\alpha_Y = 0.5$, $\beta_{Z\to X} = 0.8$, $\beta_{X\to M} = 0.9$, $\beta_{M\to Y} = 0.9$, and $\beta_{Z\to Y} = 0.8$. We generate N samples after burn-in and estimate information flows from the resulting time series.

F.2. Bivariate vs. conditional LKIF and sequential conditioning

We evaluate the bivariate information flow $\mathcal{T}_{X\to Y}(t)$ and the fully conditioned flow $\mathcal{T}_{X\to Y|M,Z}(t)$ using the conditional LKIF expression given in the Methods (Eq. 1) and Appendix A. For this stationary linear experiment, we report the time-aggregated estimates (equivalently, a single-window estimate), which preserves the interpretation of the CIS components that are defined as sums over time.

To attribute the restructuring to individual conditioning variables, we use the same sequential (incremental) definition adopted in the paper. Let $\mathcal{T}^{(0)}$ denote the bivariate flow, $\mathcal{T}^{(1)}$ denote the flow conditioned on the first variable in an ordered set, etc. For the order (M, Z), the incremental contributions are

$$\Delta IF^{(M)} = \mathcal{T}_{X\to Y|M} - \mathcal{T}_{X\to Y}, \qquad \Delta IF^{(Z)} = \mathcal{T}_{X\to Y|M,Z} - \mathcal{T}_{X\to Y|M}, \tag{36}$$

and the total differential flow is $\Delta IF = \mathcal{T}_{X\to Y|M,Z} - \mathcal{T}_{X\to Y} = \Delta IF^{(M)} + \Delta IF^{(Z)}$.

F.3. Toy-model results

The simulated system yields a substantial bivariate information flow from X to Y,

$$\mathcal{T}_{X\to Y} = 0.1481, \tag{37}$$

despite the fact that Eq. (35) contains no direct $X \to Y$ link. After conditioning on both the mediator M and the confounder Z, the conditional flow collapses to approximately zero,

$$\mathcal{T}_{X\to Y|M,Z} = -0.00468, \tag{38}$$

indicating that the apparent bivariate coupling is almost entirely explained by indirect/confounded pathways. Accordingly, the total differential information flow is

$$\Delta IF = \mathcal{T}_{X\to Y|M,Z} - \mathcal{T}_{X\to Y} = -0.1528, \tag{39}$$

Table 1: Linear toy-model LKIF results for the coupling $X \to Y$ in Eq. (35) (no direct $X \to Y$ link). All values are in nats per unit time.

Quantity	Symbol	Value
Bivariate information flow	$\mathcal{T}_{X\to Y}$	0.1481
Conditioned on mediator	$\mathcal{T}_{X\to Y\|M}$	0.04761
Conditioned on mediator and confounder	$\mathcal{T}_{X\to Y\|M,Z}$	-0.00468
Total differential flow	ΔIF	-0.1528
Sequential mediator increment	$\Delta IF^{(M)}$	-0.1005
Sequential confounder increment	$\Delta IF^{(Z)}$	-0.05230

i.e., conditioning strongly suppresses the inflated bivariate coupling.

Sequential attribution further partitions this suppression into a dominant mediator contribution and a smaller confounder contribution:

$$\Delta IF^{(M)} = -0.1005, \qquad \Delta IF^{(Z)} = -0.05230. \tag{40}$$

Thus, in this experiment, approximately two-thirds of the restructuring arises when conditioning on the mediator M, and the remaining one-third arises from conditioning on the confounder Z, consistent with the imposed structure and coupling strengths.

For convenience, Table 1 summarizes the principal quantities.

F.4. Behavior of CIS components under the linear system

We now apply the CIS components as defined in the main text to the linear toy model. Because the coupling magnitudes decrease monotonically as conditioning variables are added (Table 1), the moderation-gain terms are zero (i.e., there is no amplification of coupling magnitude induced by conditioning), so $MG_M = MG_Z = 0$ for this experiment.

The mediator-dominance index (MDI), which normalizes the absolute incremental restructuring across conditioning variables, yields

$$MDI_M = \frac{|\Delta IF^{(M)}|}{|\Delta IF^{(M)}| + |\Delta IF^{(Z)}|} \approx 0.6578, \qquad MDI_Z \approx 0.3422. \tag{41}$$

This correctly identifies the mediator as the dominant contributor to restructuring in this system.

The confounding-pressure (CP) terms (defined as reductions in magnitude relative to the bivariate flow and normalized across conditioners in the paper) give

$$CP_M \approx 0.4120, \qquad CP_Z \approx 0.5880, \tag{42}$$

reflecting that the final inclusion of Z is what collapses the coupling magnitude closest to zero (hence receiving larger CP credit under this normalization). We note that sequential attributions can depend on the conditioning order; here we use the order (M, Z) motivated by the known causal motif (mediator followed by confounder), and the qualitative conclusion (both M and Z strongly restructure the bivariate coupling, with M dominating the incremental restructuring) is robust.

Finally, the causal-sufficiency rate (CSR) is defined in the paper via a tolerance threshold τ applied to the discrepancy between bivariate and fully conditioned flows. In this toy model, the discrepancy $|\mathcal{T}_{X\to Y|M,Z} - \mathcal{T}_{X\to Y}| = 0.1528$ is large; therefore CSR is close to zero for any reasonable small tolerance (e.g., $\tau \ll 0.15$), indicating that bivariate coupling is not causally sufficient in the presence of the mediator and confounder.

Under equal weights in the CIS definition, the conditioner-specific CIS values (up to the common $(1 - CSR)$ term, which does not affect the relative ranking across conditioners) are

$$CIS_M \propto MDI_M + CP_M \approx 1.0698, \qquad CIS_Z \propto MDI_Z + CP_Z \approx 0.9302, \tag{43}$$

showing that both variables are causally important for correctly interpreting $X \to Y$, with the mediator slightly dominating overall importance in this parameter setting.

F.5. Implication

This linear toy-model experiment demonstrates that, even under a fully linear dynamical system with known dependencies, bivariate LKIF can exhibit apparent coupling induced by mediation and confounding, while conditional LKIF collapses to near-zero when the correct conditioning variables are included. Moreover, the CIS components behave consistently with their intended interpretations: the differential information flow captures the magnitude of restructuring induced by conditioning, MDI identifies the dominant restructuring variable (here the mediator), MG correctly indicates the absence of coupling amplification, and CP reflects the degree to which conditioning suppresses an inflated bivariate coupling. This provides an interpretable and transparent validation of our methodology under linear dynamics. Further experiments are presented in Siddique et al. (2026).